Advanced statistical post-processing of ensemble weather forecasts

Submitted by

Sam Allen

to the University of Exeter as a thesis for the degree of
Doctor of Philosophy in Mathematics

January 2021

Supervised by

Dr Frank Kwasniok,
Dr Christopher A. T. Ferro,
Dr Gavin R. Evans,
Dr Piers Buchanan.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signed: ....

Abstract

Today, weather forecasts are generated by evolving the current state of the atmosphere through time subject to established mathematical and physical laws. The resulting forecasts are considerably more accurate than those produced using any other approach that humans have devised to predict the weather. Nonetheless, these forecasts are imperfect. In particular, errors arise in the forecast due to limitations in both our theoretical understanding of the atmosphere and our practical ability to reproduce it. This, combined with the atmosphere's chaotic nature, means obtaining a perfect forecast of the future weather is, for practical purposes, impossible.

It is therefore imperative that a forecast is issued alongside its associated uncertainty. This is often achieved by generating an ensemble of weather forecasts that differ in their initial conditions, and possibly also the formulation of the dynamical weather model which with they are produced. However, due to errors in their construction, operational ensemble forecasts themselves possess systematic deficiencies. For this reason, it is necessary to apply an a posteriori adjustment to the ensemble forecast, so that it provides a more realistic representation of the weather that will occur.

Several statistical methods have been proposed for this purpose that can not only correct for systematic errors present in the dynamical models, but can issue forecasts that are probabilistic, thus accounting for the uncertainty inherent in the forecast scenario. Such statistical post-processing methods have become an integral component of operational forecasting suites over the last decade. Recently, however, studies have demonstrated that conventional post-processing methods can be ameliorated by leveraging additional sources of information within the statistical models. With this in mind, this thesis seeks to recognise circumstances under which the performance of dynamical weather models is expected to change, thereby indicating what information should be incorporated within statistical post-processing methods.

In particular, previous studies have indicated that the errors in dynamical weather models may depend on the occurrence of certain patterns in the synoptic-scale behaviour of the atmosphere, and we therefore postulate that these atmospheric regimes can be utilised when post-processing. A general framework for incorporating this regime information into established post-processing methods is proposed, and its merits are demonstrated in a variety of circumstances. A novel approach to evaluate the performance of forecasts is also introduced that can help to identify situations where incorporating information into post-processing methods is expected to be beneficial.

Acknowledgements

The work presented in this thesis has benefited greatly from discussions with numerous people, friends and colleagues alike. David Stephenson and Theo Economou, in particular, have always been willing to share their opinions and offer their advice, while the comments and discussions provided by Peter Challenor and Stéphane Vannitsem in reviewing this thesis have been invaluable.

I would also like to thank the entire Verification, Impacts and Post-Processing team at the UK Met Office for welcoming me during a short work placement, and I am indebted to Gavin Evans and Piers Buchanan for their guidance throughout this period.

Lastly, I cannot understate how grateful I am to Frank Kwasniok and Chris Ferro. Their supervision during this degree has not only significantly enhanced the quality of work presented in this thesis, but it will continue to influence any work I undertake in the future. Thank you.

Contents

List of Tables

List of Figures

1 Introduction

Accurate predictions of the future are fundamental for making decisions in the face of uncertainty. The weather, for example, has a profound influence on our day-to-day lives, and forecasts of the weather are used regularly to mitigate its impacts: several industries require insurance against losses incurred due to adverse weather conditions; airlines and shipping companies rely on high-quality weather forecasts in order to operate safely and efficiently; farming and agriculture depends, as it has done for generations, on predictions of crop yields; while weather forecasts also have important implications for the ever-growing renewable energy sector. Moreover, the ability to predict the occurrence of floods, heatwaves, and yet more extreme weather events has potentially life-saving consequences.

Forecasts of the weather have therefore existed in various guises throughout human civilisation, with written texts on the subject dating back as far as Aristotle's 'Meteorologica' in the 4th century BC. At the time, methods to predict the weather were hindered by the inability to routinely record and store meteorological data, and it was not until over 1500 years later that inventions such as the thermometer and barometer rendered this achievable. The ability to describe the instantaneous state of the atmosphere using measurements from such instruments marked a huge milestone in weather forecasting: in knowing the current atmospheric state, it became straightforward to reference previous instances in which similar weather occurred, in turn allowing for a greater appreciation of what may materialise in the future.

This was the primary basis of weather forecasting for hundreds of years; as described by Lynch (2008), forecasting during this period was "more of an art than a science." This changed following Sir Isaac Newton's pioneering work in the 17th century to define the laws of motion, which permitted a better understanding of the governing forces that drive the atmosphere. It was not until the early 20th century, however, that these laws were formally considered for use within weather forecasts. Lewis Fry Richardson hypothesised that Newton's laws of motion (applied to a fluid) could be used as the basis of a dynamical weather model, whereby the current atmospheric state is evolved through time according to these physical laws to derive an estimate of the future weather (Richardson, 2007). Although practical limitations meant the resulting forecast from his weather model was unrealistic, Richardson's seminal work paved the way for meteorologists to follow. In 1950, a team led by Jules Charney adapted the work of Richardson in light of recent developments in computing, and succeeded in constructing a reasonable weather forecast using these dynamical models (Charney et al., 1950). This approach, known as numerical weather prediction, has revolutionised weather forecasting.

Due to their ability to reproduce the atmosphere's underlying dynamics, numerical weather prediction models are now a cornerstone of operational weather forecasts. However, even the elaborate forecasts issued by these dynamical models become no better than guesswork when predicting the atmosphere sufficiently far into the future, suggesting there exists a limit to the atmosphere's predictability. Since the first dynamical weather forecast, weather models have become progressively more representative of the atmosphere, and incremental developments have steadily pushed back this limit of predictability, due both to an increased understanding of atmospheric processes, along with a proliferation in computational resources. As such, the quality of weather forecasts has improved substantially over this period: a forecast made one day in advance 50 years ago possessed roughly the same amount of skill as a forecast made six days in advance from a current weather prediction model (Bauer et al., 2015).

Yet, although the quality of weather forecasts has reached a level beyond which could have been imaginable even a century ago, these forecasts are still imperfect. In particular, uncertainties in the forecasting process arise due to an incomplete knowledge of the initial atmospheric conditions from which to run the numerical weather model, as well as errors in the formulation of the model itself. Due to the seemingly chaotic nature of the atmosphere, these imperfections can induce considerable biases in the resulting forecasts. Therefore, a single forecast from such an imperfect prediction system contains little information regarding the weather that will occur in the future.

Instead, in the last few decades, forecasters have realised the imperative need to account for the uncertainty in their prediction. For this purpose, Leith (1974) proposed running a numerical weather model several times, each initialised from different starting conditions, to generate a collection of distinct forecasts. This collection of forecasts, now known itself as an ensemble forecast, conveys the uncertainty surrounding the future weather: if there is strong agreement between the individual model runs, then the forecaster can be confident about the weather that will materialise, whereas high variation between the different forecasts suggests a range of possible weather events could occur (Leutbecher and Palmer, 2008).

The utility of these ensemble forecasts was soon recognised, and they began to be implemented operationally in the 1990s (Toth and Kalnay, 1993; Kalnay et al., 1996). However, ensemble forecasts are products of imperfect weather models, and are hence themselves subject to systematic biases. For example, ensemble forecasts made for surface weather variables are regularly found to be overconfident, exhibiting considerably less spread than required (e.g. Hamill and Colucci, 1997). To address this, several statistical methods have been proposed that adjust the output from numerical weather models so that it better reflects the weather likely to occur. This is referred to as statistical post-processing. Post-processing methods not only correct for recurring forecast errors, but they translate the forecasts from the discrete model phase space to the real atmosphere (Stephenson et al., 2005).

Like ensemble forecasts, this approach has been hugely successful, and numerous studies have highlighted the enhanced performance of forecasts having undergone statistical post-processing. However, in its most common form, statistical post-processing serves only to calibrate the ensemble forecast, so that it aligns with previous weather observations. Recent approaches, meanwhile, have recognised the relative ease with which extra information can be incorporated into the statistical methods, in turn addressing more nuanced structures in the errors of numerical weather models. The focus of this thesis is the amelioration of conventional statistical post-processing methods in a similar vein, where the principal question concerns what additional information should be utilised when post-processing.

In particular, we postulate that the biases in numerical weather prediction models depend on the prevailing behaviour of the atmosphere's circulation, which is often represented using distinct weather regimes. These weather regimes can be thought of dynamically as quasi-stationary equilibria in the atmosphere, defined more generally as structures in the atmosphere's circulation that tend to persist beyond the timescales of individual weather events, and recur at fixed geographical locations (Hannachi et al., 2017). Weather regimes are a well-known meteorological phenomenon, due largely to the profound influence they assert on local weather systems, and recent studies have considered the ability of numerical weather models to simulate the regime behaviour observed in reality (Dawson et al., 2012; Ferranti et al., 2015; Matsueda and Palmer, 2018). Using the results from such studies, we investigate how the occurrence of certain weather regimes affects the biases present in dynamical weather models, focusing in particular on medium-range forecasts of surface weather variables. The work presented herein then explores to what extent this can be utilised in practice to improve the effectiveness of statistical post-processing methods.

The remainder of this thesis is organised as follows. The following chapter describes in more detail the key components of weather forecasting, including numerical weather models, ensemble forecasts, and prominent statistical post-processing methods. Techniques to evaluate the performance of weather forecasts are also introduced, which relate to the field of forecast verification. This review of the extant literature is by no means exhaustive, and relevant studies are referenced throughout: the primary goal of this chapter is to introduce the methods and studies that have influenced the work presented in subsequent chapters of this thesis.

Chapter 3 presents a formal introduction to the concept of weather regimes, and introduces a framework for utilising these regimes within statistical post-processing. Regime-dependent extensions of two state-of-the-art statistical post-processing methods are described, which are later trialled on a highly idealised model of the atmosphere in which regime behaviour has previously been studied. This chapter has been published in Allen et al. (2019). Chapter 4 extends this further by applying the techniques to more realistic data sets, namely wind speed forecasts from a quasigeostrophic weather model, and from a large set of hindcasts generated as part of the Global Ensemble Forecasting System (GEFS) Reforecasting Project at the National Oceanic and Atmospheric Association (NOAA; Hamill et al., 2013). This chapter, published in Allen et al. (2020), also discusses two approaches to deduce the prevailing regime from observed or predicted weather fields, and it thus presents a self-contained example of how the regime-dependent methods can be applied in practice. A variant of the regime-dependent approach is then applied to operational wind speed forecasts from the Met Office's global ensemble prediction system in Chapter 5, and this has been published in Allen et al. (2021b). Using results from the previous chapters, an extension to the approach is administered to identify situations in which regime-dependent post-processing is expected to be most beneficial, thereby maximising the utility of the weather regimes, while also accounting for the limited amounts of data available in operational settings.

In Chapter 6, we build upon the existing functionality within the UK Met Office's soon-to-be operational post-processing suite, focusing in particular on the calibration of temperature forecast fields. Although temperature is expected to depend on the prevailing weather regime, it also exhibits a pronounced seasonal cycle. Rather than utilising weather regimes when calibrating these temperature fields, we discuss alternative extensions to conventional post-processing methods that are better suited to capture the dependence of forecast errors on the time of the year. In particular, these extensions relax the distributional assumptions that are commonly made when post-processing temperature forecasts, allowing the post-processing methods to account for changes in the shape of the temperature distribution throughout the year. This chapter has been submitted for publication in Monthly Weather Review (Allen et al., 2021a).

Finally, in Chapter 7, attention is turned towards the evaluation of forecast performance. Forecasts are commonly assessed using scoring rules, which provide a numerical value that objectively quantifies the accuracy of the forecast. It has previously been demonstrated that such scoring rules can be decomposed into constituent terms, each of which describes a different aspect of the forecast performance. In this chapter, it is shown that a further decomposition of scoring rules exists that can provide additional information regarding the properties of a forecast in certain circumstances. The decomposition proposed in this chapter sheds light on when to expect improvements by utilising alternative sources of information when post-processing, helping to explain the results presented in previous chapters. This is discussed further in some closing remarks in Chapter 8.

2 Forecasting the weather

2.1 Numerical weather prediction

2.1.1 Dynamical weather models

In theory, the instantaneous state of the atmosphere can be described by measurements of certain meteorological variables at every point in space. This atmospheric state will then change, or evolve over time, according to some underlying physical laws. The fundamental problem in weather and climate forecasting is to understand the nature of this evolution.

It is conceivable that, given exact knowledge of the atmospheric state, along with a complete understanding of the governing laws, it would be possible to obtain a perfect weather forecast for any length of time into the future. In reality, neither of these are known. Nevertheless, dynamical weather prediction models aim to replicate this procedure. To do so, they simplify the problem at hand by representing the atmospheric state using values of only a few fundamental meteorological variables at a finite number of points in space. These points are generally structured in a three-dimensional grid above the Earth's surface, where the number of vertical layers in the grid defines the model's vertical resolution, and the distance between neighbouring grid points within a vertical layer determines the model's horizontal resolution. Models that operate on a finer grid are said to exhibit a higher resolution, since they are able to resolve more intricate features of the atmosphere. Weather models may also differ in the variables they use to describe the atmosphere, though temperature, wind velocity, specific humidity and surface pressure are all typically included (Buizza, 2018). The instantaneous atmospheric state is thus represented by an array of grid points containing values of each variable at every spatial location, referred to as the model's phase space.

Given the approximated atmospheric state, the future atmospheric conditions can be obtained by evolving this initial state through time, subject to established mathematical and physical laws. These laws are comprised of several partial differential equations, governed principally by the Navier-Stokes equation - essentially Newton's second law applied to a fluid - along with formulae for the conservation of mass and energy (Vallis, 2017). It is common to look for solutions to this system of equations in spectral space (Buizza, 2018), though understanding the behaviour of such solutions is notoriously difficult. As such, a numerical, finite-difference approach is typically employed to integrate these equations forward in time.

Despite the substantial simplifications that are put in place, these dynamical models have revolutionised weather forecasting. However, these forecasts are themselves imperfect. Due to the intricacy of the atmosphere, there are physical processes that operate on scales smaller than those resolved by the model, and the model therefore cannot represent the effect these processes have on the weather. Increasing the model resolution - i.e. the number of grid points at which the underlying equations are evaluated - would alleviate this, but the additional computations required, as well as the extra storage necessary, mean the resolution of the model is restricted by the available computational resources. In fact, weather forecasting operates on the limits of our technological understanding, and recent improvements in the quality of weather forecasts over the last century are largely attributable to advances in computational capacity (Alley et al., 2019). This is likely to remain true for some time.

In the meantime, while model resolution cannot be increased, it is still necessary to account for the smaller-scale effects that are unresolved by the model. This includes, for example, convection that operates on sub-grid-scales, or the effect of local orographic properties. These convoluted features of the atmosphere are typically represented using parametrisation schemes, which model the influence the unresolved scales have on the resolved scales. These schemes do not attempt to explicitly model these sub-grid-scale phenomena, but rather try to capture the effect they have on the resolved features. Parametrisation schemes are discussed in more detail in Bauer et al. (2015), while a comprehensive overview of numerical weather prediction models is available in (Lynch, 2006).

2.1.2 Data assimilation

Although the evolution of a starting point through time according to certain mathematical rules is conceptually simple, such a starting point is difficult to deduce in practice. Furthermore, even infinitesimal errors in the starting conditions can have a significant effect on the resulting forecast. Considerable effort has thus been devoted to developing approaches that collect and combine observations from various sources, and integrate all available data into a best guess of the atmospheric state at a given time. Such approaches relate to the field of data assimilation.

Advances in data assimilation techniques have contributed to significant improvements in forecast quality this century (Alley et al., 2019), adapting to the proliferation in the volume of data available, particularly from the vast coverage afforded by satellites. However, observations alone are not sufficient to estimate the entire state of the atmosphere: the number of available atmospheric observations at a given time is typically considerably less than the dimension of a weather model's phase space (Kalnay, 2003). Therefore, it is necessary to define a background field, containing prior information regarding the atmospheric state, which constitutes a first guess for the model's initial conditions. This background field is then updated in light of the available observations.

Today, the prior information in the background field is almost always encapsulated by the output of a high resolution dynamical weather model, run over a short period of time, called the ‘assimilation window’. Since this forecast is defined in the model’s phase space, it provides complete spatial coverage over the domain, in contrast to atmospheric observations, which tend to be available at an irregular network of locations. Although this short-range forecast will be subject to errors, combining this information with observations of meteorological variables throughout the assimilation window ensures the predicted field provides a reasonable estimate of the true atmospheric state. This creates an objective analysis field (more generally known as the ‘analysis’), which can be thought of as an ongoing short-range forecast that is continually realigned to be consistent with recent observations (Kalnay, 2003).

The way in which the observations and the short-range forecast are combined depends on the data assimilation method applied, and a description of such approaches is beyond the scope of this thesis. An introduction to data assimilation techniques is available in Talagrand (1997a), while more comprehensive overviews can be found in Daley (1993) and Kalnay (2003). Regardless of the data assimilation scheme, due to the incompleteness of atmospheric observations, as well as imperfect forecasts on which to base the background field, there will always be error in the forecast’s initial conditions. This error, although small, can have a monumental impact on the resulting forecast.

2.1.3 Limits of predictability

Despite the information provided by numerical weather models, the resulting forecasts still become useless beyond a certain time. This prompted studies into the predictability of the atmosphere, and whether there was a limit to how far in advance it was possible to predict the weather before all skill in the forecast was lost (Gleeson, 1967; Leith, 1971). Thompson (1957) defines the atmosphere’s predictability as “not merely the extent to which its behaviour is predicted in practice, but the extent to which it is possible to predict it with a theoretically complete knowledge of the physical laws that govern it.” Thompson considered the case where the numerical models described above were in fact correct, but even with a perfect model formulation, two forecasts with imperceptible differences in their initial states were found to diverge with time.

This result was confounded by the seminal work of Lorenz (1963), which introduced a highly idealised model of the atmosphere, controlled by three crude convection equations, and used this to illustrate the system’s sensitivity to initial conditions. Lorenz remarks that the atmosphere is ostensibly non-periodic and hence unstable, concluding that long-range forecasts are likely no more skilful than pure guesswork unless the initial conditions are known exactly. Dynamical systems that exhibit this property, such as the atmosphere, are now referred to as chaotic systems (Lorenz, 2014). This conclusion changed the direction of weather prediction, highlighting the importance of quantifying any uncertainty around a forecast, since even a perfect model configuration would provide an inaccurate prediction if the assumed initial state of the atmosphere was wrong. Indeed, there are several factors that contribute to uncertainty in the initial conditions, including precision errors arising from the instruments used to measure weather variables, a scarcity of data meaning a complete picture of the atmospheric state is unattainable, or limitations in the method used to discern the analysis from the data at hand (Fleming, 1971). Lorenz's work therefore demonstrates that a perfect forecast of the future atmospheric state is, for practical purposes, impossible.

This reinforced a notion that had been growing throughout the previous decade: weather forecasts should not be deterministic in nature, but should utilise both statistical and dynamical concepts. That is, although the atmosphere itself may be determinate, so that one state will always evolve in the same manner given identical starting conditions, a single, deterministic forecast of the atmosphere will provide little information unless a measure of its associated uncertainty is also presented. Gleeson (1970) therefore looked to unify the fields of statistics and meteorology through the theory of statistical dynamics. Gleeson notes that it should be understood that several possible initial conditions exist, and there is no way of knowing which corresponds to the 'true' state of the atmosphere. Therefore, rather than integrating the chosen weather model forward in time only from the current best guess of the atmospheric state (the analysis), information about the entire distribution of possible initial states should be utilised.

This is corroborated by results in Epstein (1969a), which show, albeit in simplified dynamical systems, that models initialised from the best guess of the initial forecast state do not in general result in the most accurate forecast. Therefore, Epstein (1969c) proposed Stochastic Dynamic Prediction, a framework that "seeks solutions corresponding to probabilistic statements of the initial conditions" while retaining the assumption that the laws governing the atmosphere are entirely deterministic. The framework acknowledges the irremovable error in the analysis and suggests that specifying a probability distribution for the initial conditions and numerically integrating this through time, using the deterministic governing equations, would produce an appropriate probability distribution for the atmosphere's future state. Although the evolution of this initial distribution through time can be described mathematically (Fleming, 1971), solutions are intractable for the large number of dimensions present in the typical phase space of a numerical weather model (Leutbecher and Palmer, 2008). Instead, Leith (1974) proposed a more pragmatic approach, approximating this procedure using Monte Carlo simulations, an approach now known as ensemble forecasting (Lewis, 2005).

2.2 Ensemble forecasts

2.2.1 Theory

The growing need to combine statistical and dynamical approaches when forecasting the weather stimulated the development of ensemble forecasts. Ensemble forecasts, introduced by Epstein (1969c) and later Leith (1974), select a finite sample of initial points in the model's phase space, from which the numerical weather model is integrated forward in time. The result is a collection of forecasts (called an ensemble), each of which provides its own, unique simulation of the atmosphere's trajectory. Ensemble forecasting can thus be thought of as a Monte Carlo approximation to Stochastic Dynamic Prediction, whereby initial conditions are sampled from a probability distribution for the current atmospheric state, and the resulting ensemble members are assumed to represent draws from the probability distribution of the future atmosphere. That is, the ensemble members form an empirical distribution for the future atmospheric state, thereby allowing weather forecasts to transition from the deterministic realm to the probabilistic (Palmer, 2002).

However, numerically evaluating these complex dynamical weather models is time consuming and computationally expensive, imposing a practical constraint on the size of the ensemble (i.e. the number of member forecasts that comprise the ensemble). Hence, operational ensemble forecasts are generally relatively small, and the distribution realised by the ensemble is therefore rather coarse (Ferro et al., 2008). Despite this, the utility of an ensemble forecast was quickly recognised, and they began to be implemented operationally in the 1990s (Toth and Kalnay, 1993; Molteni et al., 1996). How best to extract information from an ensemble, however, was not initially known (Anderson, 1996; Stephenson and Doblas-Reyes, 2000). Since weather forecasts had hitherto been deterministic in nature, the ensemble output was at first condensed into the mean of its constituent members, generating a single forecast field. Although this ensemble mean forecast was repeatedly found to outperform the individual ensemble members, ensemble forecasts provide considerably more information than just that available from the ensemble mean.

In particular, several studies have identified a connection between the spread of the ensemble members and the performance of the ensemble mean forecast (Whitaker and Loughe, 1998). This 'spread-error' or 'spread-skill' relationship reflects the ability of the ensemble to distinguish between situations of low and high predictability, and the spread of the ensemble members therefore provides a measure of the flow-dependent uncertainty present in the current forecast: if the ensemble members all exhibit similar trajectories then the forecaster can be confident about the resulting outcome, whereas if the members show little agreement then a range of possible outcomes may occur.

However, it has repeatedly been shown that operational ensemble forecasts of surface weather variables are underdispersed, or overconfident, with the observed weather value falling outside the range of the ensemble members significantly more often than expected (Hamill and Colucci, 1997; Buizza et al., 1999). This thus suggests that errors in the prediction system grow faster than the rate at which the ensemble members diverge from one another. Therefore, although ensemble forecasts are unequivocally more informative than an individual point forecast - which is equivalent to an ensemble comprised of one member - there are sources of uncertainty in the forecast that are still ignored when generating the ensembles. These error sources are generally divided into two categories: error in specifying the forecast's initial conditions, and error in the model itself. As noted by Leutbecher and Palmer (2008), however, although it is convenient to treat them separately, the initial condition error and the model error are inherently connected, since a numerical weather model is also used to estimate the initial state (see Section 2.1.2).

2.2.2 Initial condition uncertainty

This first source of error relates to how the initial conditions of the ensemble forecast are chosen. The distribution of the initial forecast states should reflect the uncertainties in the analysis, and the forecast members initialised at these points should thus represent the most probable trajectories of the atmosphere (Magnusson et al., 2008). A variety of approaches have been proposed to generate the ensemble's initial conditions, and no single method is widely recognised as more appropriate than the others.

Initially, these methods involved perturbing the model analysis field identified using data assimilation. In this case, the deterministic forecast initialised from the unperturbed analysis field is generally also included as a member of the ensemble, and is termed the ensemble's control member. The perturbation to the analysis could simply be random noise, as was proposed by both Epstein (1969c) and Leith (1974), though it is more common to select points in phase space from which the model's trajectory is likely to diverge rapidly from that of the control member. In doing so, the distribution of these points adequately reflects the variation in the probability distribution for the initial conditions. Toth and Kalnay (1993), for example, propose 'breeding' initial perturbations based on the dispersion of a short-range ensemble forecast, thus exploiting the natural evolution of errors in the model's phase space, whereas Buizza and Palmer (1995) introduce an approach centred around singular vectors, which similarly seeks to sample the fastest-growing forecast errors (Buizza et al., 2005).

More recently, however, ensemble-based data assimilation techniques have been developed to obtain an ensemble of analysis fields (Buizza et al., 2008). In this sense, the goal of data assimilation is not only to provide the best guess of the instantaneous atmospheric state, but to generate a suitable ensemble of possible states from the probability distribution of the initial conditions. Such approaches include, for example, the ensemble Kalman filter (Evensen, 1994), the ensemble transform (Wei et al., 2008), and the ensemble data assimilation approach (Buizza et al., 2008). Precise details on these methods are omitted here, though the basic premise is to apply the data assimilation scheme multiple times - using either perturbed observations to simulate measurement errors (Houtekamer et al., 1996), or an ensemble of short-range forecasts to simulate model errors - to obtain distinct analyses. This collection of analyses then constitutes the starting points from which to initialise the ensemble members.

Ensemble-based data assimilation techniques combine the data assimilation with the generation of an ensemble of initial conditions, in contrast to the bred and singular vector approaches, though combinations of these different techniques have also been utilised operationally to good effect (Buizza et al., 2008). Several studies have looked to compare these approaches (Buizza et al., 2005; Bowler, 2006; Magnusson et al., 2008), and readers are diverted to references therein for a more complete overview of the field. A brief description of these methods is also available in Kalnay (2003).

2.2.3 Model uncertainty

The previous section highlighted that even if the physical equations governing the atmosphere were known exactly, a forecast would still be imperfect due to errors in the initial conditions. Similarly, even if it were possible to obtain a perfect formulation of the initial conditions, the trajectory of the forecast would deviate from reality since all numerical weather prediction models exhibit errors. There are several sources that contribute to this model error: an incomplete knowledge of the physical processes governing the atmospheric flow; the numerical finite-difference approach used to integrate the model forward in time; the spatial truncation of the model, meaning small-scale physical processes are left unresolved; errors in the subsequent parametrisations of these sub-grid-scale effects, and so on. Consequently, accounting for these errors that arise due to imperfections in the model is arguably more difficult than addressing the initial condition uncertainty.

Nonetheless, just as the error in the analysis field is addressed by initialising the numerical weather model at a variety of possible initial conditions, model uncertainty could be partially represented in the ensemble forecast by changing the formulation of the model with which the ensemble members are constructed. Houtekamer et al. (1996), for example, propose issuing ensemble members that differ not only in their initial conditions, but also in their representation of sub-grid-scale physical processes. That is, the parametrisation scheme used to capture the unknown scales of motion is configured differently for each ensemble member (Leutbecher et al., 2017). Buizza et al. (1999) use a stochastic approach for the same purpose, where the parametrised tendencies corresponding to each ensemble member are scaled by a random number, and these stochastic parametrisation schemes have since become a key component of operational ensemble forecasts (Slingo and Palmer, 2011).

To account for the other sources of model error, further discrepancies between the ensemble members could be incorporated. In practice, different operational centres will employ ensemble prediction systems with varying model configurations, including their horizontal and vertical resolutions, sub-grid-scale parametrisation schemes, and their methods of generating the ensemble's initial conditions. Combining forecasts from several operational models may therefore better reflect the possible outcomes that could materialise. As such, these 'multi-model' ensemble forecasts have repeatedly been found to outperform single model ensembles (Kharin and Zwiers, 2002). It is often the case that the members of each ensemble prediction system are highly correlated and indicate a high degree of confidence in the forecast, yet there tends to be disagreement between the ensembles generated from the different models. As a result, the pooled spread of the multi-model ensemble provides a better reflection of the uncertainty present in the current situation, when compared to any of the constituent ensembles. Hagedorn et al. (2005) attributed this improvement to extra information in the ensembles, gained as a result of the various models exploring different possible outcomes (Johnson and Swinbank, 2009). Even these multi-model approaches, however, do not capture all uncertainty present in the construction of the ensemble prediction system.

2.3 Statistical post-processing

2.3.1 Introduction

Enhanced data assimilation techniques, stochastic parametrisation schemes, and an increase in computational resources have all contributed to substantial improvements in the performance of weather forecasts in recent decades, due largely to a better representation of the forecast uncertainties. Nevertheless, current ensemble prediction systems still possess several shortcomings. If these shortcomings were known, then they could be addressed after having obtained the model output. For example, the spread of the ensemble members could be augmented to account for a lack of dispersion, where the extent of this increase in spread is determined by previous forecast dispersion errors. More generally, systematic forecast errors can be identified by searching for patterns in an archive of historical forecasts and observations, and an improved prediction could then be obtained by removing these errors from the current forecast. This is the basis of statistical post-processing.

Statistical post-processing combines the output from a numerical weather model with information regarding previously observed weather events to provide a more accurate estimate of what the future atmosphere will look like. In doing so, post-processing can not only remove any recurring biases in the model output, but it can also generate predictions for meteorological quantities not present in the model's phase space. This includes, for example, issuing forecasts at spatial locations not positioned on the model grid (i.e. statistical downscaling, see e.g. Scheuerer et al., 2013), or predicting auxiliary weather variables, such as road surface temperature or rainfall runoff, both of which might rely on temperature and precipitation forecasts along with knowledge of external factors. In this sense, statistical post-processing methods translate the forecast output from the discrete model phase space to the continuous real atmosphere (Stephenson et al., 2005).

Throughout the subsequent chapters of this thesis, statistical post-processing will refer to the statistical correction of weather forecasts, possibly after having first interpolated the model output to specific locations. This is the most common application of statistical post-processing, in which case it is also known as forecast recalibration. Perhaps the simplest example of post-processing in this form is removing a constant bias from a deterministic prediction. Similarly, ensemble forecasts may exhibit probabilistic biases, and addressing such errors requires consideration not only of the ensemble's location, but also of the dispersion and shape of the forecast distribution.

Moreover, it may be the case that the forecaster is aware of the limitations of their weather model, but cannot address them explicitly when constructing the prediction system, due to a lack of computational resources, for example. In this sense, post-processing can be thought of as a temporary solution to address these errors inherent in the forecast system, until it is feasible to include further information directly into the formulation of the prediction system. Recent studies have extended established post-processing methods so that they not only recalibrate the ensemble forecast, but also use other sources to augment the information content present in the prediction. These methods, as well as more conventional approaches, will be discussed further throughout this section. However, this section seeks only to introduce the most prominent post-processing approaches, focusing in particular on those to be employed in the remainder of this thesis: we do not provide a comprehensive overview of all statistical post-processing techniques. To this end, a considerably more thorough review of statistical post-processing methods to recalibrate weather forecasts can be found in Vannitsem et al. (2018).

2.3.2 Predictive distributions

Due to the seemingly chaotic nature of the atmosphere and the inherent uncertainty when specifying the forecast's initial state, a single point forecast contains little information regarding the atmosphere's evolution. Instead, it is also necessary to provide a measure of the forecast's uncertainty. Recent decades have thus seen a permanent shift from deterministic to probabilistic weather predictions (Gneiting et al., 2007), emanating from the introduction of ensemble forecasts. An ensemble forecast provides a sample of the possible outcomes that could materialise, and thus constitutes an empirical probability distribution for the response variable. However, ensembles are necessarily finite in size and thus inherently unreliable (Buizza and Palmer, 1998; Weigel et al., 2007; Ferro et al., 2008). Instead, statistical post-processing methods have been proposed that issue forecasts in the form of smooth predictive distributions, which represent the conditional distribution of the outcome given the ensemble forecast. Although there has been recent interest in the application of non- or semi-parametric post-processing methods (Taillardat et al., 2016; Henzi et al., 2019), it is common to attain a smooth forecast distribution by assuming that the response variable follows a suitable parametric distribution.

Consider forecasts made for a single weather variable (e.g. temperature, wind speed, precipitation), at a given point in space and time, and for a particular length of time into the future - known as the lead time, or forecast horizon. The variable being forecast is denoted by Y and is referred to as the outcome, predictand or response variable. This outcome is assumed to be a random variable that follows an unknown distribution, and the value of the outcome that materialises is termed the observation or verification, denoted y. The corresponding forecast for this variable is an ensemble comprised of M members $\boldsymbol{x} = (x_1, ..., x_M)$, with mean $\bar{x}$ and variance s_x^2. The ensemble members are assumed here to be known, fixed values, though they could also be treated as random quantities to account for the uncertainty in their construction (Siegert et al., 2016b). In reality, a sequence of forecasts and observations is typically available at several spatial locations, and further subscripts could be used to denote the time, location and variable under consideration, though this is omitted in the following for ease of notation.

2.3.2.1 Non-homogeneous Regression

The simplest forecast in the form of a predictive distribution is the unconditional distribution of the response variable, better known in the meteorological community as the variable's climatology. Such a forecast might assume that the outcome follows a particular distribution. Temperature, for example, is commonly modelled using a Gaussian distribution (Von Storch and Zwiers, 2001):

$$Y \sim N(\mu, \sigma^2), \tag{1}$$

where μ and σ^2 denote mean and variance parameters to be estimated from previous observations. However, due to the abundance of weather data, the empirical distribution of the observed values is generally fairly smooth itself, and hence relevant quantiles or functionals of the distribution can be reliably estimated without having to specify a particular parametric distribution. The unconditional distribution depends only on previous observations of the weather variable of interest. Therefore, although it encapsulates all uncertainty present in the predictand, the climatological forecast distribution contains no information regarding the current atmospheric state, and hence does not change throughout time, nor for different forecast horizons. This forecast is thus commonly implemented as a reference forecast to which more sophisticated methods are compared (Mason, 2004). Temporal and spatial information can be added to this forecast, however, by restricting the data on which to estimate the distribution (see Section 2.3.3.1).

More advanced statistical modelling techniques such as time series (Wilks, 2019), or linear regression-based approaches (McCullagh, 2018) could be used to identify temporal, spatial and inter-variable relationships between meteorological observations, which could in turn generate forecasts that more accurately capture the prevailing dependencies in the atmosphere. However, in the short- and medium-range, numerical weather models contain considerably more information than these purely statistical methods, and the model output should hence be utilised in the forecast. Wilks (2002) suggest smoothing ensemble forecasts by issuing Gaussian distributions centred on the ensemble mean, and with variance equal to the ensemble variance:

$$Y|\boldsymbol{x} \sim N(\bar{x}, s_x^2), \tag{2}$$

potentially after having first applied a suitable transformation to the data. The dependence of the forecast on the ensemble members is made explicit in Equation 2. Conceptually, a smooth predictive distribution can be thought of as an ensemble comprised of infinite members, and statistical distributions therefore provide a computationally feasible way of increasing the ensemble size (Roulston and Smith, 2003). These parametric distributions are therefore particularly useful in the case of small ensembles. However, smoothing the ensemble accounts only for the finite size of the ensemble, rather than correcting any shortcomings in the numerical model.

Instead, the model output could be used in conjunction with the linear regression framework, where the ensemble mean is used as the sole predictor:

$$Y|\boldsymbol{x} \sim N(\alpha + \beta\bar{x}, \sigma^2). \tag{3}$$

This approach is typically referred to as Model Output Statistics (MOS; Glahn et al., 2009). Unlike Equation 2, Equation 3 relies not only on the ensemble forecast, but also on additional post-processing parameters that can account for conditional and unconditional biases in the ensemble mean.

The ensemble mean filters out the contrasting information provided by the various ensemble members (Kalnay, 2003), but collapsing the information provided by the ensemble into just its mean assumes that the members exhibit identical error characteristics. If this is the case then the ensemble members are said to be statistically exchangeable. When the ensemble members are not exchangeable, as is the case in multi-model ensembles or when a control member is present, each ensemble member, or the mean of each group of exchangeable members, should be included individually as inputs to the linear regression model (Fraley et al., 2010). Alternatively, a preliminary aggregation of the ensemble members could be performed to obtain a suitable deterministic forecast to be used as a predictor in Equation 3, in place of the ensemble mean. So-called 'consensus forecasts' are generally constructed by a weighted affine combination of the ensemble members (Greybush et al., 2008), where the optimal weights are parameters to be estimated from historical data. Rather than treating the ensemble members equally, consensus forecasts allow more accurate forecasts to be assigned higher leverage when constructing the deterministic prediction. More elaborate methods to blend forecasts are also available (Vannitsem et al., 2020). Regardless of the predictors, however, MOS translates this deterministic prediction to a forecast in the form of a predictive distribution.

Furthermore, although now commonly applied to ensemble output, linear regression was used to recalibrate output from numerical weather models prior to the advent of ensemble forecasts (Klein et al., 1959; Glahn and Lowry, 1972). In this case, the predictor of the regression model (Equation 3) is simply the single deterministic prediction, perhaps along with additional input variables, highlighting how statistical methods can be used to provide probabilistic forecasts without first generating an ensemble (Leutbecher and Palmer, 2008). Unlike deterministic predictions, however, ensemble forecasts contain useful flow-dependent information regarding the atmosphere's predictability. Therefore, Gneiting et al. (2005) extend MOS by including the ensemble variance as a predictor for the forecast variance, resulting in a heteroscedastic distribution:

$$Y|\boldsymbol{x} \sim N(\alpha + \beta\bar{x}, \gamma + \delta s_x^2). \tag{4}$$

The forecast distribution in this form is known as Non-homogeneous Gaussian Regression (NGR). As will be discussed in Section 2.3.2.3, if different assumptions are made about the choice of distributional family, then the method is more generally referred to as Non-homogeneous Regression (NR), or Ensemble Model Output Statistics (EMOS). Equation 4 constitutes a very general class of post-processing models that encompasses more primitive methods, including Equations 1 - 3. It can be extended further by including the forecast time as a predictor, making it more suitable for post-processing seasonal and decadal predictions where it is often necessary to remove time-dependent trends in the forecast biases (Sansom et al., 2016).

2.3.2.2 Bayesian Model Averaging

Although Non-homogeneous Regression utilises the spread of the ensemble members by including it as a predictor for the forecast variance, the trajectories of the ensemble members often separate into distinct groups, each exploring a different region of the model's phase space. In this case, specifying a unimodal forecast distribution centred at the mean of these members, as in Non-homogeneous Regression, would not replicate the information present in the ensemble. Instead, a multi-modal distribution would be more appropriate (Wilks, 2002).

Multi-modal distributions can easily be achieved using Ensemble Dressing techniques, so-called because they dress each ensemble member with its own error distribution, akin to kernel density estimation methods (Wilks, 2019). This was first proposed by Roulston and Smith (2003), who use past errors of the 'best' ensemble member, defined as that which has previously exhibited the smallest forecast error, to perturb the original ensemble members, creating a significantly larger ensemble. Wang and Bishop (2005) ameliorated the approach by using smooth distributions centred around each (bias corrected) ensemble member. The final forecast thus takes the form of a mixture distribution - an equally-weighted combination of component distributions associated with each ensemble member - rather than a finite ensemble. However, due to restrictions placed on the variance of the component distributions, this approach can only increase the ensemble spread, and hence cannot be used to calibrate overdispersed forecasts.

A more general variant of this is Bayesian Model Averaging (BMA; Raftery et al., 2005), a method used extensively in economics and the social sciences prior to its introduction in a weather forecasting context (Raftery et al., 1997). BMA similarly yields predictive distributions corresponding to each ensemble member, but unlike the approach of Wang and Bishop (2005), no constraint is put on the variance parameter; instead, it is included as a coefficient to be estimated. Assuming Gaussian component distributions, BMA becomes

$$Y|\boldsymbol{x} \sim \sum_{m=1}^{M} w_m N(\alpha_m + \beta_m x_m, \sigma_m^2). \tag{5}$$

In this case, separate weights w_m and coefficients can be estimated for each ensemble member, though if groups of exchangeable members exist, then the weights and parameters assigned to members of the same group should be equal (Fraley et al., 2010). BMA is thus a more general Ensemble Dressing approach. Moreover, since it can issue distributions that are multi-modal, Equation 5 is also more flexible than Non-homogeneous Gaussian Regression (Equation 4). However, this flexibility comes at the expense of parsimony, and hence estimating the coefficients of this mixture distribution is typically more cumbersome than for unimodal forecast distributions.

2.3.2.3 Choice of distribution

Thus far, only Gaussian predictive distributions have been considered. Of course, MOS is a linear regression model, and hence can easily be extended to utilise alternative distributions via the generalised linear modelling framework (McCullagh, 2018). One example of a generalised linear modelling approach commonly used to statistically post-process weather forecasts is logistic regression (Lemcke and Kruizinga, 1988; Hamill et al., 2004). Rather than estimating an entire distribution for the response variable, logistic regression issues probabilistic forecasts for binary outcomes. Successive applications of logistic regression could be used to forecast the probability that the response variable falls below a sequence of thresholds, thus approximating a complete predictive distribution. However, since this requires fitting several independent logistic regression models, the forecast probabilities are not constrained to increase as the threshold increases, potentially leading to illogical forecast distributions.

Wilks (2009) therefore proposes an extension whereby an increasing function of the threshold is included as a predictor in the regression equation. The resulting distribution depends smoothly on the chosen thresholds, thus alleviating any inconsistencies, while also allowing the estimation of a complete distribution using just one regression equation. Messner et al. (2014) show that when the increasing function of the threshold is chosen to be the threshold itself, the extended logistic regression framework proposed by Wilks (2009) is a particular example of issuing a forecast for the response variable in the form of a logistic distribution. Messner et al. (2014) extend this further to include the mean and spread of the ensemble as predictors for the location and scale of the logistic predictive distribution, and, in doing so, show that the approach aligns with the Non-homogeneous Regression (NR) framework:

$$Y|\boldsymbol{x} \sim L(\alpha + \beta\bar{x}, \gamma + \delta s_x), \tag{6}$$

where the logistic distribution has been employed in place of the Gaussian distribution in Equation 4, with $L(\mu, \sigma)$ denoting the logistic distribution with location μ and scale σ.

This is an example of the flexibility of NR, and the ease with which it can be applied to forecasts of different weather variables through the choice of an appropriate parametric distribution. BMA, similarly, can be employed with alternative choices for the component distributions. These distributions represent the conditional distribution of the observations given the ensemble output, and considerable research has been devoted to identifying statistical distributions suitable for post-processing a range of weather variables. This distribution is typically chosen to be similar in form to the marginal, or unconditional distribution of the variable being forecast, though this need not be the case. The response variable could exhibit a complicated structure, but if the forecast reproduces this behaviour, then the conditional distribution might be reasonably modelled using simpler families of distributions. However, in practice, it is convenient to assume a forecast distribution that is similar in shape to the variable's marginal distribution. In doing so, non-zero probabilities are assigned only to outcomes that can occur, and the forecast tends towards the predictand's climatology as the information in the ensemble decreases.

Representing temperature observations using Gaussian distributions, for example, is a commonplace assumption in the meteorological community (Von Storch and Zwiers, 2001). As such, a Gaussian distribution is typically employed in NR and BMA, and extensions thereof, to post-process temperature forecasts. This approach has, however, recently come under scrutiny (Gebetsberger et al., 2018, 2019, see Chapter 6). Wind speed, on the other hand, is non-negative and methods to post-process wind speed forecasts thus require predictive distributions with non-negative support. Gamma distributions have been proposed in both the BMA (Sloughter et al., 2010) and NR (Scheuerer and Möller, 2015) frameworks for this purpose, whereas alternative choices include distributions truncated below at zero (Thorarinsdottir and Gneiting, 2010; Baran, 2014; Messner et al., 2014).

The choice of parametric distribution becomes particularly difficult when forecasting precipitation accumulation, which follows a distribution with positive support when it exceeds zero, but may also be exactly zero. The problem can be decomposed into two stages: first estimating the probability of observing any precipitation - using, for example, logistic regression (Sloughter et al., 2007) or probability mapping-based approaches (Flowerdew, 2014) - and then fitting a positive predictive distribution only to non-zero precipitation events. Though, a more unified approach involves censoring distributions at zero, with any mass assigned to negative values instead attributed to zero precipitation (Messner et al., 2014; Scheuerer, 2014; Scheuerer and Hamill, 2015a).

These three weather variables have been studied in most detail, and forecasts of other weather variables can generally be post-processed using similar techniques. One notable exception is cloud cover, which is measured on a discrete scale and thus requires alternative approaches (Hemri et al., 2016). Regardless, the choice among distributions will depend on the time period, spatial region, and weather model under consideration, with no distribution invariably being more appropriate than others. As is the case in the wider field of statistical modelling, diagnostic model checking techniques should be employed (see Section 2.4) to verify that the post-processing model is appropriate for the present scenario.

Furthermore, to ensure that the data being considered align with the assumptions made by a particular distribution, the forecasts and observations could be transformed prior to post-processing. Wilks (2002), for example, models square-root transformed wind speeds using a Gaussian distribution, while Hemri et al. (2015) employ the more general Box-Cox transformation (Box and Cox, 1964) for rainfall runoff. In doing so, these transformed variables can be post-processed using more familiar approaches (Wilks, 2019). Scheuerer and Hamill (2015a), on the other hand, argue that transforming variables distorts the scale of the variables, and could thus remove some of the information contained within the raw ensemble forecasts.

Alternatively, it is possible to combine information from several possible parametric distributions, yielding a more flexible forecast. Lerch and Thorarinsdottir (2013), for example, propose a combination of distributions to post-process wind speed forecasts, where the choice of distribution is determined by the behaviour of the ensemble. In particular, if the ensemble median exceeds a predefined threshold then the generalised extreme value (GEV) distribution is applied, with a truncated normal distribution implemented otherwise. Baran and Lerch (2015) later modify this approach to use a log-normal distribution in place of the GEV distribution, in the hope of increasing computational efficiency. Such an approach assumes that the log-normal or GEV distribution is better suited to capture the behaviour in the upper tail of the predictive distribution, which is often poorly represented when post-processing (Friederichs et al., 2018). A similar method is employed by Gebetsberger et al. (2017) to recalibrate precipitation forecasts, applying a separate post-processing model depending on the fraction of ensemble members that predict zero precipitation.

The resulting forecast distribution is thus a mixture of two component distributions. These combination approaches depend, however, on a suitable criterion to decide when to switch between distributions. Instead, Baran and Lerch (2016) extend the combination approach to include a weight parameter that is estimated from the data, which provides a data-driven way to choose how much emphasis to put on each distribution. However, Baran and Lerch (2016) find estimating the weight and the coefficients for both predictive distributions simultaneously to be computationally expensive and at times numerically unstable. Both distributions could instead be estimated separately, and the optimal weight estimated conditionally on these two distributions in a second step, though Gneiting et al. (2013) prove that the linear combination of two calibrated forecast distributions is necessarily uncalibrated. The authors go on to provide approaches that can combine the distributions in such a way that the resulting mixture model is also calibrated, and these methods have been applied in a post-processing context in Baran and Lerch (2018).

In the combination approach of Lerch and Thorarinsdottir (2013) the forecast distribution can be expressed as a mixture model, where the weight is an indicator function that takes the value one when the median of the ensemble members exceeds the predefined threshold, and zero otherwise. The component distributions of the mixture model in this case are then estimated conditionally on the weight function, which thus directs the post-processing model towards particular regions of the historical data on which to estimate the different component distributions. The choice of this weight is therefore of considerable interest, and Chapters 3-5 of this thesis similarly consider using a weight function to direct a mixture model-based post-processing method towards particular instances in which the performance of the prediction system is expected to vary. In doing so, the component distributions of the mixture model can each separately address different error characteristics of the ensemble forecast.

2.3.2.4 Input variables and link functions

Previous applications of MOS to deterministic weather forecasts often utilised several predictor variables (Glahn and Lowry, 1972), whereas established ensemble post-processing methods tend to use as predictors only information provided by the ensemble member forecasts for the weather variable under consideration: Non-homogeneous Regression and Bayesian Model Averaging as presented above are common examples of this. Although modern ensemble forecasts should themselves be more informative than previous weather forecasts, forecasters have recently realised once again the benefit to be gained by including alternative sources of information into the statistical models.

Messner et al. (2017), for example, propose a boosting algorithm to select relevant predictor variables for Non-homogeneous Regression models. Letting $\boldsymbol{u}$ and $\boldsymbol{v}$ denote (row) vectors of possible predictor variables, Equation 4 can be generalised to

$$Y|\boldsymbol{u}, \boldsymbol{v} \sim N(\boldsymbol{u}\boldsymbol{\beta}, \boldsymbol{v}\boldsymbol{\gamma}), \tag{7}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are coefficient (column) vectors corresponding to $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Non-homogeneous Gaussian Regression is then a particular example of Equation 7 where $\boldsymbol{u} = (1, \bar{x})$ and $\boldsymbol{v} = (1, s_x^2)$. To select the most relevant predictors to use in Equation 7, the boosting algorithm implemented by Messner et al. (2017) iteratively updates the coefficient estimates in such a way that only the most useful variables are assigned non-zero coefficients. In doing so, the approach removes any unnecessary predictors, while automatically selecting those most relevant for use within the post-processing model.

Equation 7 can be generalised further to use potentially nonlinear functions, g_μ and g_σ, of the predictors to estimate the mean and spread of the predictive distribution:

$$Y|\boldsymbol{u}, \boldsymbol{v} \sim N(g_\mu(\boldsymbol{u}), g_\sigma(\boldsymbol{v})). \tag{8}$$

There are several ways of constructing such nonlinear functions, though arguably the most familiar approach in the field of statistical modelling is to use linear combinations of nonlinear basis functions; splines, for example, are commonly used within the generalised additive modelling (GAM) framework (Hastie and Tibshirani, 1990). Compared with the wider field of statistical modelling, the use of splines as a tool for forecast recalibration is relatively sparse, though they have been used in the context of post-processing to construct smooth functions across time and space (Dabernig et al., 2017; Gebetsberger et al., 2019; Lang et al., 2019a), to incorporate information regarding circular weather variables (Eide et al., 2017; Lang et al., 2019b), and to extend quantile regression (Bremnes, 2019).

Alternatively, Rasp and Lerch (2018) employ a neural network to model the nonlinear functions for the location and scale of the predictive distribution in Equation 8. Neural networks are powerful machine learning tools that also allow additional predictors to be seamlessly integrated into post-processing models. Although not a recent concept, they have become progressively more accessible with advances in computational capabilities, and, as a result, have been applied recently to several problems in meteorology (McGovern et al., 2017). Rasp and Lerch (2018) implement a neural network with only one hidden layer, and hence the mean and standard deviation of the forecast distribution in this case are obtained via a bias corrected linear combination of the input variables, followed by the application of a nonlinear activation function.

However, the neural network proposed by Rasp and Lerch (2018) still assumes that the conditional distribution of the response variable belongs to a known statistical family. To allow for more flexibility when constructing the predictive distribution, non- or semi-parametric post-processing methods exist that make fewer distributional assumptions, and are thus applicable when forecasting a variety of weather variables. Taillardat et al. (2016), for example, propose post-processing weather forecasts using quantile regression forests. Quantile regression forests are comprised of a series of consecutive decision trees, each of which considers binary events associated with thresholds of possible predictor variables. Given the current prediction, the forecaster can evaluate each of the decision trees, and identify previous forecasts that have taken a similar path through the forest, thereby exhibiting similar characteristics to the current forecast. The observations corresponding to these previous forecasts then constitute an ensemble of possible outcomes that could occur given the current forecast (Taillardat et al., 2016).

Although the use of statistical learning techniques (Hastie et al., 2009), such as quantile regression forests and neural networks, within post-processing is still in its infancy, considerable work is currently ongoing in this area (Taillardat and Mestre, 2020; Vannitsem et al., 2020). Such methods rely heavily on the available data to estimate relationships between predictor variables, and this is yet more pertinent for non-parametric approaches that make no distributional assumptions. The nature of the archive of available meteorological data is thus of considerable importance when post-processing.

2.3.3 Training post-processing methods

2.3.3.1 Training data

Post-processing methods learn relationships between the weather model and the atmosphere through an archive of historical forecasts and observations, commonly referred to as the training, or learning data set. Ideally, the training data should reflect the present situation, in that the current forecast is expected to exhibit similar error characteristics to those in the training data. As such, removing these historical biases from the current forecast should enhance its predictive performance. Choosing an inappropriate data set, on the other hand, may result in the post-processing method addressing biases that do not materialise in the forecast, potentially hindering its performance relative to that of the raw ensemble.

In practice, selecting a suitable training data set is not straightforward. It is impossible to measure exactly the instantaneous atmospheric conditions, and there is thus a decision to be made regarding how to define the ‘observations’ when training post-processing methods. The most accurate estimate of the prevailing weather is provided by recordings from suitable sensors, or measuring devices, positioned in an irregular network. However, the global coverage afforded by this network is generally fairly sparse, with particularly few recording stations situated in regions with complex terrain and over seas (Hamill, 2018). Post-processing methods that are trained using weather recordings at synoptic stations only address the systematic forecast errors present at these locations, and hence, if these stations are not representative of the entire model domain, then biases specific to other locations will remain in the forecast, meaning an entire calibrated forecast field is unobtainable.

Instead, since the numerical weather models produce forecasts at specific grid points over the model domain, it would be desirable to have observations similarly available on a grid. Although this is generally not the case for empirical weather measurements, an alternative option is to train the post-processing methods using the analyses from the weather model at the time for which the forecast is being made (the forecast validation time). As described in Section 2.1.2, the model analysis represents the best guess of the atmospheric state, constructed by assimilating a short-range forecast with large volumes of data recordings. Errors in the data assimilation, however, mean the analysis itself exhibits potentially large errors, and, as a result, using analyses to train post-processing methods tends to underestimate the uncertainty present in reality (Feldmann et al., 2019). On the other hand, model analyses provide appropriate spatial coverage, obviating some of the shortcomings of post-processing with weather measurements. Therefore, the choice of observations to utilise when post-processing depends on the problem at hand: model analyses are often used to post-process predictions when a calibrated forecast field is desired, while weather measurements are preferred for more localised bias corrections.

Regardless of how the observations are defined, statistical post-processing seeks to remove systematic deviations between these observations and the forecasts. Given all available historical forecasts and observations, it is still necessary to select the most relevant subset of this data on which to train today's post-processing model. The magnitude of forecast errors may depend on a variety of factors, such as the time of the year, spatial location, or the forecast values themselves, and to allow the post-processing model to account for such dependencies, the training data should vary accordingly. For example, to account for the recent behaviour of model errors, a rolling or sliding window, comprised only of the most recently available forecast-observation pairs, is often used operationally to train post-processing models (Gneiting et al., 2005). Longer windows provide larger data sets on which to estimate post-processing coefficients, whereas smaller windows are more tailored to recent model behaviour.

Alternatively, post-processing models could be trained using a fixed window. Such a window does not adapt continually in light of more recent data, alleviating the need to continually update post-processing coefficients. Hence, one set of coefficients can be estimated and applied repeatedly to different forecasts. Fixed windows are less suited to accounting for potential seasonality in model errors, but generally provide much larger amounts of data than time-adaptive windows. Therefore, to account for seasonality whilst using a fixed window, smooth functions of the time of the year can be included directly in the post-processing model (Hemri et al., 2014; Messner et al., 2017; Dabernig et al., 2017; Lang et al., 2019a).

Larger data sets on which to estimate post-processing coefficients also reduce the chances of overfitting the training data (Wilks, 2019). Overfitting refers to when the statistical model captures not only the signal between the observations and forecasts, but also the noise specific to the training data. Post-processing methods that overfit the training data therefore perform well when evaluated using this training data, but this forecast quality does not translate to out-of-sample predictions, where the method is poorly equipped to adapt to the new data. The amount of data used to train statistical post-processing models is thus an inherent compromise between obtaining enough data on which to reliably estimate model coefficients, and using little enough so as to incorporate only the most relevant data for the current forecast.

This is exacerbated further by the fact that operational weather models are continually updated, leading to changes in their error characteristics. Although these updates generally improve model performance, they come at the expense of robust statistical post-processing: after introducing a new model version, properties of the resulting forecasts may not yet be well understood, and it thus becomes difficult to identify previous forecasts that are believed to behave similarly to those from the updated model version. In order to exploit characteristics of the updated model, it is necessary to wait for an extended period of time before sufficient data is collected from which these characteristics can be deduced. In the meantime, a choice must be made between foregoing the benefits of post-processing, and generating hindcasts from the new prediction system to understand its performance. Alternative approaches that can adapt to updates in the model when post-processing are in short supply, and further work on this subject is thus required (Demaeyer and Vannitsem, 2020).

As well as displaying pronounced seasonal dependencies, forecast errors may also vary considerably in space. Pooling information regarding all stations into the training data, referred to as ‘global’ post-processing, might therefore not account for locally-varying biases. Conversely, training data sets could be designed separately for every location of interest, each comprised of only forecast-observation pairs obtained at that location. However, fitting a separate post-process method at every location under consideration is not only computationally expensive, particularly when considering large networks of stations, but it also reduces the amount of data available on which to train the statistical models, increasing the chances of overfitting. Moreover, such an approach is at times redundant. Although the biases should not be assumed stationary in space, it is reasonable to believe that the errors at neighbouring locations will often exhibit similar characteristics (Scheuerer and Hamill, 2015a). Proximity in space is not the only factor that should be accounted for. Hamill et al. (2017) group together synoptic stations in the US according to several properties that were believed to influence biases in precipitation forecasts, while Lerch and Baran (2017) consider a more empirical approach based on the past behaviour of both the observations and forecasts at each location. In this sense, the current forecast is trained using previous forecasts at other locations that are believed to exhibit similar qualities.

The quality of a post-processing method depends heavily on the data on which it is trained. Although calibrated forecasts could be achieved without considering particular times of the year or spatial locations separately, incorporating these dependencies allows for a more targeted, and hence informative, forecast. This highlights how information can be added to post-processing methods, not only by including additional input variables when recalibrating (as in the previous section), but also by incorporating these dependencies when constructing the training data. In fact, some post-processing approaches do not issue forecasts in the form of parametric distributions, but rely on the training data to such an extent that the observed values in the training data are themselves treated as an ensemble forecast for the present scenario. The training data in this case must be chosen carefully, using 'analogues' of today's forecast.

2.3.3.2 Analogue-based post-processing

Forecast analogues refer to previous instances where the state of the atmosphere was comparable (ideally analogous) to its currently predicted state. The observations that occurred in these instances are then assumed to be representative of the possible outcomes that could materialise given the current forecast. In turn, the distribution of these historical observations constitutes an empirical estimate of the conditional distribution of the response variable given today's prediction. Issuing as a forecast this ensemble of previous observations can thus be thought of as an analogue-based post-processing method (Hamill and Whitaker, 2006). The use of analogues as a tool to recalibrate forecasts dates back to Van den Dool (1989), though similar approaches were used as a general forecasting technique prior to the advent of numerical weather prediction (Lynch, 2008). Since then, analogues have also become common in other branches of post-processing, most notably statistical downscaling (Zorita and Von Storch, 1999).

Analogue-based methods are non-parametric and can thus be applied to a wide range of weather variables (perhaps simultaneously) without any distributional assumptions.

Moreover, this flexibility means they have previously been used to post-process both deterministic predictions (Delle Monache et al., 2011, 2013) as well as forecasts in the form of an ensemble (Hamill and Whitaker, 2006). Clearly, however, the approach is reliant on a suitable method of defining forecast analogues. Typically, this is achieved using some measure of the distance from previous forecasts to the current prediction, the exact form of which will depend on the type of forecast to be post-processed, and also the number of variables under consideration. Historical forecasts that yield a similarity measure that falls below a predefined threshold are then defined as analogues to today's forecast, where the choice of threshold represents a trade-off between larger ensembles and more targeted forecasts. In this sense, the quantile regression forest post-processing approach proposed by Taillardat et al. (2016) (see Section 2.3.2.4) can also be viewed as an analogue-based approach, where the forecast analogues are defined using a series of decision trees.

These approaches are, however, limited by the range of observations that have previously materialised, and they thus depend on a suitably large archive of data to capture the occurrence of more extreme weather events. Instead, methods have been proposed that extend analogue-based post-processing methods to yield forecasts in the form of a predictive distribution, either by implementing parametric post-processing methods on the training data defined by the analogues, or by applying kernel density estimation to the resulting observations (Junk et al., 2015; Barnes et al., 2019; Taillardat et al., 2019).

Such methods address shortcomings in the analogue-based approach owing to the finite number of available observations. Alternatively, effort could be made to augment the set of historical forecasts and observations from which to define analogues. This led to the idea of retrospective forecasting (reforecasting; Hamill et al., 2004). Reforecasts are predictions of past events (hindcasts) made by initialising the presently operational weather model at historical analyses, for the purpose of learning the error characteristics of the current weather model. To ensure the prediction system is as similar as possible to that currently in place, the model analyses should similarly be re-computed using the current data assimilation scheme (Kalnay et al., 1996; Compo et al., 2011; Dee et al., 2011), prior to generating the hindcasts. Hamill et al. (2004) demonstrate that using newly found computational resources to generate reforecasts can be more beneficial than increasing the resolution of the operational weather model, since statistical post-processing methods can use these larger archives of data to provide a more reliable estimate of the future weather.

2.3.3.3 Parameter estimation

Having procured a suitable set of training data, one which is streamlined enough to contain previous forecasts and observations particularly relevant for the current situation but also large enough to avoid overfitting, relationships between the numerical weather model and the atmosphere can be learned. For the parametric methods described herein, this requires the estimation of suitable post-processing parameters. There are several ways to obtain such estimates, but due to the complexity of the post-processing models, this typically involves using numerical optimisation routines to minimise a specified loss function over the training data. That is, estimating the parameter vector $\boldsymbol{\theta}$ that minimises a loss, or penalty function $l(\boldsymbol{\theta}; F, y)$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i} l(\boldsymbol{\theta}; F_i, y_i), \tag{9}$$

where i is an index over all forecasts F and observations y in the training data. This process is generally referred to as M estimation. For statistical post-processing, Gneiting and Raftery (2007) propose the optimum score estimation framework, which belongs to the more general class of M estimation, but it restricts the loss function l in Equation 9 to be a proper scoring rule. Scoring rules are functions designed to assess the quality of forecasts, and proper scoring rules exhibit certain desirable characteristics that are discussed in detail in Section 2.4. Hence, the premise behind optimum score estimation is that the parameters to be used in post-processing models should be those that produce the most accurate forecasts over the training data, as measured using a proper scoring rule.

The choice of scoring rule to employ in Equation 9 is not trivial, though for forecasts in the form of a parametric predictive distribution, two functions are preferred in the literature. The first is the logarithmic score (Good, 1952), which is equivalent to the negative of the logarithm of the likelihood function. As a result, minimising the logarithmic score is equivalent to maximum likelihood estimation. Since likelihood-based estimation techniques are abundant in the field of statistics, the resulting parameter estimates are well understood, and known to exhibit appealing properties. Moreover, in some cases, including linear regression models, maximum likelihood estimates are available analytically, obviating the need for numerical optimisation methods. In other cases, algorithms have been fine-tuned to efficiently obtain maximum likelihood estimates in potentially complicated situations, since they are required so frequently in practice. One example of this is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), which is often used to estimate the post-processing coefficients in BMA (Raftery et al., 2005).

The second preferred scoring rule is the continuous ranked probability score (CRPS), introduced in Matheson and Winkler (1976) as a measure to evaluate forecasts in the form of continuous distributions, but first used as a loss function for post-processing in Gneiting et al. (2005). Since then, the CRPS has become commonplace in weather forecasting studies, both in parameter estimation and as a tool for verification. The CRPS differs from the logarithmic score in that it depends on the entire forecast distribution, whereas the logarithmic score is 'local', and hence relies only on the forecast probability of the weather event that materialises. As a result, Gneiting et al. (2005) argue that the logarithmic score is more sensitive to outliers than the CRPS, and can hence lead to overdispersed forecasts.

The logarithmic score and the CRPS are defined mathematically in Section 2.4.3. Regardless of which score is chosen, optimising coefficients through the minimisation of some loss function does not take into account any uncertainty in the parameter estimation. However, since the training data sets used to post-process weather forecasts are generally fairly small, this parameter uncertainty should not be neglected. In the case of linear regression models, a Student's t-distribution could be used in place of the normal distribution to account for the additional uncertainty induced by the parameter estimation (Glahn et al., 2009; Siegert et al., 2016a), though for more complicated post-processing methods, such an analytical extension may not exist. Instead, Siegert et al. (2016a) introduce a way of accounting for the uncertainty in parameter estimation via a bootstrap resampling technique.

Alternatively, Bayesian post-processing methods have also been proposed to explicitly model this parameter uncertainty. For this purpose, a prior distribution must be specified for the coefficients, which is evolved using Bayes theorem in light of more recent observations. The resulting posterior distribution can then be utilised to generate a forecast distribution for the response variable, often in the form of a large sample obtained via Markov Chain Monte Carlo (MCMC) techniques (Krüger et al., 2016). Although several Bayesian post-processing methods exist (e.g. Stephenson et al., 2005; Di Narzo and Cocchi, 2010; Siegert et al., 2016b; Barnes et al., 2019), plug-in approaches have received considerably more attention. As such, parameter uncertainty in post-processing methods is often overlooked.

2.3.4 Future research

Univariate statistical post-processing methods are currently well-adapted to recalibrate the output from numerical weather models, whether deterministic or an ensemble, without losing the information provided by the dynamical model. As a result, it is possible to obtain reliable forecasts of several weather variables at individual points in time and space. However, as numerical weather prediction models become more accurate, operating at higher model resolutions and incorporating more sophisticated stochastic physics, they begin to account for several sources of uncertainty that were not represented by previous model versions, reflecting a better capacity to predict the weather purely through the dynamical models. If this continues, then, just as the numerical weather models are evolving, so too should statistical post-processing methods.

The statistical learning algorithms discussed in Section 2.3.2.4 provide a promising, data-driven approach to identify more nuanced structures in the forecast errors, objectively selecting from a large set of possible predictors those most relevant for use within post-processing. Thus far, work on such approaches indicates that the additional information provided by these input variables can enhance forecast performance, and, as a result, considerable research is now ongoing to develop, trial and compare new statistical learning methods within post-processing (Taillardat and Mestre, 2020; Vannitsem et al., 2020).

However, even without the additional flexibility provided by these more data-driven approaches, post-processing methods have been shown to consistently offer vast improvements upon the raw ensemble forecast. Yet, despite the obvious need for post-processing, and the abundance of methods proposed in the literature (see e.g. Vannitsem et al., 2018), their inclusion within operational forecasting suites is not yet widespread (Gneiting and Katzfuss, 2014; Taillardat and Mestre, 2020). One possible reason for this is the inability of parametric post-processing methods to capture the various dependencies that manifest in the atmosphere. The post-processing methods described herein are all univariate: they act exclusively on a specific variable, time, lead time, and often location. Of course, however, the atmosphere is a multifaceted phenomenon that exhibits spatial, temporal and inter-variable dependencies. Therefore, although the use of statistical learning methods to incorporate additional information into the forecast is a current direction in the field of post-processing, an equally pressing, but not unrelated, problem is that of capturing the various dependencies present in the atmosphere. The problem arises because post-processing variables at specific times and locations separately using parametric methods, such as BMA and NR, does not necessarily reproduce the correlation structure present in the atmosphere, making a calibrated forecast field, as is commonly desired in practice, difficult to attain. There is thus the need to extend these developed and well-understood univariate techniques to higher dimensions.

Multivariate distributions could be used within a similar parametric post-processing framework to that discussed for the univariate setting. Work on the simultaneous calibration of several different weather variables is comparatively under-developed, due perhaps to the lack of development of verification methods with which to assess and compare multivariate forecasts (Gneiting et al., 2008; Thorarinsdottir and Schuhen, 2018). As such, parametric post-processing of multiple variables is limited primarily to the joint forecasting of temperature and wind speed (Baran and Möller, 2017), and the recalibration of bivariate wind velocity forecasts (Pinson, 2012; Schuhen et al., 2012; Lang et al., 2019b). However, there are fewer multivariate parametric distributions to choose from when post-processing, while estimating the coefficients for such distributions in high dimensions can become numerically unstable.

Instead, parametric multivariate forecast distributions can be achieved using copulas. Copulas allow multivariate distributions to be expressed as a combination of several marginal distributions, that are subject to a specified dependence structure. Copula-based post-processing methods therefore permit the use of univariate approaches to be applied along each margin (i.e. separately for each variable, lead time and location), before imposing a certain dependence template on the post-processed output. The crux of the copula is then to specify a suitable dependence structure that adequately reflects the relationships between the various margins. A Gaussian copula is commonly used for this purpose, which, if the marginal forecast distributions are all themselves Gaussian, is equivalent to issuing a multivariate normal predictive distribution (Schefzik and Möller, 2018). As such, the Gaussian copula has been applied to capture dependencies between different forecast lead times (Pinson and Girard, 2012; Hemri et al., 2015), meteorological variables (Schuhen et al., 2012; Pinson, 2012; Baran and Möller, 2017; Lang et al., 2019b), and spatial locations (Feldmann et al., 2015) when post-processing.

However, the dependencies in the atmosphere are typically more complicated than those that can be described by simple parametric copulas, and, for this reason, it is often preferred to generate a forecast in the form of an ensemble, rather than an entire predictive distribution (Pinson, 2012). A calibrated ensemble forecast could be obtained by sampling from the predictive distribution issued by parametric post-processing methods, typically using evenly spaced quantiles of the distribution (Bröcker, 2012b). Although the resulting ensemble forecast will still not reflect the dependencies in the atmosphere, an empirical copula can be applied to reorder the ensemble members according to a dependence template, in order to reproduce the correlation structure observed in reality. The approach is equivalent to the parametric copula-based framework described above, but empirical copulas obtain a dependency structure from a finite sample of possible weather scenarios, and are thus considerably more flexible than parametric alternatives. The two most prominent approaches are the Schaake Shuffle (Clark et al., 2004) and Ensemble Copula Coupling (ECC; Schefzik et al., 2013; Schefzik, 2016), which differ only in the choice of data from which to assume the structural dependencies arise: the Schaake Shuffle uses historical weather observations, while ECC exploits the raw ensemble itself. These empirical copula approaches are commonly used in operation, allowing the use of well-tested univariate post-processing techniques whilst retaining the structural dependencies present in either previous atmospheric observations or the raw ensembles.

Alternatively, post-processing methods exist that are designed to directly produce forecasts in the form of calibrated ensembles, rather than entire predictive distributions (Van Schaeybroeck and Vannitsem, 2015; Williams, 2016). As Wilks (2018) remarks, such methods typically yield ensemble forecasts $\boldsymbol{x}^* = (x_1^*, ..., x_M^*)$ whose members are expressed as a linear bias correction of the ensemble mean, and a stretching of the distance between the original ensemble member and the ensemble mean:

$$x_m^* = (\alpha + \beta \bar{x}) + \gamma(x_m - \bar{x}). \tag{10}$$

These 'member-by-member' post-processing approach thus address simultaneously both the bias and dispersion errors in the raw ensemble, without imposing any distributional assumptions. Moreover, they do not alter the ordering of the original ensemble members, and therefore retain the dependence structure in the raw ensemble forecast, without the need for a subsequent reordering. Although this means such an approach is unable to correct errors in the ensemble's dependence structure, the multivariate trajectory provided by a particular ensemble member is a realistic simulation of the entire atmosphere.

Both empirical copula-based approaches and member-by-member post-processing methods are powerful tools for the generation of calibrated, multivariate forecast fields that capture the dependencies in the atmosphere (Lerch et al., 2020). However, methods that permit forecasts in the form of smoother predictive distributions while accounting for correlations between the predictive distributions along each margin would be yet more effective at representing the uncertainty inherent in the forecast scenario. In particular, smooth predictive distributions provide considerably more information regarding the tails of the forecast distribution, especially in comparison with a relatively small ensemble.

Moreover, high-impact weather is often a product of compound weather events, referring to the simultaneous occurrence of extreme values of multiple weather variables, and hence forecasts in the form of predictive distributions that can capture the dependence structure between several times, spatial locations, and meteorological variables will provide a more suitable foundation on which to base the prediction of extreme weather events. That being said, further work is yet required to devise post-processing methods that accurately represent the tails of the conditional distribution of the response variable, even in the univariate case. For example, due perhaps to small amounts of training data in operation, meaning less exposure to rare events, the predictive distributions issued by NR and BMA tend to be inadequate when it comes to forecasting extreme weather (Friederichs, 2010; Friederichs and Thorarinsdottir, 2012). Pantillon et al. (2018) show that post-processing can even hinder the performance of more extreme events, since the error characteristics in the training window are often inappropriate for forecasts of such events. Methods to attain better forecasts of more extreme weather events are few and far between. Friederichs (2010) and Lerch and Thorarinsdottir (2013) utilise results from extreme-value theory to provide a better representation of the tails of the predictive distribution, though further work in this area is certainly required, especially given the high-impact nature of these events.

2.4 Forecast verification

2.4.1 Introduction

Although forecasts in the form of ensembles or predictive distributions acknowledge the uncertainty in their prediction, they can still be misleading if this uncertainty is specified incorrectly. Therefore, just as important as issuing forecasts is understanding how they perform. In doing so, forecasters gain a greater awareness of the strengths and limitations of their predictions, and, in turn, learn how they can be improved (Jolliffe and Stephenson, 2012). A framework to achieve this is the goal of forecast verification.

There are several components to consider when evaluating forecasts, but, intuitively, a good forecaster is one whose predictions consistently agree with the event that materialises (Murphy, 1993). This is true not only for weather forecasts, but also for predictions made in other settings. For example, if a forecast predicts the occurrence of an event with probability p, then this event should occur with probability p when such a forecast is issued. A forecaster who achieves this is said to be (auto-)calibrated, or reliable. Calibration is thus a joint property of the forecasts and the observations, which asserts that the conditional distribution of the observations given the forecast must be equal to the forecast itself. Assessing forecasters via their calibration therefore aligns with the 'distributions-oriented' approach to forecast verification proposed in Murphy and Winkler (1987).

Murphy and Winkler (1987) claim that all information of interest when assessing forecasts is provided by the joint distribution of the forecasts and observations, and its factorisations into conditional and marginal components. These distributions describe the statistical characteristics of the forecasts, the observations, and the relationship between them, and they thus provide a logical platform on which to base forecast evaluation. Moreover, the distributions-oriented framework permits several aspects of the forecast performance to be assessed, which is necessary in order to fully realise the practical use of the forecast (Stephenson and Doblas-Reyes, 2000).

This practical utility defines the 'value' of the forecast to its user, and, in this sense, the value of a forecast will depend on how it is to be used. Therefore, it is desirable for a forecaster (and the forecast user) to know in what situations their predictions do, and do not, perform well. To do so, properties of the joint distribution are typically inspected graphically. Graphical verification tools aid the visualisation of forecast performance, rendering it easy to identify deficiencies in the prediction system. These visual representations of forecast qualities present an overview of the strengths and limitations of the forecaster, but they do not provide an overall, objective measure of forecast performance. To this end, they are suited to characterising the performance of a particular forecast system, but they alone are unable to provide an objective comparison of competing prediction schemes.

To complement graphical displays for this purpose, various functions exist to objectively measure the agreement between forecasts and observations. These functions, known as scoring rules, take a set of historical forecasts and observations and output a numerical measure of the forecaster's accuracy. In doing so, different forecast systems can be objectively ranked and compared. Several scoring rules, or functions have thus been derived to evaluate deterministic, ensemble and probabilistic forecasts. Just as the forecasts must be calibrated in order to be trustworthy, it is widely accepted that these scores must exhibit desirable properties, most notably propriety (Gneiting and Raftery, 2007). Proper scoring rules take into account not only whether or not the forecaster is calibrated, but also their ability to distinguish between instances in which different observations will occur. This latter quality refers to the information content in the forecast, often called the forecast resolution.

Despite their widespread use, scoring rules do not directly relate to the distributions-oriented framework for forecast verification. However, decompositions exist that divide such scores into components that each assess a different aspect of the forecast, such as its resolution and reliability (Murphy, 1973b; Bröcker, 2009), with the terms typically related to different factorisations of the joint distribution of the forecasts and observations (Mitchell, 2020). These decompositions, discussed in detail in Chapter 7 of this thesis, demonstrate that the single numerical measure provided by the scoring rule implicitly incorporates the behaviour of the joint distribution of the forecasts and observations (Murphy and Winkler, 1987).

Furthermore, the general goal of probabilistic forecasting is to maximise the information in the forecast, while ensuring a reliable prediction. This is commonly expressed as issuing predictive distributions that are sharp, subject to being calibrated (Gneiting et al., 2007). Both graphical displays and proper scoring rules can help to understand to what extent these criteria are satisfied, and these two approaches are discussed further in the proceeding section. As in Section 2.3, however, we focus here on the verification tools to be employed throughout the following chapters of this thesis: considerably more extensive reviews of forecast verification are available in Jolliffe and Stephenson (2012) and Thorarinsdottir and Schuhen (2018).

2.4.2 Calibration

As has been discussed, a forecaster must be calibrated in order for their predictions to be considered useful, and establishing whether or not a forecaster is calibrated should therefore be a priority when evaluating predictive performance. There are several ways of analysing a forecaster's calibration, many of which employ graphical techniques. Reliability diagrams, for example, are the dominant technique for assessing the calibration of probabilistic forecasts of binary events (also called probability forecasts). Reliability diagrams plot the forecast probability of an event against the conditional frequency of the event occurring given the forecast, and a forecaster is said to be reliable if these two quantities always coincide, in which case the points on the reliability diagram lie along the line of equality (Bröcker and Smith, 2007a). Reliability in this sense is equivalent to auto- or conditional calibration (Gneiting et al., 2013; Tsyplakov, 2013), which also appears in well-known decompositions of proper scoring rules (Bröcker, 2009).

In order to estimate the auto-calibration in practice, the reliability diagram requires that there are only a finite number of possible values that the forecast can assume. If this does not hold, then a binning of the forecast into discrete groups is necessary, from which the conditional event frequencies can be determined. Estimating the auto-calibration of a forecaster who issues forecasts in the form of predictive distributions is yet more difficult - rather than binning the forecast values, it may be necessary to group together forecast distributions that share similar location, scale, and shape characteristics, potentially determined using intervals of post-processing parameters. Therefore, it is more common to assess the probabilistic calibration of such forecasts (Thorarinsdottir and Schuhen, 2018).

Although generally a weaker form of calibration, probabilistic calibration is equivalent to auto-calibration for forecasts of dichotomous outcomes (Gneiting et al., 2013; Strähl et al., 2017). However, probabilistic calibration has a natural interpretation in terms of the distribution of the probability integral transform (PIT) values. In particular, the PIT theorem states that if a random variable U follows a distribution with cumulative distribution function (CDF) F, then the random variable $Z = F(U)$ (i.e. the CDF of U evaluated at U) follows a standard uniform $\mathcal{U}(0, 1)$ distribution. From this, having specified a forecast distribution with CDF F, and having obtained an observation y, the PIT value defined as $z = F(y)$ should represent a draw from a uniform distribution if the observation does in fact follow distribution F. Properties of the distribution of PIT values can therefore be considered to assess whether or not the forecast is calibrated (Dawid, 1984). For example, the mean of a standard uniform distribution is 1/2, and its variance is 1/12. Checking whether the mean and variance of the sampled PIT values lie close to these values therefore constitutes a rough test for calibration (Gneiting et al., 2008).

More commonly, however, these PIT values are grouped together into a discrete number of bins, determined by evenly spaced intervals across the range of zero to one, and displayed in a histogram. If the PIT values follow a standard uniform distribution then they are equally likely to fall into any bin, meaning the resulting PIT histogram appears uniform (subject to sampling variations). If this is the case, then the forecast is said to be probabilistically calibrated. Statistical hypothesis tests exist to assess whether there is significant deviation from uniformity in the PIT histogram (Hersbach, 2000; Candille and Talagrand, 2005; Pinson and Girard, 2012; Wilks, 2019), though this is typically assessed only visually.

PIT histograms are the preferred tool to assess probabilistic calibration because there is a natural analogue that applies to ensemble forecasts, allowing a seamless comparison between forecasts of the two types. In particular, rank histograms display the frequency with which the observation assumes each rank when pooled alongside the ensemble members (Anderson, 1996; Talagrand, 1997b; Hamill and Colucci, 1997). This assumes that the ensemble members are independent samples from the underlying distribution of the response variable, and hence the observed evolution of the atmosphere is considered a plausible member of the ensemble. If this assumption is met, then the observation is equally likely to assume each rank in the ensemble, and the resulting rank histogram will be uniform. Furthermore, in cases where the forecast is not probabilistically calibrated, rank and PIT histograms can also be used to decipher what biases are present in the forecast (Hamill, 2001).

These histogram-based verification tools have become a central component in the evaluation of ensemble and probabilistic forecasts, so much so that statistical post-processing methods exist that focus almost solely on their output (Hamill and Colucci, 1997). As a result, they have been extended for use in multivariate settings, despite there being no natural way to order the ensemble members in this case (Gneiting et al., 2008). As noted by Hamill (2001), however, a uniform rank or PIT histogram is a necessary but insufficient criterion for a useful forecast. Therefore, to supplement these graphical displays, forecasters often also desire an objective measure of their forecast's performance.

2.4.3 Scoring rules

An objective measure of forecast performance is typically represented by a numerical value that quantifies the degree of association between the forecast and the corresponding verification. For point forecasts that take the form of just a single value, there exist several statistical measures of this ‘distance’ between forecasts and the observation (Jolliffe and Stephenson, 2012). In the context of forecast verification, these measures are typically referred to as scoring functions. Scoring functions take a forecast-observation pair and map this to a numerical output, allowing an effortless comparison of the performance of various forecasts.

For probabilistic forecasts, on the other hand, the problem is not so trivial. In particular, the forecast and the observation assume different forms: the forecast is a predictive distribution, whereas the observation is a single value (or a vector of values in the case of multivariate forecasts). Nonetheless, a function is similarly desired that maps from the joint space of predictive distributions and observations to the real numbers. Such functions are referred to as scoring rules. More formally, let Ω denote the set of all possible observations and let $\mathcal{F}$ represent the set of all possible forecast distributions. A scoring rule is then defined as the map

$$S \colon \mathcal{F} \times \Omega \to \mathbb{R} \cup \{\infty\}. \tag{11}$$

Scoring rules can be either positively oriented functions that we wish to maximise (i.e. utility functions), or negatively oriented functions that we wish to minimise (i.e. loss functions). Henceforth, interest is placed only on scores of the latter form, though this choice is arbitrary since a negatively oriented scoring rule can trivially be obtained by negating a positively oriented scoring rule.

Just as the forecasts should exhibit desirable properties in order to make correct inferences about the outcome, such as calibration, so too should the scoring rule possess certain characteristics that allow for a trustworthy assessment of the forecast accuracy. One particular property that is widely accepted as necessary is propriety. A scoring rule is said to be proper if the following criterion holds:

$$E_G[S(G, Y)] \leq E_G[S(F, Y)], \qquad \forall F, G \in \mathcal{F}, \tag{12}$$

where Y is the response variable which follows distribution G, and E_G denotes the expectation with respect to G (Gneiting and Raftery, 2007). That is, if the observations are believed to arise according to the distribution G, then the expected score is minimised by issuing G as the forecast. In this sense, forecasters are rewarded for issuing their true beliefs, and thus discouraged from issuing a forecast that they do not believe in but that they know would certainly yield a better score on average than their actual beliefs (Bröcker and Smith, 2007b) - proper scoring rules ensure that no such forecast exists. Conversely, a scoring rule that does not satisfy propriety (i.e Equation 12) is inconsistent, in that there exists a forecast distribution that would obtain a better score than the distribution from which the observations actually arose. Strict propriety implies that the inequality in Equation 12 is strict, so that the minimum score is attained only for the distribution G.

Numerous proper scores exist and readers are diverted to Gneiting and Raftery (2007) for a comprehensive review. Arguably the most frequently used scoring rule is the Brier, or probability score (Brier, 1950). The Brier score is used to assess categorical forecasts, in which case the outcome can assume only a finite number K of possible values:

$$\mathrm{BS}(\boldsymbol{P}, y) = \sum_{k=1}^{K} (P_k - y_k)^2. \tag{13}$$

The forecast distribution $\boldsymbol{P}$ in this case is a vector containing a probability P_k assigned to each category, while y_k takes the value one if the observation belongs to category k and $y_j = 0 \, \forall j \neq k$. The Brier score is thus equal to the sum of squared differences between the forecast probabilities and the outcome indicator, and low values of this score imply greater correspondence between the forecasts and observations. Commonly, the Brier score is used to assess forecasts of dichotomous events, such as the occurrence of precipitation or a weather variable exceeding a certain threshold. In this case, $K = 2$ and the Brier score is generally halved, since both components of the sum in Equation 13 are equal. This is typically also referred to as the Brier score, though it is more technically defined as the half-Brier score.

The Brier score considers unordered, or nominal, categories, though it can easily be extended to ordered categories via the (discrete) ranked probability score (Epstein, 1969b; Candille and Talagrand, 2005; Wilks, 2019). One example of ordered categories is forecasting whether the outcome exceeds a sequence of increasing thresholds. In the limiting case, as the number of thresholds tends to infinity, the forecast takes the form of a complete cumulative distribution function (CDF) over the range of possible outcomes. Integrating the Brier score over these values yields the continuous ranked probability score (CRPS):

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{u \geq y\}]^2 du \tag{14}$$

(Matheson and Winkler, 1976; Hersbach, 2000), where $\mathbb{1}\{\cdot\}$ denotes the indicator function that takes the value one if the statement inside is true, and zero otherwise. The CRPS can thus be interpreted as the squared difference between the predictive distribution function F and the empirical CDF associated with the observation, for all possible values of the response variable.

The CRPS is commonplace in weather forecasting studies, due in part to its relationship to the Brier score, and also to the quantile score (Friederichs and Thorarinsdottir, 2012). The CRPS also reduces to the mean absolute error for a deterministic forecast (Gneiting et al., 2005), and there exists a convenient representation of the score when assessing forecasts in the form of an ensemble (Grimit et al., 2006). As such, the CRPS operates on the same scale as the observations, giving it an intuitive interpretation. Moreover, just as the CRPS extends the Brier and discrete ranked probability scores to continuous forecasts, the energy score provides a further generalisation for multivariate forecasts (Gneiting and Raftery, 2007). The Brier score, the discrete and continuous ranked probability scores and the energy score therefore comprise a family of scores that can assess a multitude of forecast types.

The CRPS exploits the predictive distribution function and therefore considers the entire distribution when evaluating the forecast. As such, it is a non-local score. A scoring rule is said to be local if it assesses a forecast only through the probability density assigned to the verifying observation. That is, two distributions that assign the same probability mass to the verifying value should be given the same score, regardless of the behaviour of their distributions at other values. The motivation for using local scores is fuelled by the idea that the score assigned to a forecast should depend only on the value that actually occurs, and not on the possible values that could have occurred but did not (Bernardo, 1979). Benedetti (2010) argues that non-local scores should be considered inadmissable, such is the importance of being local.

Bernardo (1979) goes on to show that the only smooth, continuous scoring rule that satisfies this criterion is the logarithmic score, or affine transformations thereof:

$$\mathrm{LS}(F, y) = -\log f(y), \tag{15}$$

where f is the predictive density function associated with the CDF F. The logarithmic, or negative log-likelihood, score utilises the natural logarithm in Equation 15, meaning it is proportional to the ignorance score (Roulston and Smith, 2002) and thus has a strong connection with information theory, giving it an appealing basis on which to assess forecasts (Benedetti, 2010). Furthermore, Benedetti (2010) shows that the Brier score is a second-order approximation to the logarithmic score, and claims that this explains some of the Brier score's desirable properties, while also highlighting its limitations. In particular, the steepness of the logarithmic function at values close to zero means the logarithmic score imposes a much harsher penalty than the Brier score when the forecast assigns little (or no) mass to an event that subsequently occurs. The

Brier score, in comparison, is more conservative.

However, although locality is theoretically desirable, there are practical issues concerning the exclusive use of the logarithmic score to assess forecasts. For example, there is often non-negligible error in the observation, due in part to precision errors in measuring equipment. Non-local scores, such as the CRPS, do not judge a forecast only on the probability it assigns to the outcome that occurs, and hence forecasters are also encouraged to assign high probability density to outcomes close to the observation. In this sense, they encourage smoothness of the forecast distribution (Bröcker and Smith, 2007b), and, in doing so, are less sensitive to observation error than local scores. Moreover, the logarithmic score depends on the probability density function of the forecast distribution, rather than its CDF, and hence forecasts in the form of an ensemble cannot directly be assessed using this score. Hence, since different scoring rules themselves exhibit distinct properties, it is typically recommended that numerous scores be employed to assess forecasts in order to draw robust conclusions regarding their predictive performance (Gneiting and Raftery, 2007; Scheuerer and Hamill, 2015b).

2.4.4 Comparing forecasters

The proper scoring rules introduced in the previous section are defined for one particular forecast instance. In reality, to compare various prediction systems, it is necessary to consider the expected value of the scores. That is, we wish to evaluate $E_G[S(F,Y)]$, for a forecast distribution F and outcome distribution G. By definition, for proper scores this is minimised when the forecast and realised distributions coincide. In practice, the distribution G is not known, nor can we even be sure that such a distribution exists. Instead, only a finite sample of materialisations of a weather variable is available. Assuming these observations are independent, the expected value of the score can be estimated using the sample mean as an unbiased estimator:

$$\hat{E}_G[S(F,Y)] = \frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i), \tag{16}$$

where subscript i provides an index over the set of n forecasts and observations, and a hat is used to denote that this is an estimator of the expectation. Two forecast schemes can then be compared through their sample mean scores, with a lower score indicative of a better prediction system. Naturally, this average score is subject to sampling variations, and statistical hypothesis tests can thus be used to confirm or deny the significance of any difference between scores (Diebold and Mariano, 2002; Wilks, 2016).

Rather than comparing the performance of forecasters via their expected scores, they can be directly compared using skill scores. The skill of a forecaster measures the accuracy of their forecasts relative to those from another prediction scheme. There are several ways of defining a skill score (Murphy, 1973a), though for a score S, its skill score is most commonly defined as

$$SS(F, F_{ref}, Y) = \frac{\hat{E}_G[S(F_{ref}, Y)] - \hat{E}_G[S(F, Y)]}{\hat{E}_G[S(F_{ref}, Y)] - \hat{E}_G[S(F_{perf}, Y)]}, \tag{17}$$

where F_{ref} denotes a reference forecast scheme, while F_{perf} represents a perfect forecast (Wilks, 2019). Skill scores therefore provide a measure of how much better (or worse) a forecaster F is, relative to a reference, or baseline prediction scheme. This reference forecaster is often taken to be the unconditional, or climatological forecaster, which is known to be calibrated yet uninformative (Mason, 2004). Unlike the scores themselves, skill scores are positively oriented and bounded above by one, while a skill score below zero indicates the forecaster is performing worse than the reference scheme. Testing for the significance of a difference in forecast performance can again be achieved using statistical tests, typically via nonparametric bootstrap resampling methods (Efron, 1982).

3 Regime-dependent statistical post-processing of ensemble forecasts

3.1 Introduction

The previous chapter introduces numerical weather prediction (NWP) models and discusses the imperfections inherent in their construction. Although ensemble prediction systems can address some of these deficiencies, statistical post-processing methods, such as Bayesian Model Averaging (BMA; Raftery et al., 2005) and Non-homogeneous Regression (NR; Gneiting et al., 2005), are necessary to obtain a more realistic forecast given the output from the NWP model. Not only do these statistical post-processing methods provide probabilistic forecasts in the form of predictive distributions, but they exploit relationships between historical forecasts and observations to correct for systematic model errors.

However, the relationship between forecasts and observations may change under different circumstances. If such circumstances are known then it is often possible to incorporate this additional information into established post-processing methods. For example, to account for seasonal biases in the NWP model, it is common to implement a rolling training window (Gneiting et al., 2005), which may be comprised of only forecast-observation pairs from particular locations to account for locally-varying biases (Thorarinsdottir and Gneiting, 2010; Lerch and Baran, 2017; Hamill et al., 2017). In this chapter, we postulate that the relationship between the dynamical weather model and the atmosphere changes depending on the concurrent behaviour of the atmospheric circulation, which is generally represented by a finite number of atmospheric, or weather, regimes.

Numerous studies have investigated how weather regimes influence the biases of synoptic-scale model output, finding that forecast errors are dependent on the underlying regime (Koch et al., 1985; O'Lenic and Livezey, 1989; Stoss and Mullen, 1995). In comparison, limited work has examined how forecasts of surface weather variables rely on the atmosphere's flow. Weather regimes have, however, previously been used to calculate ensemble member weights in a consensus forecast for temperature (Greybush et al., 2008), and to calibrate and blend short-range precipitation forecasts (Kober et al., 2014). Lorenz (1969) remarks that the low-frequency circulation has a much larger range of predictability than shorter-scale flows, and hence predictions of the large-scale information could be used to assist forecasts of high-frequency, noisy events, such as the weather. Scheuerer (2014) therefore suggests that weather regimes could provide a suitable basis on which to train post-processing methods.

The aim of this chapter is to illustrate how this could be achieved. The weather regime paradigm is introduced in more detail in the following section, though readers are diverted to Hannachi et al. (2017) and references therein for a more thorough review of the extant literature. Approaches to include weather regime information into statistical post-processing methods are discussed in Section 3.3, before Section 3.4 introduces a highly idealised representation of the atmosphere, the Lorenz (1996) system, in which these approaches are tested. The post-processing methods and forecast verification techniques are presented in Section 3.5, with the corresponding results for the simulation study displayed in Section 3.6. Section 3.7 discusses potential extensions to operational forecasts, while also concluding.

3.2 Weather regimes

3.2.1 Introduction

Although the chaotic nature of the atmosphere renders its instantaneous state (i.e. the weather) almost impossible to predict with certainty, features of the long-term atmospheric behaviour (i.e. the climate) are considerably more predictable. In this sense, weather forecasting represents an attempt to uncover the small-scale, rapid fluctuations around the predictable atmospheric signal. It is thus in the interest of forecasters to identify the recurrent patterns in the atmosphere's circulation that comprise this predictable signal.

Seasonal trends in the atmosphere are well understood, but also of interest to meteorologists is the behaviour of the atmosphere on time-scales between those of daily weather fluctuations and the seasonal cycle; that is, the low-frequency, or intraseasonal atmospheric variability. Although the low-frequency atmosphere is governed by nonlinear and chaotic dynamics, it is widely accepted that this component of the atmosphere frequently arranges itself into familiar patterns (Charney and DeVore, 1979). These patterns, often known as weather regimes, persist beyond the timescales of individual weather events (typically around one week on average), and tend to manifest repeatedly at the same geographical locations. As such, these regimes explain a remarkably high proportion of the low-frequency atmospheric variability. The continuous evolution of the atmosphere can thus be reasonably well described by transitions between distinct weather regimes (Franzke et al., 2011).

Although there does not appear to be a precise, formal definition of an atmospheric regime, they can be thought of dynamically as quasi-stationary, or metastable equalibria in the atmosphere's phase space (Reinhold and Pierrehumbert, 1982; Franzke et al., 2008). Notable examples of such regimes include teleconnection patterns (Wallace and Gutzler, 1981) - highly correlated weather variables situated at distant spatial locations

- and persistent anomalies in geopotential height fields (Dole and Gordon, 1983), such as atmospheric blocking (Woollings et al., 2018).

The concept of weather patterns, or weather types has existed in one form or another throughout the history of weather forecasting, though Baur et al. (1944) are widely credited with formalising the notion by introducing a set of weather types ('grosswetterlagen') that influence the weather over central Europe. Since then, it has become common for operational weather centres to maintain a catalogue of weather patterns that arise in the surrounding area (e.g. Ferranti and Corti, 2011; Neal et al., 2016). However, weather types, or patterns differ from atmospheric regimes in that they do not necessarily persist. In persisting for prolonged periods of time, regimes assert a higher influence on local weather systems, and are thus often linked to the occurrence of extreme weather events (Carrera et al., 2004).

It is widely recognised that atmospheric regimes have a significant effect on local weather systems, and, as such, considerable effort has been devoted to identifying regimes from archives of historical atmospheric data. Nonetheless, despite their long history, the underlying causes behind the formation of these regimes are still today not well understood (Hannachi et al., 2017). Diagnosing the prevailing weather regime is therefore challenging from a dynamical viewpoint (Vautard, 1990). Instead, several statistical methods have been proposed for this purpose.

3.2.2 Statistical representation

It has become common to define weather regimes in terms of the statistical artefacts that arise from the application of a classification routine to the anomaly field of a large-scale circulation variable, such as mean sea level pressure or geopotential height. The anomaly field in this setting represents the difference between a spatial field and its mean state, both of which are typically defined using high resolution model (re)analyses. Since the spatial scale of these atmospheric regimes is generally reasonably large, it is often desirable to first reduce the dimension of the spatial anomaly fields, prior to searching for regimes. This is typically achieved using principal component analysis (PCA; also known as empirical orthogonal function analysis). PCA works by transforming the anomaly fields to a new set of orthonormal variables that are linear combinations of the original circulation variable on the spatial domain under consideration, where the linear mapping is typically designed such that the transformed variables explain a large proportion of the variation in the spatial fields through a comparatively small number of variables (Wilks, 2019).

The transformed variables themselves, called the principal components, do not generally resemble weather regimes: they are the transformations of the atmospheric fields that explain the largest amount of variation in the data, which typically arise from an amalgamation of several different weather patterns (Horel, 1981). Horel demonstrates that a further linear transformation of the leading principal components could be used to isolate the contributing patterns, though it is often preferred to use the initial, untransformed principal components as a basis on which to perform alternative statistical procedures to identify the regimes. In doing so, the methods can identify nonlinear features in the spatial structures of the weather regimes (Corti et al., 1999).

Examples of such statistical methods include analysing the correlation between atmospheric fields to measure the similarity between two synoptic conditions (Horel, 1985), or searching for "local maxima in probability density evaluated in the system's phase space" (Kimoto and Ghil, 1993), in turn assuming that the different regimes account for distinct modes of the atmosphere's attractor. Machine learning approaches have also recently been proposed to identify weather regimes (Deloncle et al., 2007). The most popular approach, however, is to define regimes as clusters in the archive of transformed anomaly fields.

Cheng and Wallace (1993), for example, propose a hierarchical clustering algorithm that iteratively groups together (transformed) anomaly fields such that a collective measure of the homogeneity of the resulting clusters is minimised. The process is stopped when a reproducibility statistic of the clusters falls below a predefined threshold, and the groups, or clusters obtained at the end of this process are defined as weather regimes. Alternatively, Michelangeli et al. (1995) suggest using k-means clustering to identify regimes. Whereas hierarchical clustering approaches use a stopping criterion to estimate the optimum number of clusters given the data, k-means clustering groups the anomaly fields into a predetermined number of clusters, such that some distance between the distinct clusters is maximised. Like the hierarchical approach, however, k-means clustering assigns each data point to one, and only one, cluster. Smyth et al. (1999), on the other hand, introduce a method that fits a mixture model to the transformed variables. The weather regimes are thus defined as statistical distributions in the space of the leading principal components. This approach therefore constitutes a probabilistic clustering routine in which each point is assigned a probability of belonging to each cluster, thereby addressing the uncertainty when identifying the prevailing regime.

All of these clustering approaches, however, tend to focus only on recurring spatial structures in the anomaly fields, rather than their temporal properties. As such, although the weather regimes identified using these approaches are the most recurrent patterns in the atmosphere's phase space, they do not necessarily persist. Majda et al. (2006) therefore introduce hidden Markov models (HMMs) as a tool to detect weather regimes. Hidden Markov models use the underlying dynamics of the system to detect regimes that are persistent as well as recurrent (Franzke et al., 2008). They are therefore capable of detecting regime-like behaviour despite the atmosphere's probability density exhibiting unimodal statistics. To achieve this, HMMs extend the mixture model clustering approach of Smyth et al. (1999) to include a transition matrix that documents the preferred transitions between the identified states. Therefore, like mixture model clustering, HMMs have the added benefit of issuing a probability of residing in each state, while this probability depends not only on the observed spatial field, but also on the previous regime sequence.

However, to accurately estimate the transition matrix, HMMs are dependent on a relatively long time series of anomaly fields. As a result, alternative clustering methods are often more convenient to apply, and therefore tend to be implemented more frequently to define weather regimes in practice (Falkena et al., 2020). Furthermore, such approaches are also more readily applicable to deduce the regime from the output of a numerical weather model. In doing so, several studies have been able to evaluate the ability of weather and climate models to simulate the regime structure observed in the atmosphere.

3.2.3 Regimes in Numerical Weather Prediction models

In applying the statistical methods outlined in the previous section to global model reanalyses, several studies have found evidence that weather regimes exist in the atmosphere. Due to the simplifications necessary to construct numerical weather models, however, it is not necessarily the case that the forecasts generated by these models will reproduce this regime-like behaviour. Dawson et al. (2012) show that the ability of a numerical weather model to simulate atmospheric regimes depends on the model's resolution: numerical models that operate on coarse spatial grids, such as climate models, cannot accurately replicate the observed regimes, whereas weather models, which are typically run at much higher resolutions for considerably shorter periods of time, are capable of mimicking the regime behaviour. As the authors acknowledge, it is not clear whether there is an exact threshold of the resolution beyond which the weather models are unable to simulate the regimes. Dawson and Palmer (2015) therefore use the ability to capture these regimes as a way of assessing the performance of numerical weather and climate models.

However, Dawson et al. (2012) recognise that although the regimes identified by the high resolution weather models resemble those observed in reanalyses, they do not necessarily exhibit identical characteristics. For example, numerical weather models tend to prefer transitions between certain weather regimes, leading to fallacious climatological regime frequencies compared with those estimated from reanalyses (Ferranti et al., 2015; Neal et al., 2016; Matsueda and Palmer, 2018); in particular, the numerical weather models seem to inadequately simulate the onset, maintenance, and decay of atmospheric blocking events (Tibaldi and Molteni, 1990). Therefore, although the flow-dependent spread of an ensemble forecast might capture the well-established dependence of the atmosphere's predictability on the initial flow regime (Thompson, 1957; Leutbecher and Palmer, 2008), regime-dependent forecast errors may arise from the weather model failing to predict the regime that materialises. In this case, the forecast biases would depend not only on the regime that is forecast, but also on the regime that occurs in reality: if the weather variable depends strongly on the regime then forecast errors would be expected to be larger when the regime predicted by the numerical model does not agree with that which manifests in the atmosphere.

Moreover, the spatial structures of the forecast regime centres of action may be shifted or distorted somewhat relative to those observed in reanalyses, while the intensity of the associated pressure systems may also not be perfectly represented (Dawson et al., 2012). As a result, the NWP model may not accurately capture the relationship between the atmospheric regimes and the weather variable(s) being forecast, so that forecast errors will depend on the prevailing regime even if the NWP model accurately predicts the regime that will occur. For example, even if the forecast correctly simulates the formation and persistence of an anticyclone over the North Atlantic Ocean, errors in the exact position of this blocking regime could have a significant influence on the behaviour of the Northern Hemisphere polar jet stream, which is linked heavily to the surface weather over much of Europe (Woollings et al., 2010, 2018). Therefore, the quality of forecasts emitted from a numerical weather model will likely depend on the prevailing atmospheric regime, regardless of whether or not the regime itself is accurately predicted. Statistical post-processing methods to recalibrate NWP model output should thus utilise this regime information. This chapter seeks to investigate how this could be achieved, and in what circumstances this is most beneficial.

3.3 Methodology

In particular, we focus on extending established methods of statistically post-processing ensembles of weather forecasts, which generate forecasts in the form of a predictive density function f, with corresponding distribution function F, conditional on an ensemble forecast $\boldsymbol{x} = (x_1, x_2, ..., x_M)$ comprised of M members. We consider the case where the response variable Y (and its corresponding observation y) is univariate, though the method could easily be extended for multivariate post-processing, which might also address spatial structures in the biases that arise due to the prevailing regime.

Firstly, note that it is possible to include information regarding the atmosphere's circulation into post-processing methods without utilising the concept of weather regimes. If the circulation can be quantified by some continuous metric, ρ, then the predictive distributions could simply be extended to include this metric as an additional input variable during the recalibration. Using a continuous measure allows the flow to be represented on a spectrum and, rather than binning the circulation into a finite number of regimes, it permits a degree of membership to several states to be quantified. In reality, although indices exist that measure how much the atmosphere resembles commonly recognised weather regimes such as the Notrh Atlantic Oscillation (Hurrell, 1995; Hurrell and Deser, 2009), the Arctic Oscillation (Thompson and Wallace, 1998; Baldwin, 2001) and the Pacific-North American pattern (Leathers et al., 1991), there is no recognised method of objectively condensing the flow over a spatial domain into a single continuous metric.

Moreover, although there has been some debate on the irrefutable presence of atmospheric regimes (Stephenson et al., 2004), they are a useful feature in this framework. Defining regimes to exhibit persistence renders the time spent transitioning between states negligible compared to time spent in the regimes, and the regime states thus form a mutually-exclusive, collectively-exhaustive (MECE) partition of the atmosphere's phase space. Suppose then that a finite number, R, of regimes in the atmosphere are identified. If an underlying regime can accurately be attributed to a forecast, then recalibration can be performed conditional on this atmospheric state. For example, when forecasting in each regime, the post-processing methods could use a separate set of model parameters, or even specify a distinct distribution. More generally, the forecast distribution can be written as a mixture of predictive distributions that depend on the regime:

$$f(y|\boldsymbol{x}) = \sum_{r=1}^{R} w_r f_r(y|\boldsymbol{x}, r), \tag{18}$$

where w_r is the weight associated with regime r, allowing the model to account for uncertainty present when attributing the forecast to a regime. It is important to note that the weights in this case are a function of time, rather than a parameter, as such, highlighting that the weight will change depending on the prevailing behaviour of the atmosphere. The forecast then takes the form of a predictive distribution, which is itself composed of component predictive distributions $f_r(y|\boldsymbol{x}, r)$ that depend on the prevailing weather regime. This chapter focuses on these regime-based extensions: the idea of introducing a continuous metric to measure circulation is not investigated. Discretising the flow like this places fewer restrictions on any post-processing parameters, allowing for more flexibility in the statistical recalibration methods.

Moreover, although we focus here on weather regimes, this approach is suitable for any grouping of the forecasts in which different model biases might be expected. Similar extensions to statistical post-processing methods have been implemented previously in the hope of attaining more skilful forecasts of extreme wind speed events. Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015), for example, apply a regime-switching approach that issues a separate predictive distribution depending on whether or not the ensemble median lies above some threshold, suggesting that biases in the forecasts depend on the predicted values themselves. Rather than using a fixed threshold, Baran and Lerch (2016) extends this idea further by utilising a continuous mixture of predictive distributions, with weight parameters that are estimated simultaneously with the coefficients of the component distributions. Although the regime-switching approaches implicitly assume that forecast biases differ between two or more distinct configurations of the atmosphere, they do not necessarily refer to weather regimes. Gneiting et al. (2006), however, find that skilful short-range forecasts of wind speed are obtained when separate recalibration models are fitted depending on the prevailing local wind direction.

Statistical post-processing corrects systematic errors in the raw ensemble by exploiting relationships between archived forecasts and their corresponding verifications. Thus, a training data set of historical forecasts and observations - forecast-observation pairs - is required, from which relationships can be identified. Continual adjustments to NWP models often limit the training data available to operational forecasters, whereas the low-frequency circulation is a product of the atmosphere only and does not depend on the forecast. Therefore, the regimes need not be estimated from the training data, they can be discerned from a much larger set of observations. However, regimes are latent and must be inferred from other, observable variables. As discussed in the previous section, this can be circumvented by converting these dynamical phenomena to statistical artefacts. There has been extensive work on regime detection and the framework presented here assumes only that the statistical representations are reasonable approximations of their dynamical counterparts - beyond this, the choice among methods is arbitrary. The regimes are hereafter assumed to be known.

The regime-dependent approaches rely on the ascribing of forecasts to an underlying regime. In doing so, the training data set can be stratified into R MECE subsets. Relationships can then be identified between the forecasts and the observations from each of the separate subsets (Scheuerer, 2014). However, the regime of a forecast is not unique: the underlying regime may change throughout the forecast trajectory and thus a time at which to define the regime must be chosen. We seek the time at which the disparities in forecast errors between the regimes are largest. That is, when the relationship between the numerical weather model and the atmosphere changes most drastically between the regimes. There are two intuitive options: the forecast's initialisation time or its validation time.

It is well-established that the predictability of the atmosphere depends on the initial flow regime (Thompson, 1957), and hence biases in the forecast may themselves depend on the atmospheric regime at the forecast initialisation time. In this case, the training data can be stratified into subsets depending on the regime deduced from the forecast model analysis and separate predictive distributions can be estimated over each subset. This assumes that all ensemble members estimate the same regime at the initialisation time and that a small perturbation to the analysis field is not sufficient to alter the large-scale state of the atmosphere. Any new forecast would then be assigned to a regime in the same way and post-processed using the predictive distribution estimated from the corresponding training subset.

However, it may be the case that the spread of the ensemble members captures this flow-dependent uncertainty (Leutbecher and Palmer, 2008). Moreover, since the length of a medium-range weather forecast may exceed the average duration of a weather regime, the atmospheric state will often change throughout the forecasting period. Therefore, conditioning on the regime at the initialisation time may result in the loss of information regarding the occurrence of different weather events in different regimes, contradicting some of the reasons for believing this method may be successful, such as extreme events occurring more frequently in certain regimes.

The regime of the atmosphere at the forecast validation time, on the other hand, does not suffer from these problems and therefore may be associated with more heterogeneous relationships between the model and the atmosphere. However, in order to exploit past regime-dependent relationships, the regime of a new forecast should be defined in the same way as those in the training data. Therefore, because the regime at the forecast validation time is not known prior to issuing the forecast, it must instead be estimated using information that is available to the forecaster at the initialisation time. Since the dynamical weather model simulates the trajectory of the atmosphere, each ensemble member provides an estimate of the future atmospheric state. Therefore, provided forecast fields are of the same spatial scale as the regimes, each ensemble member can be matched with the regime that is statistically the closest (Neal et al., 2016), generating an ensemble forecast for the regime. This ensemble can then be used to estimate the probability of the atmosphere residing in each regime at the forecast validation time, which can in turn be employed as mixture model weights in Equation 18. In this case, since every forecast-observation pair would not necessarily be assigned to exactly one regime, rather than stratifying the training data into subsets for each regime and fitting a separate predictive distribution to each subset, a model averaging technique could be applied in which all component distributions are estimated simultaneously.

Alternatively, methods that post-process each ensemble member separately, such as Ensemble Dressing techniques (Roulston and Smith, 2003; Wang and Bishop, 2005), Bayesian Model Averaging (Raftery et al., 2005), and member-by-member approaches (Van Schaeybroeck and Vannitsem, 2015; Williams, 2016), could calibrate each ensemble member using its own regime prediction. This accounts for the fact that the nature of the biases in the NWP model will depend on whether or not the model accurately predicts the regime that will occur, and ensemble members that correctly identify the future regime can thus be post-processed differently to the members that do not. In this sense, ensemble members that forecast different regimes are assumed not to be exchangeable.

Furthermore, it would be possible to define the regime using the state of the atmosphere at any intermediate time of the forecast, or even at any time prior to forecasting if such information were available, though there is little reason to believe that forecast errors would depend more strongly on the atmospheric regime at these instances. As such, Section 3.5 reintroduces Non-homogeneous Gaussian Regression (NGR), also commonly referred to as Ensemble Model Output Statistics (EMOS), and Bayesian Model Averaging (BMA), and offers examples of possible extensions to these familiar statistical post-processing methods using the regime paradigm. A separate extension is considered when defining the regime at the initialisation time and at the validation time.

3.4 Lorenz '96 system

The methodology described in the previous section is implemented in a highly idealised model of the atmosphere, the Lorenz (1996) system, whose chaotic nature lends itself to simulations of weather forecasts and the trialling of statistical post-processing methods (Roulston and Smith, 2003; Wilks, 2006; Williams et al., 2014). A coupled system containing both larger-scale variables, X_k, and sub-grid-scale variables $W_{j,k}$ is used to emulate the atmosphere:

$$\begin{aligned} \frac{dX_k}{dt} &= -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + D - \frac{hc}{b}\sum_{j=1}^{J} W_{j,k}; \\ \frac{dW_{j,k}}{dt} &= -cbW_{j+1,k}(W_{j+2,k} - W_{j-1,k}) - cW_{j,k} + \frac{hc}{b}X_k, \end{aligned} \tag{19}$$

for $k = 1, ..., K$ and $j = 1, ..., J$, where each system exhibits cyclic boundary conditions, $X_k = X_{k+K}, W_{j,k} = W_{j,k+K}$ and $W_{j+J,k} = W_{j,k+1}$.

The parameter values used here are $K = 8, J = 32, D = 20, h = 1, b = 10$ and $c = 10$ and the system is numerically integrated forward in time using a fourth-order

Parameter	ξ_0	ξ_1	ξ_2	ξ_3	ξ_4
Estimate	0.209	1.45	-0.0127	-0.00728	0.000312

Table 1: Parameter estimates for the quartic polynomial in the surrogate of the NWP model (Equation 22).

Runge-Kutta scheme with a time-step of $dt = 0.001$. Christensen et al. (2015) show that with these parameters, the system exhibits regime-like behaviour, transitioning between two distinct states. The regimes are defined using a prespecified diagnostic:

$$\sum_{k=1}^{\frac{K}{2}} \text{cov}\big(X_k, X_{k+\frac{K}{2}}\big) \tag{20}$$

where $\text{cov}(X_i, X_j)$ denotes the covariance between the i-th and j-th components of the vector of state variables $\boldsymbol{X}$, calculated over a time-series of length one model time unit (MTU; corresponding to 5 days) directly preceding the time of interest. The system is said to reside in regime A if this covariance diagnostic is positive, and regime B if it is negative. As such, regime A is characterised by high amplitudes of wavenumber 2, whereas regime B is dominated by wavenumber 1. This diagnostic allows a regime to be known with certainty, and thus removes the need to account for any uncertainty regarding the current state of the system.

We represent the NWP model in this framework using the equations that resolve only the large scales, since this is a common simplification of dynamical weather models:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + D. \tag{21}$$

In an effort analogous to improving the NWP model, this equation can be extended by including a quartic polynomial of the resolved variable, which acts as a kind of sub-grid model, or parametrisation scheme to account for the effect of the neglected variables $W_{j,k}$:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + D - (\xi_0 + \xi_1 X_k + \xi_2 X_k^2 + \xi_3 X_k^3 + \xi_4 X_k^4). \tag{22}$$

The parameters $\xi_0, \xi_1, ..., \xi_4$ are estimated by minimising the mean squared difference between true and parametrised tendencies (Wilks, 2005; Kwasniok, 2012). The resulting coefficient estimates are shown in Table 1. This model is numerically integrated through time using a fourth-order Runge-Kutta scheme with a time-step of $dt = 0.005$.

To trial the regime-dependent statistical post-processing approach, a training data set is generated, comprised of forecasts initialised at points 0.15 MTU apart, from which predictive distributions are estimated. A fixed training data set is used throughout, consisting of 20,000 forecast-observation pairs, and trajectories up to a lead time of 3 MTU (15 days) are considered. The resulting statistical post-processing methods are evaluated over 50,000 ensemble forecasts initialised at intervals of 50 MTU, akin to Wilks (2006). Along each margin, k, ensembles are generated by adding a stochastic perturbation to the initialisation points, governed by a $N(0, 0.1^2)$ distribution, and integrating the NWP model through time starting at these perturbed points. Ensembles of size 20 are used throughout, though the results were found not to depend on the ensemble size. To allow for exchangeable members, these ensembles do not contain a control, or analysis, forecast.

There are two different systems being considered: the 'true' system imitating the atmosphere (Equation 19) and a deterministic NWP model with which ensemble forecasts can be generated (Equation 22). Table 2 shows the average persistence time of the regimes, along with the corresponding proportion of time the system spends residing in each regime. In the true system, regime A persists for 6.23 MTU (31 days) on average, and regime B only 1.60 MTU (8 days). The NWP model captures the mean persistence time of regime B but severely overestimates the persistence of regime A. Therefore, the weather model spends a larger proportion of time in this state than the atmosphere.

The numerical weather model is used here to predict two different quantities. The system is invariant under translation and hence all margins of $\boldsymbol{X}$ are statistically identical. Therefore, since we are interested in univariate post-processing approaches, only X_1 is considered. Secondly, the mean squared value of all X_k variables is also forecast. This quantity is labelled E since it is proportional to the total energy of the system (Lorenz, 1996):

$$E = \frac{1}{K} \sum_{k=1}^{K} X_k^2. \tag{23}$$

To visualise the regime-like behaviour, Figure 1 shows a year-long time-period (73 MTU) of the predictands, X_1 and E, along with the covariance diagnostic and the corresponding regime. Large spells in regime A with intermittent periods in regime B reinforce the features displayed in Table 2. There is no obvious disparity in the behaviour of X_1 depending on the regime of the system and this is confirmed by a plot of the empirical distributions of the observations in Figure 2. E, on the other hand, does appear to vary with the regime, with lower values coinciding with the occurrence of regime B. Although the distributions of X_1 are similar between the two regimes, nothing can be deduced from Figure 2 about the behaviour of the forecasts nor the predictability of the system in each regime. Therefore, regime-dependent post-processing may still benefit forecasts made for this variable. One particularly interesting attribute is that the covariance diagnostic appears less erratic during prolonged spells in regime B, perhaps

	Mean duration		% of time	
	Reg. A	Reg. B	Reg. A	Reg. B
True system	6.23	1.60	80	20
NWP model	12.13	1.61	88	12

Table 2: Average duration (MTU) of regimes A and B and the proportion of time the systems spend in each regime.

implying that the system is more settled in this regime.

3.5 Statistical post-processing

Consider again a raw ensemble forecast $\boldsymbol{x} = (x_1, x_2, ..., x_M)$ comprised of M members. Numerous techniques exist to statistically post-process the ensemble, most of which are variants of the two most eminent methods, Bayesian Model Averaging (BMA) and Non-homogeneous Gaussian Regression (NGR). Whereas an ensemble forecast constitutes a collection of point forecasts - instantaneous realisations of phase space - BMA and NGR generate probabilistic forecasts in the form of parametric predictive distributions. These methods both assume that each observation, or verification, y, is a realisation of a random variable, Y, that follows a proposed statistical distribution conditional on the M point predictions issued by the raw ensemble forecast.

Despite deviations from Gaussianity in the marginal distributions of the observed values (Figure 2), suitable diagnostic checks, such as the quantile-quantile plots of the standardised residuals in Figure 3, show that the normal distribution is an appropriate choice for the predictive distribution for both $Y = X_1$ and $Y = E$. E, by construction, is a positive quantity and using a normal predictive distribution issues a non-zero probability of observing a negative response. In this case, however, this probability is always negligibly small. A Gamma EMOS model (Scheuerer and Möller, 2015) was also implemented when forecasting E, but was found to perform worse than NGR (not shown).

3.5.1 Bayesian Model Averaging

BMA entails specifying a mixture of weighted component distributions that are centred around a linear adjustment of each ensemble member (Raftery et al., 2005). Here, we assume that all M members are interchangeable and hence equally weighted:

$$Y|\boldsymbol{x} \sim \frac{1}{M}\sum_{m=1}^{M} N(\alpha + \beta x_m, \sigma^2). \tag{24}$$

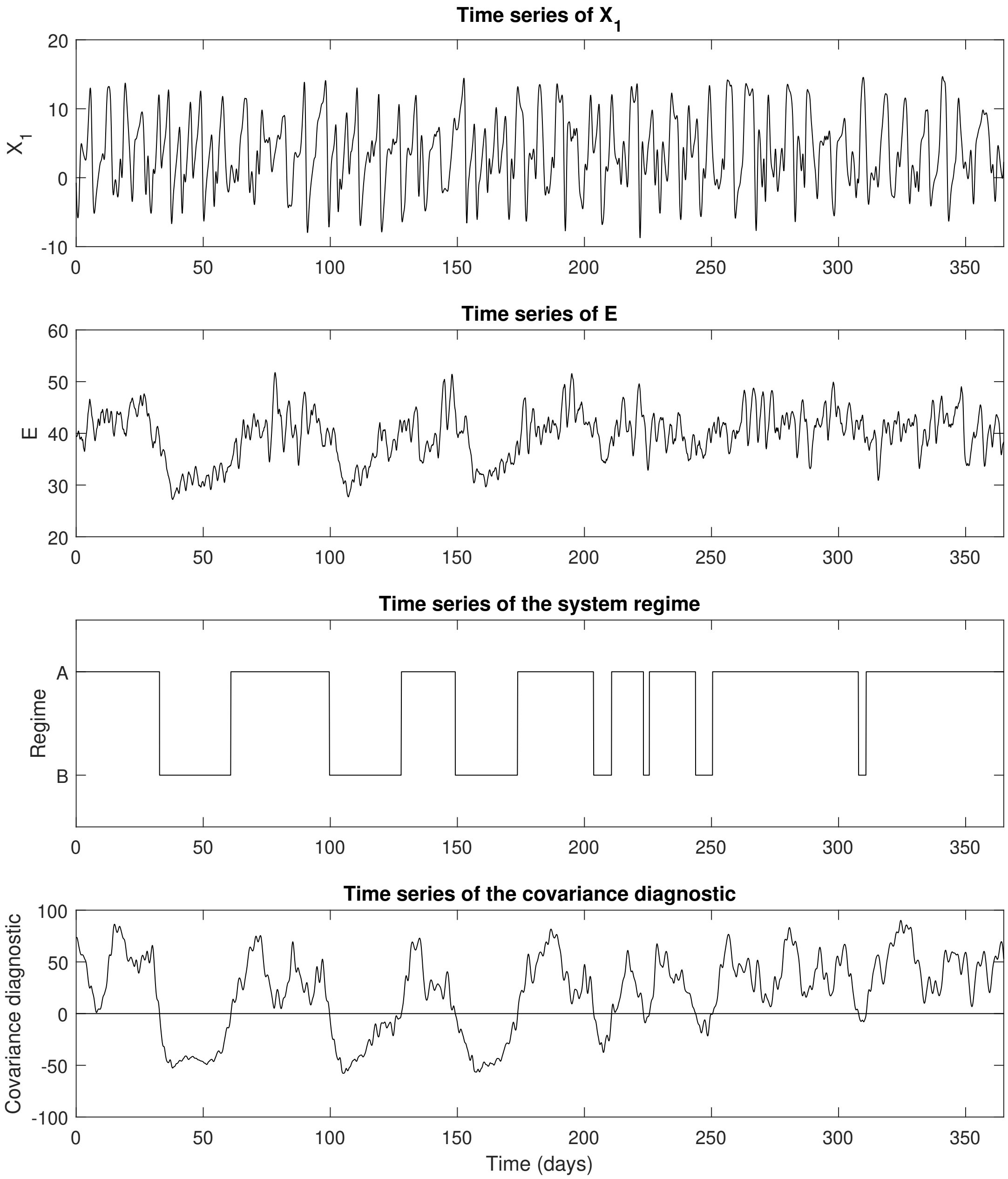

Figure 1: Time series of the observed predictands for a year-long time period, along with the concurrent regime and the associated value of the covariance diagnostic.

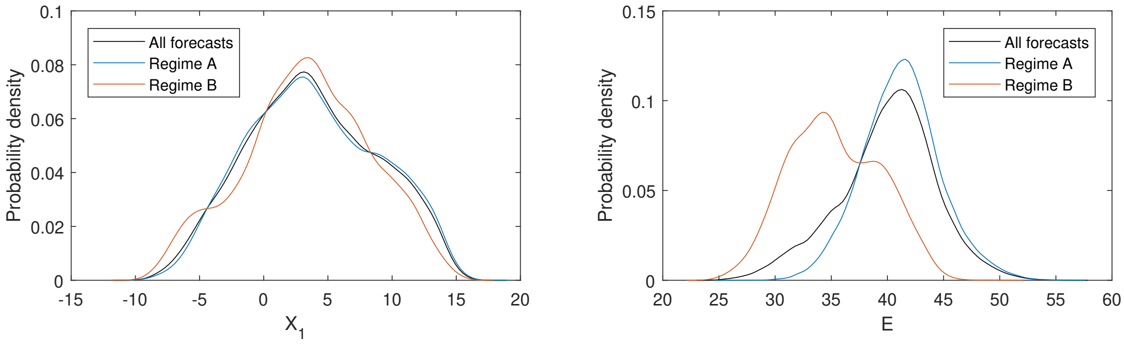

Figure 2: Empirical distribution of X_1 (left) and E (right) when the true system resides in each regime.

The individual component distributions are Gaussian, and the parameters $(\alpha, \beta, \sigma^2)$ are estimated by numerically maximising the likelihood function, or, equivalently, minimising the logarithmic score (LS). The LS for a mixture distribution as in Equation 24 is

$$\mathrm{LS}(F_{BMA}, y) = -\log\left[\frac{1}{M}\sum_{m=1}^{M}\phi\left(\frac{y-\alpha-\beta x_m}{\sigma}\right)\right], \tag{25}$$

where $\phi(\cdot\,)$ is the standard Gaussian probability density function, y is the corresponding observation, and F_{BMA} is the cumulative density function (CDF) corresponding to Equation 24. Equation 25 is then averaged over all forecasts in the training data to obtain the overall logarithmic score.

In the regime paradigm, we propose two different extensions to BMA depending on when the regime of the forecast is defined. If the atmospheric regime is defined

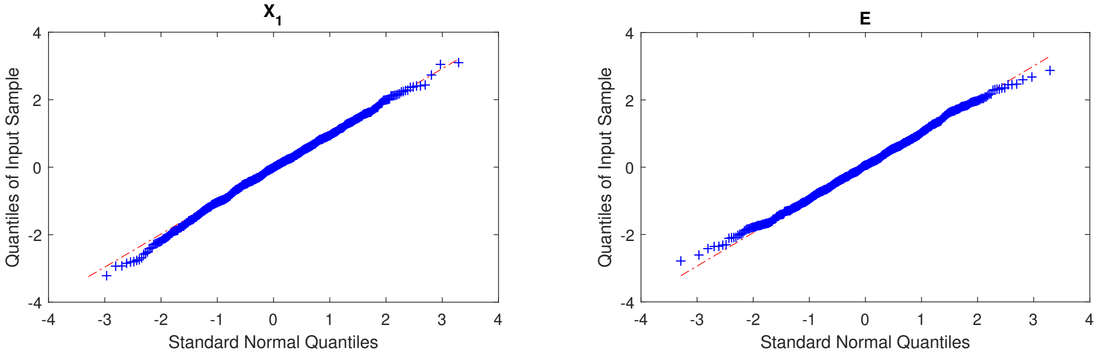

Figure 3: Quantiles of a random sample of 1,000 standardised residuals from the NGR forecasts plotted against the quantiles of a standard normal distribution.
Shown when predicting X_1 (left) and E (right) at a lead time of three days. Similar results are seen for forecasts associated with each regime, and also at other lead times.

at the forecast initialisation time then the training data can be divided into subsets based upon the regime of the atmosphere at the forecast's initialisation time, and a separate set of parameters can be estimated for each regime ($\alpha_r, \beta_r, \sigma_r^2$ for $r = 1, ..., R$) by minimising the logarithmic score over each training subset. A new forecast could then simply be conditioned on one regime:

$$Y|\boldsymbol{x}, r \sim \frac{1}{M} \sum_{m=1}^{M} N(\alpha_r + \beta_r x_m, \sigma_r^2). \tag{26}$$

This method is referred to as RDBMA-init.

Alternatively, if the regime is defined at the validation time, then, since BMA specifies a separate distribution around each ensemble member, every member can be post-processed conditional on its own regime prediction. Members corresponding to the same regime are assumed to be statistically indistinguishable and hence an extension of BMA to include groups of exchangeable ensemble members is implemented (Fraley et al., 2010):

$$Y|\boldsymbol{x} \sim \sum_{r=1}^{R} w_r \sum_{m \in M_r} N(\alpha_r + \beta_r x_m, \sigma_r^2), \tag{27}$$

where M_r denotes the set of indices of the ensemble members that predict regime r, and w_r is the probability that the true system is in that regime at the validation time, with $\sum_{r=1}^{R} w_r = 1$. Fraley et al. (2010) estimate this probability using maximum likelihood via the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977), but here the groups of exchangeable ensemble members are determined by the outputs of the NWP model and hence are not known prior to forecasting. As a result, M_r changes for each ensemble. Taking the weight to be the fraction of ensemble members predicting that regime, $w_r = |M_r|/M$, thus allows the weight to vary for each forecast, providing a more flexible weight function that was found to produce more skilful predictions.

In this case, forecast-observation pairs cannot be assigned to exactly one regime and therefore the parameters corresponding to each regime must be estimated simultaneously. This method is termed RDBMA-val and the associated objective function is

$$\text{LS}(F_{BMA}^{val}, y) = -\log \left[\frac{1}{M} \sum_{r=1}^{R} \sum_{m \in M_r} \phi\Big(\frac{y - \alpha_r - \beta_r x_m}{\sigma_r} \Big) \right], \tag{28}$$

where F_{BMA}^{val} denotes the CDF associated with Equation 27.

3.5.2 Non-homogeneous Gaussian Regression

Recognising the presence of a spread-skill relationship, Gneiting et al. (2005) introduce Non-homogeneous Gaussian Regression to extend the normal linear regression framework to include a variance which is dependent on the spread of the ensemble members. The mean and variance of the predictive distribution are linear functions of the ensemble mean, $\bar{x}$, and variance, s_x^2, respectively. The result is a heteroscedastic distribution of the form

$$Y|\boldsymbol{x} \sim N(\alpha + \beta\bar{x}, \gamma + \delta s_x^2). \tag{29}$$

To estimate the parameters ($\alpha, \beta, \gamma, \delta$; with γ and δ constrained to be positive) in the regression equation, Gneiting et al. (2005) acknowledge that the coefficients should be those that minimise a proper score and therefore propose minimum continuous ranked probability score (CRPS) estimation. This aligns with the optimum score estimation framework in Gneiting and Raftery (2007) (see also Section 2.3.3.3). Gneiting et al. (2005) showed the CRPS for a forecast in the form of a Gaussian predictive distribution to be

$$\text{CRPS}(N(\mu, \sigma^2), y) = \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\}, \tag{30}$$

where $\Phi(\cdot)$ is the standard Gaussian CDF, and the total CRPS is again the average of this score computed over all forecasts in the training data.

Similarly to BMA, if the regime is defined at the initialisation time then each forecast is in either regime A or regime B and the model (labelled RDNGR-init) becomes

$$Y|\boldsymbol{x}, r \sim N(\alpha_r + \beta_r\bar{x}, \gamma_r + \delta_r s_x^2) \tag{31}$$

for $r = 1, ..., R$. Again, parameters are estimated by stratifying the training data set using the regime of the system at the forecast initialisation time and minimising the CRPS separately for each training subset.

However, if the regime is defined at the forecast validation time then it cannot be determined with certainty and hence a probabilistic approach is applied. Let w_r again denote the proportion of ensemble members that predict regime r. Then, a mixture model of R component distributions could be implemented, with weights determined by w_r. The predictive distribution is of the form

$$Y|\boldsymbol{x} \sim \sum_{r=1}^{R} w_r N(\alpha_r + \beta_r\bar{x}, \gamma_r + \delta_r s_x^2). \tag{32}$$

This approach, referred to as RDNGR-val, is essentially a model averaging technique that exploits the regime predictions of the ensemble members to calculate the model weights. The CRPS for a forecast in the form of a mixture of Gaussian distributions is

$$\begin{aligned}\text{CRPS}\left(\sum_{j=1}^{J} p_j N(\mu_j, \sigma_j^2), y\right) = \\ \sum_{j=1}^{J} p_j A(y - \mu_j, \sigma_j^2) - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J} p_j p_k A(\mu_j - \mu_k, \sigma_j^2 + \sigma_k^2),\end{aligned} \tag{33}$$

where

$$A(\lambda, \eta^2) = 2\eta\phi\left(\frac{\lambda}{\eta}\right) + \lambda\left[2\Phi\left(\frac{\lambda}{\eta}\right) - 1\right]$$

(Grimit et al., 2006). For the conditional distribution in Equation 32, J is equal to the number of regimes R, and the weights p_j are given by the proportion of ensemble members that predict each regime, w_r.

3.5.3 Forecast verification

These statistical post-processing methods are applied to a sample of point forecasts to obtain a predictive distribution conditional on the ensemble output. Forecasters have come to seek predictive distributions that are sharp, subject to being calibrated, and both of these qualities can be assessed by verifying forecasts using proper scoring rules (Gneiting and Raftery, 2007). In the following section, the CRPS is used to verify forecasts. NGR forecasts are assessed using the same loss function with which parameters were optimised in the training data, while Equation 33 can also be used to evaluate the forecast distributions generated from BMA. Although this might appear to favour NGR, since parameters are estimated using the same score that is used to verify the forecasts, similar results are obtained when using the logarithmic score to assess forecasts, and also when BMA parameters are optimised using minimum CRPS estimation.

These scores outline the overall forecast performance but concern lies more on the improvement gained from the new methodology than on the raw scores themselves. Therefore the continuous ranked probability skill score (CRPSS) is also applied. Whereas skill scores are typically implemented with a simple benchmark such as climatology (Mason, 2004), the reference forecast is taken here to be the equivalent forecast obtained via NGR or BMA at the same lead time. For example, if we let H denote the predictive distribution obtained from regime-dependent post-processing, F denote that obtained from conventional post-processing, and y the corresponding observation, then

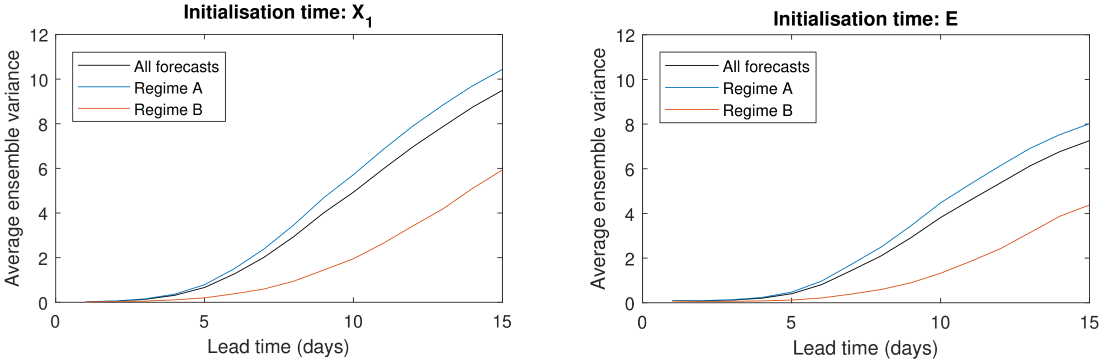

Figure 4: Average ensemble variance for forecasts of X_1 (left) and E (right) against lead time, shown when each regime occurs at the forecast initialisation time.

the continuous ranked probability skill score (CRPSS) is

$$\text{CRPSS}(H, F, y) = \frac{\langle \text{CRPS}(F, y) \rangle - \langle \text{CRPS}(H, y) \rangle}{\langle \text{CRPS}(F, y) \rangle} = 1 - \frac{\langle \text{CRPS}(H, y) \rangle}{\langle \text{CRPS}(F, y) \rangle}, \tag{34}$$

with $\langle \cdot \rangle$ denoting the average CRPS over forecasts in the test data set (Wilks, 2019, see also Section 2.4.4). The skill score in this case can thus be interpreted as the relative improvement in score upon current post-processing methods, gained from regime-dependent post-processing.

3.6 Results

3.6.1 Forecasting X_1

The statistical properties of the forecasts indicate that there are in fact disparities in the forecast behaviour between the two regimes. Figure 4 shows that the ensemble variance, computed from forecasts in the training data, is much smaller on average when the system resides in regime B than in regime A. This is true when predicting X_1 or E. Such differences in the variance suggest either that weather events are more predictable, or that the ensemble forecasts suffer more from overconfidence, when the system resides in regime B.

Ensemble forecasts assume that the ensemble members arise from the same generation mechanism as the observation, and hence the rank of the verification when pooled with the ensemble members should be uniformly distributed. This assumption can be evaluated by using verification rank histograms to visualise the distribution of the ranks across all forecasts in the test data (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1997b). Rank histograms displayed in Figure 5 indicate that the raw ensemble forecasts are highly overconfident, with observations falling outside the range

X_1		α_r	β_r	γ_r	δ_r
NGR		0.301	0.884	4.476	2.713
RDNGR-init	Reg. A	0.411	0.855	6.028	2.202
	Reg. B	-0.004	0.966	1.510	6.639
RDNGR-val	Reg. A	0.437	0.857	6.071	2.232
	Reg. B	-0.140	0.976	1.212	4.414

X_1		α_r	β_r	σ_r^2
BMA		0.450	0.861	8.260
RDBMA-init	Reg. A	0.569	0.831	9.499
	Reg. B	0.091	0.961	4.065
RDBMA-val	Reg. A	0.603	0.829	9.583
	Reg. B	-0.106	0.985	2.831

Table 3: Post-processing parameters for NGR, BMA, and for both regime-dependent extensions at a lead time of one week when forecasting X_1.

of ensemble members for the vast majority of forecasts. This is yet more prevalent for those initialised in regime B.

Figure 5 also displays Probability Integral Transform (PIT) histograms (Dawid, 1984) for the predictive distributions issued by NGR and RDNGR-init for the same lead time. PIT histograms record the frequency with which values of the forecast CDF evaluated at the verification, $F(y)$, fall into a finite number of equally-sized bins. To ensure comparability between the rank and PIT histograms, 21 bins between zero and one were chosen. Likewise, a uniform PIT histogram implies calibrated forecasts. The PIT histograms show that post-processing the forecasts using NGR yields considerably more uniform histograms, and hence considerably better-calibrated forecasts, than the raw ensembles. However, Hamill (2001) demonstrates that uniform rank histograms can be obtained from a combination of poorly-calibrated forecasts, emphasising that the uniformity of rank histograms is a necessary but not sufficient condition for reliable predictions. In this case, forecasts in regime B become largely overdispersed as a result of post-processing, indicating the forecasts are not calibrated conditional on the initial regime. Estimating a new set of parameters for forecasts initialised in regime B, as in RDNGR-init, reduces the underconfidence of these forecasts.

Table 3 presents the parameters estimated over the training data at a lead time of one week for both NGR and BMA, and all regime-dependent extensions. Regardless of the time at which the regime of the forecast is defined, the post-processing parameters are noticeably different between the two regimes. For BMA, the parameter controlling the variance decreases dramatically when forecasting an event in regime B, supporting the belief that the system is more predictable in this regime. The regime A parameters, on the other hand, are generally similar to those obtained via standard post-processing.

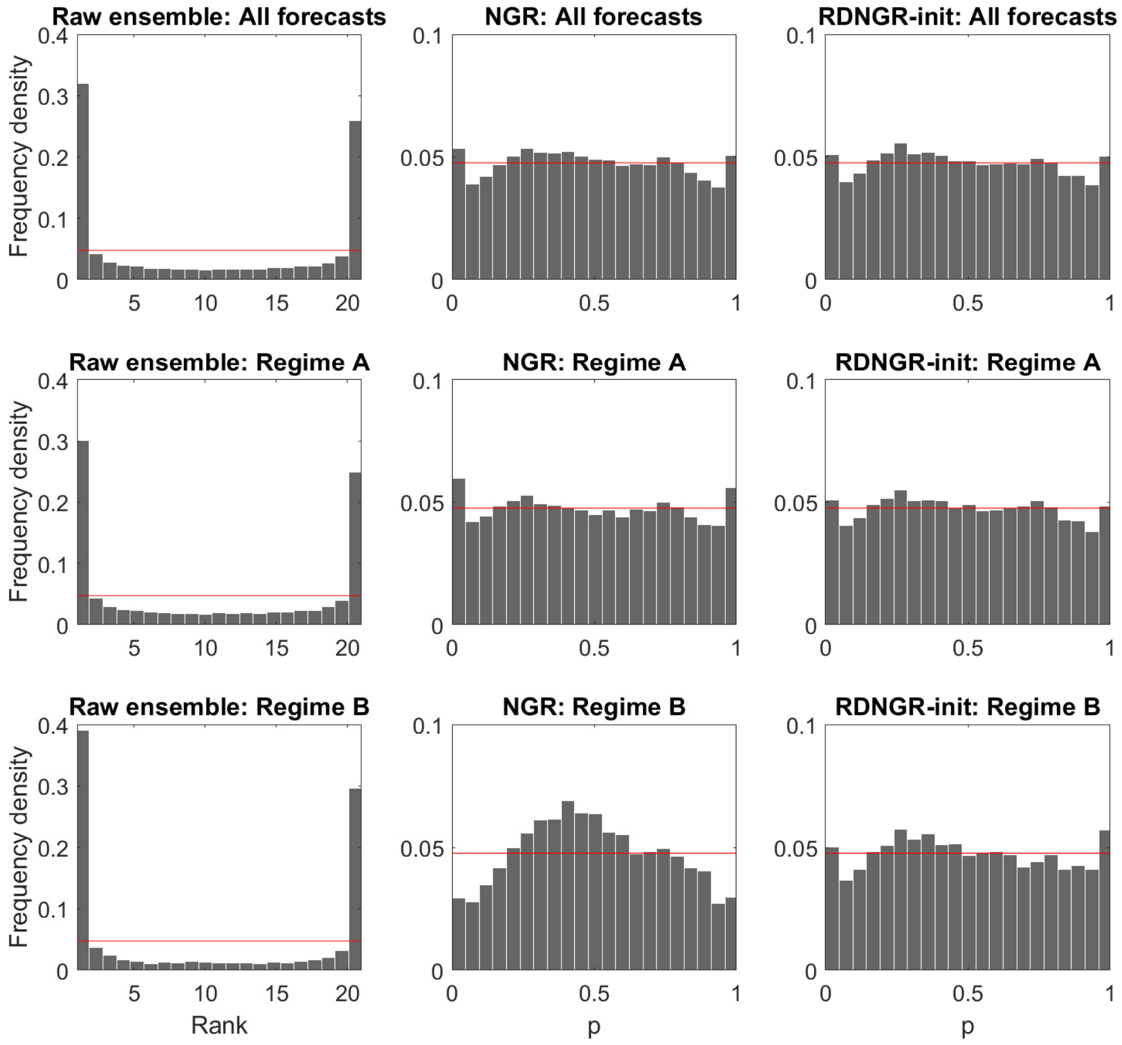

Figure 5: Rank histograms for the raw ensemble forecasts and PIT histograms for NGR and RDNGR-init forecasts at a lead time of one week.

Histograms are displayed for forecasts in each regime at initialisation time. The red line corresponds to perfect uniformity.

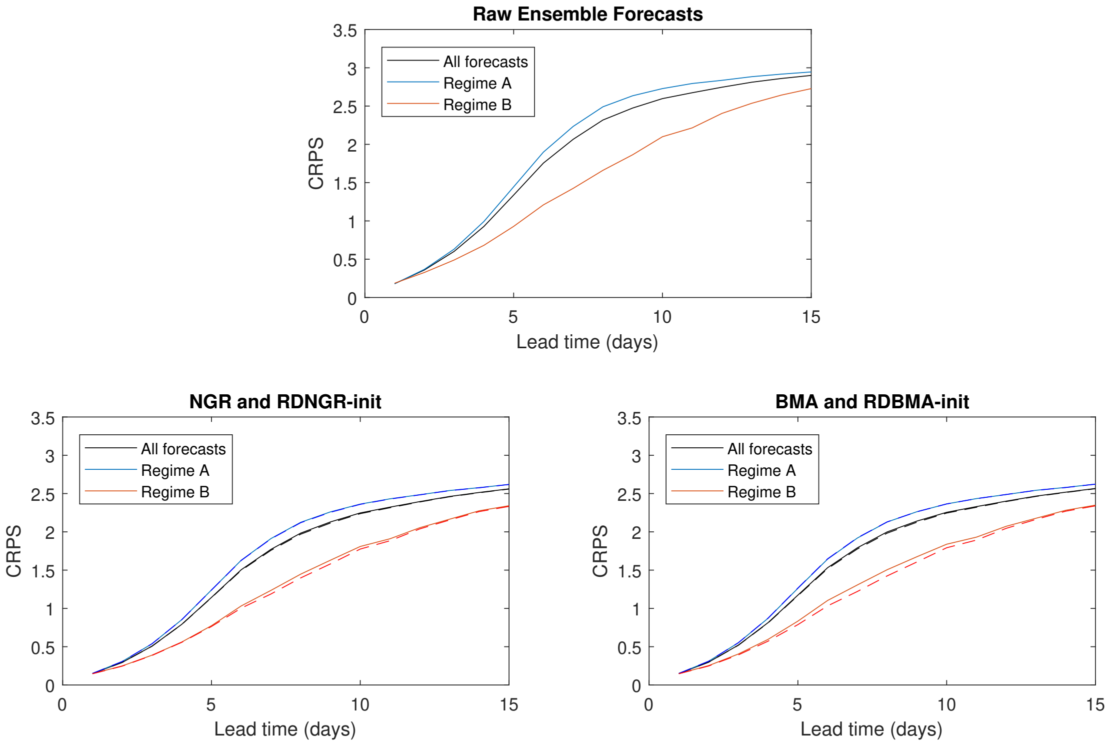

Figure 6: CRPS for the raw ensemble forecasts of X_1 and for NGR and BMA (unbroken), and RDNGR-init and RDBMA-init (dashed) against lead time when the forecast is initialised in each regime.

This is not surprising given that the system spends 80% of its time in regime A (Table 2) and hence the vast majority of forecasts in the training data are defined to be in this regime.

The regime-dependent NGR methods appear to adjust the variance of their predictive distribution differently for the two regimes. The modest ensemble spread in regime B forecasts is augmented by a larger variance inflation factor δ, whereas the variance of regime A forecasts is increased by a larger additive, or nudging, parameter γ, thus implying the presence of a stronger spread-skill relationship in regime B. There are also slight differences between the parameters dictating the forecast mean; values of β close to one, along with additive constants α close to zero, suggest the raw ensemble mean contains more information when the forecast is defined to be in regime B. These differences therefore support our theory that the relationship between forecasts and observations changes depending on the system's regime. It is difficult, however, to deduce the time at which this relationship is most varied, since there do not appear to be large discrepancies between the regime-dependent approaches.

Having seen how the models are behaving in the different regimes, attention is turned to formally assessing the forecasts. Figure 6 exhibits the CRPS against lead time for the raw ensembles and for NGR, RDNGR-init, BMA and RDBMA-init forecasts, along with the breakdown of those defined to be in regime A and B at initialisation time. The scores are considerably lower for forecasts initialised in regime B than they are for regime A, but since only 20% of forecasts are attributed to regime B, the score calculated across all forecasts is more similar to that for forecasts in regime A. The post-processed forecasts unsurprisingly yield scores much lower than those for the raw ensemble forecasts, while the extra improvements gained from regime-dependent post-processing are noticeable in regime B but appear negligible for forecasts initialised in regime A, rendering the overall improvement relatively unpronounced. The CRPS for all methods at a lead time of seven days is displayed in Table 4.

Figure 7 displays the skill score for the regime-dependent forecasts defined to be in each regime at the initialisation time, where the conventional NGR and BMA forecasts are used as baselines. Figure 7 further reinforces what has already been seen: regime B forecasts improve by as much as 6% upon standard post-processing, while those initialised in regime A experience little improvement. Regime B forecasts are thus responsible for the majority of improvement but the dominance of regime A means the relatively large improvements seen in regime B forecasts account for only 20% of the

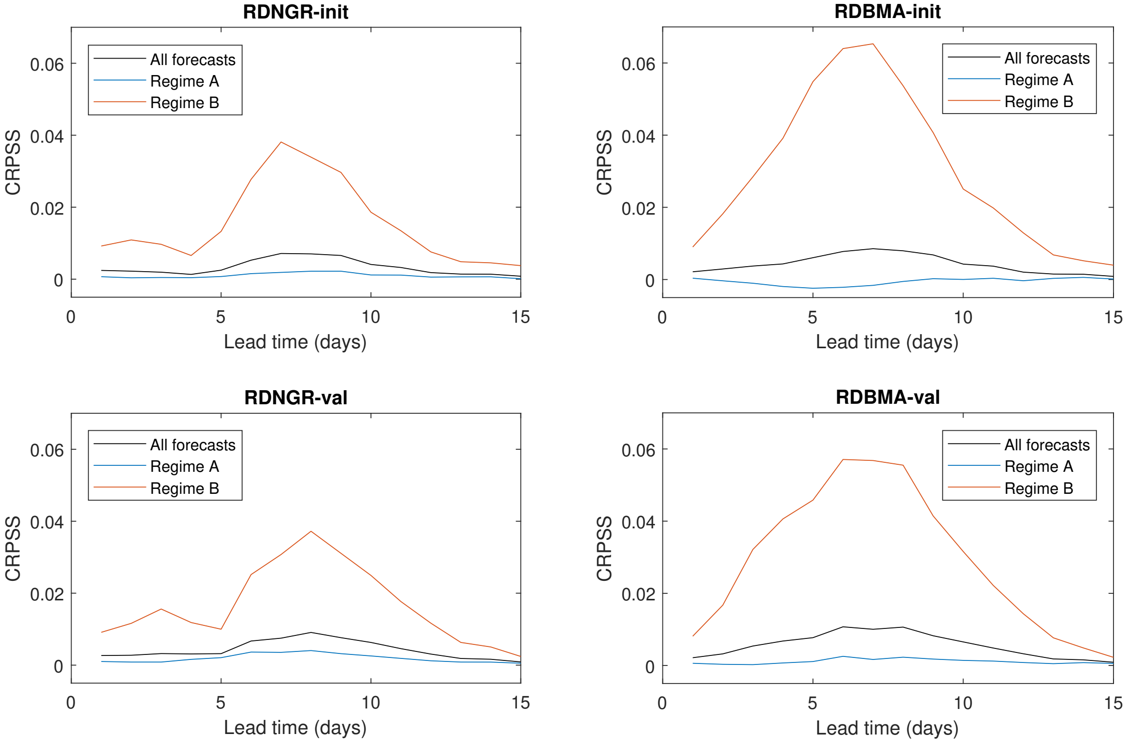

Figure 7: CRPSS against lead time for both regime-dependent NGR and both regime-dependent BMA approaches using NGR and BMA, respectively, as a reference forecast when predicting X_1.

X_1	Total	Regime A	Regime B
Raw	2.072 (0.007)	2.241 (0.008)	1.429 (0.011)
NGR	1.779 (0.005)	1.921 (0.006)	1.238 (0.008)
RDNGR-init	1.768 (0.005)	1.916 (0.006)	1.202 (0.009)
RDNGR-val	1.766 (0.005)	1.914 (0.006)	1.202 (0.008)
BMA	1.796 (0.005)	1.926 (0.006)	1.301 (0.008)
RDBMA-init	1.782 (0.005)	1.927 (0.006)	1.227 (0.009)
RDBMA-val	1.779 (0.005)	1.922 (0.006)	1.232 (0.008)

Table 4: CRPS for all forecasts of X_1 and the breakdown between those identified to be in regime A and regime B at initialisation time.
The scores are shown for the raw ensembles and for NGR, BMA post-processed forecasts, and all regime-dependent extensions, at a lead time of one week. The corresponding standard errors are calculated using the central limit theorem and are displayed in brackets next to the score.

total improvement. Therefore, the maximum overall percentage improvement is little over 1%.

3.6.2 Forecasting E

Figure 8 displays the evolution of BMA and RDBMA-init parameters over forecast lead time, when E is the predictand. The variance coefficients exhibit similar behaviour to before, with σ^2 significantly lower when the atmosphere is in regime B at initialisation time. However, as seen in Figure 2, the location of the empirical distribution of E in regime A is different to that in regime B, which is not the case for observations of X_1. There are now much larger distinctions in the location parameters, α and β, between the regimes, indicating the NWP model exhibits both spread and location biases that vary with the regime.

As a result, much larger improvements are gained from regime-dependent post-processing. Figure 9 displays the CRPS for the raw ensemble forecasts, NGR, BMA and their extensions that utilise the regime at the forecast validation time. The forecasts issued by RDNGR-val and RDBMA-val perform considerably better (in terms of the CRPS) than those generated by NGR and BMA, respectively, particularly in regime B. This improvement is also maintained for forecasts at longer lead times. The corresponding skill scores are displayed in Figure 10. When the regime is defined at the validation time, forecasts in regime B can improve by almost as much as 20% upon NGR and BMA forecasts, with overall improvements close to 7% at lead times between six and nine days.

Initially it was believed that RDBMA-val would have a slight advantage over its NGR counterpart since it post-processes each ensemble member separately, not com-

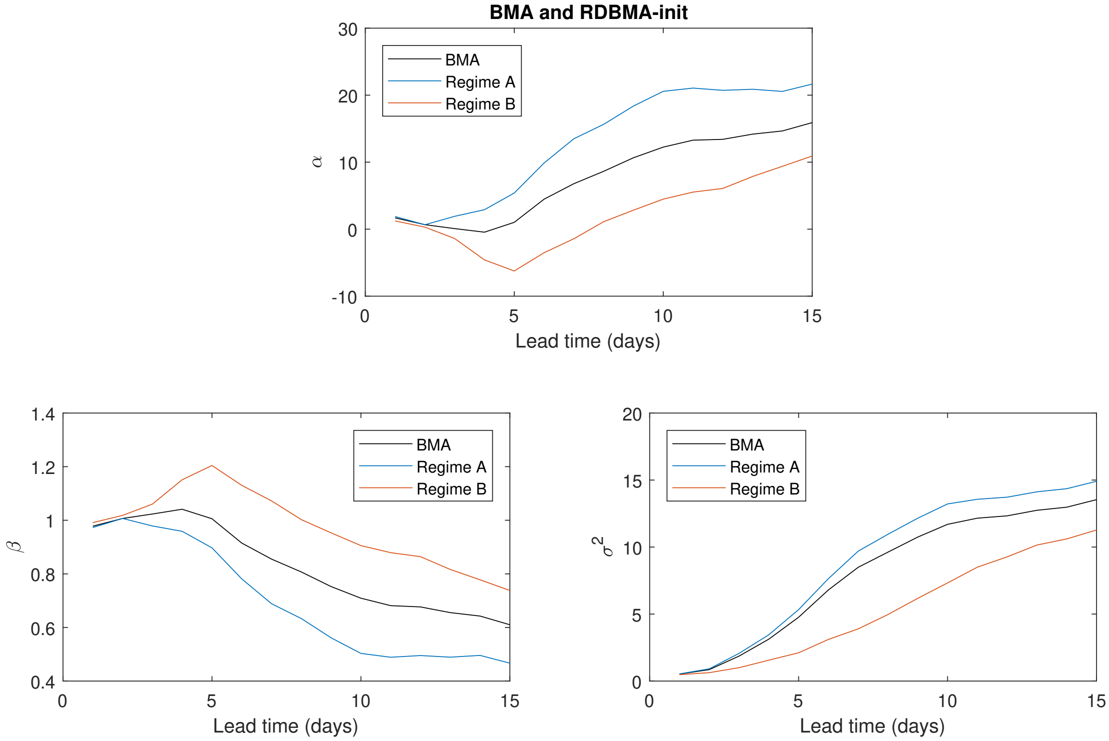

Figure 8: BMA parameters against lead time when forecasting E. RDBMA-init coefficients are also shown for forecasts initialised in each regime.

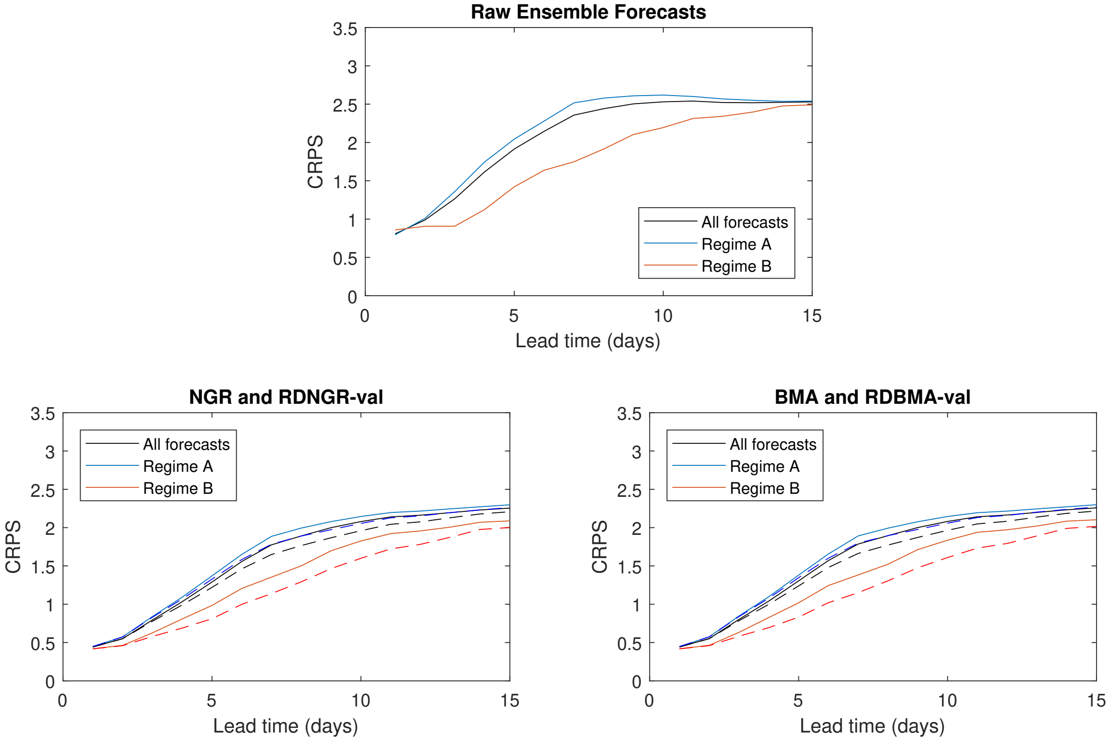

Figure 9: CRPS for the raw ensemble forecasts of E and for NGR and BMA (unbroken), and RDNGR-val and RDBMA-val (dashed) against lead time when the forecast is initialised in each regime.

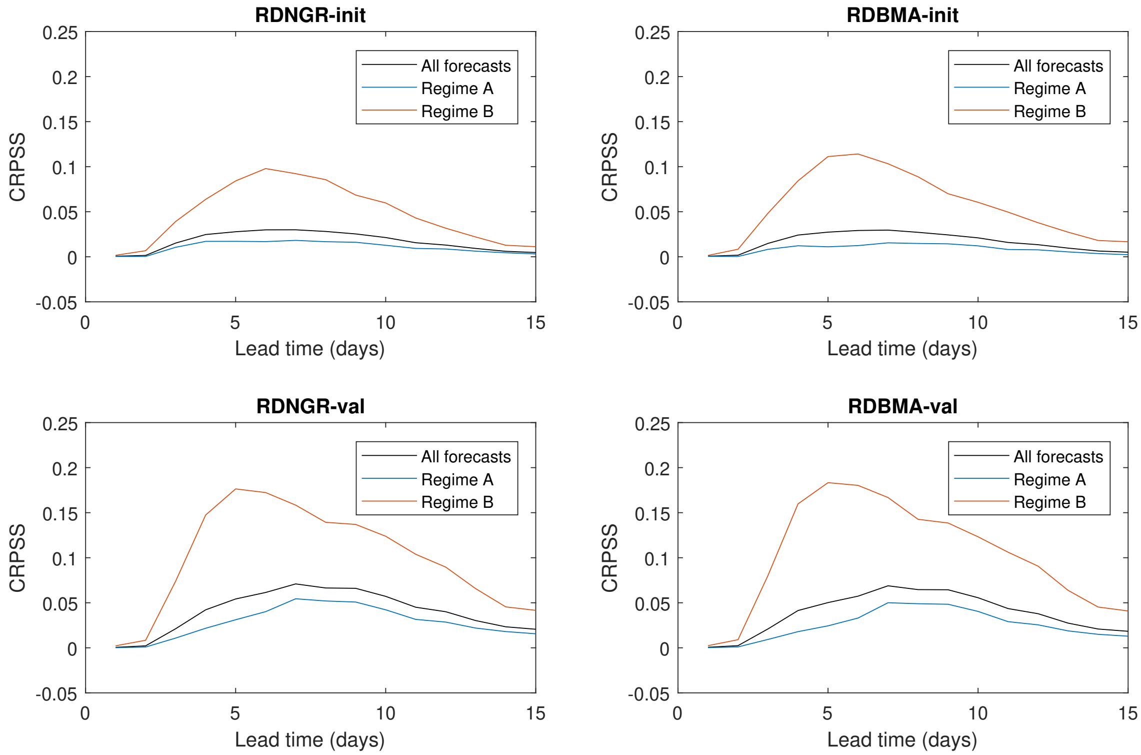

Figure 10: CRPSS against lead time for both regime-dependent NGR and both regime-dependent BMA approaches using NGR and BMA, respectively, as a reference forecast when predicting E.

pressing all the regime information into a single weight. Figures 9 and 10, however, suggest the improvements are similar for the two methods. When using the regime at the initialisation time, if the forecasts issued by BMA performed better when the system resided in one regime but NGR forecasts were preferable in the other, then it would be possible to calibrate subsets of forecasts using separate post-processing methods depending on the regime (e.g. apply NGR to all forecasts in regime A and BMA to all forecasts in regime B).

Given that regime A dominates the upper tail of the response distribution of E (Figure 2) and regime B the lower, we might also expect regime-dependent post-processing to produce more informative predictions of extreme weather events. The Brier score, or mean squared error of a probability forecast for a binary response (Brier, 1950), can be used to assess the probability of the response falling above or below some threshold of the data. Table 5 displays the Brier score, at lead times of three, five and ten days, for the predicted probability of the verification falling below the first percentile of all observations in the training data. This is hence a measure of the forecasts' performance when predicting the occurrence of extremely low values of E. Again, when the regime is defined at the validation time regime-dependent statistical post-processing noticeably

E	3 days	5 days	10 days
Raw	4.59 (0.23)	7.14 (0.32)	12.25 (0.42)
NGR	4.08 (0.20)	6.52 (0.28)	11.44 (0.44)
NGR-init	3.46 (0.18)	5.79 (0.26)	11.34 (0.44)
NGR-val	3.30 (0.17)	4.95 (0.22)	10.94 (0.40)
BMA	4.09 (0.20)	6.52 (0.27)	11.42 (0.44)
RDBMA-init	3.49 (0.18)	5.61 (0.24)	11.34 (0.44)
RDBMA-val	3.31 (0.18)	4.91 (0.22)	10.92 (0.40)

Table 5: Brier score for forecasts of the occurrence of extremely low values of E, for NGR, BMA and both regime-dependent extensions.
Extremely low values correspond to values below 29.1, the first percentile of the observations. Scores are shown at lead times of three, five and ten days, with the associated standard errors in brackets alongside. All values have been scaled by 10^3.

improves upon current post-processing approaches. Since the marginal distribution of X_1 varies less between the regimes, similar forecasts of extremely low values of X_1 exhibit less improvement, comparable to results seen for all forecasts in Figures 6 and 7.

3.7 Discussion

This chapter acknowledges that the inability to distinguish between distinct relationships linking the NWP model and the atmosphere is a potential weakness of statistical techniques of calibrating ensemble forecasts. In particular, it is proposed that, under certain circumstances, the relationship between the model and the atmosphere changes, and if such circumstances are identified, then post-processing forecasts conditional on this extra information could yield more informative prognoses. Although the methodology presented here extends to other appropriate and justifiable conditions, past literature suggests that the occurrence of particular weather regimes is an example of such a circumstance. This chapter therefore investigates how best to utilise regime information when post-processing.

The preferred approach is to implement a mixture of predictive distributions, with a separate component distribution designed to calibrate forecasts assigned to each regime. Two different methods are used to assign forecasts to regimes. The first defines the regime as that which materialises at the forecast initialisation time, while the second uses that at the forecast validation time. For the former choice, the weight in the mixture distribution is an indicator function. In this case, the forecast-observation pairs in the training data are stratified depending on their associated regime, and the component forecast distributions are fit separately over the resulting training subsets.

The regime at the forecast validation time, on the other hand, cannot be known with certainty. As such, the weights in the mixture model should be chosen to reflect the probability that the system resides in each regime. There are numerous ways to obtain a probabilistic forecast for the prevailing weather regime, and the proportion of ensemble members that forecast each regime is used here to define the mixture model weights. In this case, a forecast-observation pair in the training data cannot be attributed to exactly one regime, and hence, rather than stratifying the data into MECE training subsets, all coefficients for the mixture distribution are estimated simultaneously. Although the predictions of the regime made by the ensemble forecast may themselves exhibit biases, these could easily be addressed using statistical methods, akin to an application of statistical post-processing to the regime forecast. Future work could consider the benefit of such an approach.

Two extensions of both Non-homogeneous Gaussian Regression and Bayesian Model Averaging are presented that are suitable for the regime paradigm, one corresponding to each way of assigning forecasts to regimes, and these methods are trialled in the Lorenz (1996) system, a highly idealised surrogate of the atmosphere that is known to exhibit regime-like behaviour (Christensen et al., 2015). In particular, the system favours two states: regime A and regime B. The results were compared for forecasts made for two different variables, X_1, a state variable of the system, and E, proportional to the total energy in the system (Lorenz, 1996). The distribution of X_1 does not change considerably between the two regimes, whereas the opposite is true for E.

For both variables, regime-dependent post-processing methods improve the performance of forecasts relative to conventional approaches. Since regime A occurs on the majority of occasions, regime-dependent post-processing methods have less of an effect on forecasts defined to be in this regime, whereas forecasts corresponding to regime B improve substantially. This is particularly the case for forecasts of E, which are found to exhibit systematic regime-dependent errors in both the location and dispersion of the ensemble. Forecasts of more extreme values of E also benefit significantly from regime-dependent post-processing, since the lower tail of the empirical distribution is dominated by regime B, and the upper tail by regime A. Regime-dependent approaches therefore address the tendency of conventional post-processing methods to focus on the bulk of the data, which has been found to result at times in a decrease in the predictive skill of forecasts for extreme weather events (Pantillon et al., 2018).

Furthermore, regardless of the predictand, the ensemble forecasts perform well at short lead times, being highly concentrated around the ensemble mean (Figure 4), resulting in CRPS values close to the optimum value of zero (Figures 6 and 9). In this case, the ensemble forecasts already capture the effect the regimes have on the predictands, and, as such, the improvements gained by regime-dependent post-processing are small. As the lead time increases, however, the performance of the raw ensemble worsens, and errors arise in the forecast owing to the regimes. Therefore, the benefit of regime-dependent post-processing is considerable for forecasts at lead times between roughly four and eight days. After this, the ability of the ensemble members to identify which regime occurs at the forecast validation time deteriorates. As a result, the conditional distribution of the outcome given the ensemble forecast begins to depend less on the forecast regime, and hence the post-processing coefficients corresponding to each regime become progressively more similar to those estimated over the entire training data set. The improvements gained by regime-dependent post-processing therefore tend towards zero as lead time increases. Due to their additional flexibility, the regime-dependent methods should, in theory, always perform at least as well as conventional approaches, provided there is sufficient training data on which to reliably estimate the post-processing parameters.

The ensemble variances in Figure 4 indicate that the raw ensemble forecast captures the flow-dependent uncertainty in the system, with ensembles considerably less dispersed in regime B than in regime A. Furthermore, the regime-dependent approaches were also applied within the Model Output Statistics post-processing framework, which is equivalent to NGR without incorporating the ensemble spread as a predictor for the forecast variance, so that the δ coefficient in Equation 29 is constrained to equal zero (Glahn and Lowry, 1972; Glahn et al., 2009). The forecasts issued by Model Output Statistics benefited more from the regime information than those generated using NGR, reinforcing the idea that the spread of the ensemble captures some of the uncertainty present in the forecast situation owing to the prevailing weather regime. Moreover, notwithstanding the tendency of the numerical weather model (Equation 22) to prefer regime A (Table 2), the ensemble members provide a skilful prediction of the future regime for all lead times considered here. The regime at the forecast initialisation time, on the other hand, is less adept at predicting the future regime, despite the persistence of the regimes in this system (not shown). As a result, the improvements in CRPS from regime-dependent post-processing tend to be larger when the regime of the forecast is defined at the forecast validation time, rather than at the initialisation time.

The results presented in this chapter therefore indicate that regime-dependent post-processing can significantly enhance weather forecasts, especially when the variable being forecast exhibits a pronounced dependence on the regimes. In particular, if severe weather events occur more frequently in some regimes than others, such as extreme temperatures during prolonged blocking episodes, then incorporating this regime-dependency when calibrating forecasts could lead to refined predictions of these extreme events. Larger improvements tend to occur when the forecast accurately predicts the regime at the forecast validation time, suggesting the regime-dependent forecast errors in this system arise predominantly from the numerical weather model incorrectly simulating the relationship between the regime and the variable being forecast. It is not clear, however, how this will translate to forecasts issued using operational numerical weather models (i.e. not in a highly idealised system), and further investigation into this is necessary.

Operational forecasters often suffer from a lack of historical data available, and it has become common to use a sliding training window to estimate parameters. These windows consist of forecast-observation pairs from a relatively short number of days directly preceding the time of forecasting. The choice of the length of this window is a compromise between using enough data from which reliable parameter estimates can be obtained and using a length that is small enough for the training window to reflect the seasonality and recent behaviour of the weather. It could be argued that knowing how the model behaves in different regimes is more valuable when estimating model coefficients than knowing how forecasts behaved more recently in potentially very different atmospheric conditions. For example, if the atmosphere resides in an anticyclonic regime then the model biases will likely be similar to occasions in previous years when this pattern has occurred, rather than to the errors, say, twenty days prior to forecasting when a different regime were present. The regime-dependent framework may thus be better suited to retrospective forecasting (reforecasting) approaches that run high resolution weather models from historical reanalyses to generate a large number of hindcasts (Hamill et al., 2004).

Moreover, estimating all regime-dependent parameters simultaneously can be significantly more computationally demanding than estimating BMA and NGR coefficients. This has been identified in previous implementations of mixture models in a post-processing context (Baran and Lerch, 2016, 2018). On the other hand, despite the statistical models being more elaborate, implementing RDBMA-init and RDNGR-init was no more computationally expensive than the standard post-processing approaches in this study. Nonetheless, given that statistical post-processing is typically performed off-line, after having integrated the numerical weather model, none of the regime-dependent approaches considered here are expected to be prohibitively expensive. Regardless, these approaches estimate more parameters and hence require larger amounts of training data in order to attain reliable parameter estimates. Methods that can account for parameter uncertainty in the post-processing models (Siegert et al., 2016a), or augment the training data (Lerch and Baran, 2017; Hamill et al., 2017) are thus particularly desirable in the regime paradigm. An excessively large amount of training data was used in this simulation study to remove the necessity of such methods, though smaller archives of data drew the same conclusions.

Finally, we note that this study was repeated using a stochastic alternative to the numerical weather model considered here. The primary goal of Christensen et al. (2015) was to study the effects that stochastic parametrisations have on capturing the regime structure of the Lorenz (1996) system. The result was that the introduction of a red-noise stochastic parameter to the deterministic NWP model (Equation 22) allows a good representation of the regimes. Similar patterns emerged to those identified here, though the improvements were slightly more pronounced when using the deterministic model: the method was better at correcting poor forecasts than improving the higher-quality model. Indeed, this behaviour is intuitive for atmospheric data; the circulation dictates the weather that we experience, so the distribution of the observations would be expected to vary between the regimes, and hence we would anticipate more improvement from regime-dependent post-processing if the NWP model output did not do the same.

4 Recalibrating wind speed forecasts using regime-dependent Ensemble Model Output Statistics

4.1 Introduction

The previous chapter presents a motivating example in which the regime-dependent framework greatly outperforms established post-processing approaches when applied to simulated data from a highly idealised atmospheric model. Hence, the aim of this chapter is to investigate how the framework performs in more realistic settings. In particular, we consider a regime-dependent extension of a method to post-process wind speed forecasts, both in a simulation study involving a quasigeostrophic model of the Northern Hemisphere, and using the output from a recently operational ensemble forecasting system over the Euro-Atlantic region.

High-quality forecasts of wind speed are particularly valuable due to their application in decision making in areas such as transportation, insurance, and renewable energy production. Therefore, several statistical post-processing methods have been proposed to deal with systematic errors present in wind speed forecasts. Particular examples include quantile regression (Bremnes, 2004, 2019) and implementations of Bayesian Model Averaging (BMA) and Ensemble Model Output Statistics (EMOS) with various choices of parametric family: truncated normal (Thorarinsdottir and Gneiting, 2010; Baran, 2014), gamma (Scheuerer and Möller, 2015; Sloughter et al., 2010; Eide et al., 2017), and truncated logistic distributions (Messner et al., 2014; Scheuerer and Möller, 2015), for example. Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) blend together different choices of the parametric family in an attempt to improve the upper tail of wind speed forecasts, and the resulting predictive distributions are particular examples of the mixture distributions considered by Baran and Lerch (2016) and Baran and Lerch (2018).

Since the aim of forecast recalibration is to alleviate systematic biases in the dynamical model output, it is common to use only the ensemble forecast for the variable being predicted as an input variable when post-processing. Recently, however, techniques have been proposed that utilise more predictors, highlighting the potentially useful information that can be gained from other sources. Scheuerer (2014) and Scheuerer and Hamill (2015a), for example, exploit predictions at neighbouring grid points when recalibrating precipitation forecasts, while Eide et al. (2017) employ wind direction as an additional predictor for wind speed. More data-driven approaches have also been proposed that can deal with a large set of possible inputs, and automatically select those most relevant for post-processing (Taillardat et al., 2016; Messner et al., 2017;

Rasp and Lerch, 2018).

The underlying reason for adding predictors is that the additional variables provide helpful indications as to when the relationship between the forecast and the observation might vary. It may be the case that forecast accuracy is affected by the weather situation at hand. Weather forecasters often adjust their predictions depending on the prevailing large-scale atmospheric flow (Roebber, 1998) and incorporating the flow directly into forecast recalibration methods serves as a way of automating this procedure. Synoptic-scale patterns in the atmosphere's circulation can also explain relationships between certain weather variables and locations. Integrating the circulation into post-processing therefore allows information from alternative weather variables to be utilised, without including them directly in the calibration.

The atmosphere's circulation is intimately connected to the Earth's winds and therefore forecasts of wind speed might be susceptible to improvements if weather regime information were incorporated into the post-processing. The previous chapter introduced regime-dependent statistical post-processing, proposing that if statistical techniques can specify a probability model for the regime, then current post-processing methods can be conditioned on the underlying weather regime:

$$f(y|\boldsymbol{x}) = \sum_{r=1}^{R} w_r f_r(y|\boldsymbol{x}, r), \tag{35}$$

where w_r is the probability of residing in regime r and $f(y|\boldsymbol{x})$ is the conditional distribution of the predictand given the ensemble forecast $\boldsymbol{x} = (x_1, ..., x_M)$. The forecast in this case takes the form of a predictive distribution, but, rather than specifying just one forecast distribution, a separate distribution must be estimated for each regime. The regime-dependent framework would then be expected to be beneficial if the component predictive distributions $f_r(y|\boldsymbol{x}, r)$ were to change between the regimes.

The post-processing framework, including the regime-dependent approach considered in this chapter, is introduced in the following section. The post-processing methods are first implemented in a three-layer quasigeostrophic (QG) model of the Northern Hemisphere in Section 4.3. The QG model used here is sufficiently realistic that it is capable of generating atmospheric patterns that are present in climate reanalyses, but is simple enough that a large amount of data can be simulated, allowing an extensive investigation of regime-dependent approaches. In Section 4.4, the same approach is trialled on retrospective wind speed forecasts (reforecasts) over the Euro-Atlantic region, taken from the National Centers for Environmental Prediction Global Ensemble Forecasting System (GEFS; Hamill et al., 2013). The GEFS reforecasts are generated from a higher resolution model than that used in the QG setting, yet still provide sufficient data from which regime-dependent forecast distributions can reliably be constructed and assessed. Section 4.5 concludes and discusses the results presented throughout this chapter.

4.2 Methodology

4.2.1 Statistical post-processing

To capture the relationship between the model and the atmosphere, statistical post-processing relies on a set of historical forecasts and observations from which predictive distributions can be estimated. This training set consists of pairs of data $(\boldsymbol{x}, y)$, where $\boldsymbol{x} = (x_1, ..., x_M)$ denotes an ensemble forecast comprised of M members, and y is the corresponding verification. Regime-dependent post-processing methods extend this such that the training data pairs become triples of the form $(\boldsymbol{x}, y, \rho)$, where ρ represents some information regarding the atmospheric flow associated with that forecast and observation. This could be one weather regime, the probabilities of residing in each identified regime, or a continuous measure of the atmospheric flow, for example. Post-processing can then utilise this additional information. In the previous chapter, we discuss ways of including the flow in forecast recalibration, arguing that partitioning the phase space into a discrete number of regimes can allow for more flexible forecast distributions. This approach is similarly considered here.

Thorarinsdottir and Gneiting (2010) introduce an Ensemble Model Output Statistics (EMOS) approach that extends truncated regression models to include a non-constant variance term. The method suggests that, given an ensemble forecast, the observed wind speed is a realisation of a random variable Y that follows a normal distribution truncated below at zero:

$$Y|\boldsymbol{x} \sim N_0(\alpha + \beta\bar{x}, \gamma + \delta s_x^2). \tag{36}$$

The location and spread of the distribution are then linear functions of the ensemble mean $\bar{x}$ and ensemble variance s_x^2 respectively. This approach was found here to outperform alternative families of statistical distributions.

We employ Equation 36 in the regime-dependent framework by using a mixture of truncated normal forecast distributions that depend on the coinciding weather regime:

$$Y|\boldsymbol{x} \sim \sum_{r=1}^{R} w_r N_0(\alpha_r + \beta_r\bar{x}, \gamma_r + \delta_r s_x^2), \tag{37}$$

where w_r represents the probability of the atmosphere residing in regime r at the forecast validation time. This method involves estimating post-processing parameters

		True Regime			
		NAO+	NAO-	AR	EB
Initial Regime	NAO+	0.740	0.070	0.094	0.095
	NAO-	0.105	0.664	0.147	0.083
	AR	0.163	0.137	0.565	0.134
	EB	0.102	0.129	0.113	0.657

Table 6: Matrix of conditional probabilities of each regime occurring 48 hours after observing a given regime at the forecast initialisation time.
The regimes under consideration are those introduced in Section 4.4.1. As lead time increases every row tends to the climatological frequencies of the regimes. NAO+ and NAO- represent, respectively, the positive and negative phases of the North Atlantic Oscillation, AR denotes an Atlantic Ridge, while EB is European Blocking.

$(\alpha, \beta, \gamma, \delta)$ for each of the regimes. The weight, on the other hand, is a function of time, rather than a coefficient to be estimated.

4.2.2 Mixture model weights

It is, however, necessary to define these mixture model weight functions, w_r. The motivation for regime-dependent approaches assumes that there are differences in model biases that depend on the prevailing weather regime. Results in the previous chapter highlighted that, since the weather that materialises is dependent on the atmospheric regime, the weights should provide probabilities that the atmosphere will reside in each regime at the forecast validation time. The weights can thus be thought of as predictions of the future atmospheric state.

Three choices for the weight are compared here. A first choice defines the weights by a persistence forecast: if s is the regime present at the forecast initialisation time then $w_r = 1$ when $r = s$, and $w_r = 0$ when $r \neq s$. These will be called ‘initial regime’ weights. The disadvantage of this approach is that as forecast horizon increases, so does the probability of transitioning to another regime. The initial regime would thus not be representative of the atmospheric conditions at the validation time. When the forecast lead time is long relative to the regime persistence times, model biases would not be expected to vary depending on the initial regime and hence the regime-dependent mixture model would revert back to the conventional truncated normal distribution in Equation 36, offering little improvement despite its added flexibility.

Since it is possible to determine which regime actually materialised for forecasts in the training data, a second choice is to find conditional probabilities of each regime occurring given the initial regime. That is, given a certain regime occurs at the initialisation time, one can calculate the proportion of instances in which each regime materialises at the validation time. An example of this for the GEFS reforecast data in

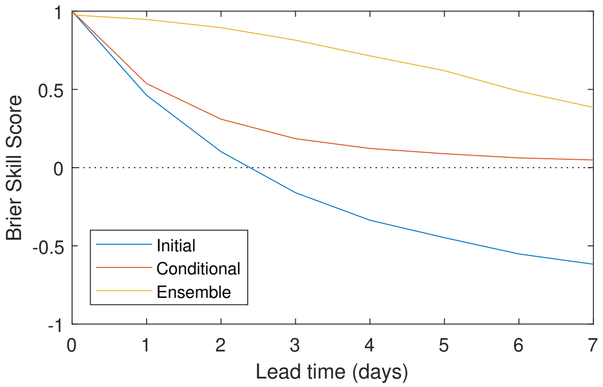

Figure 11: Brier skill score relative to climatology for different forecasts of the future weather regime, plotted against lead time.

Section 4.4 is shown in Table 6. The data and identified regimes are described in detail in Section 4.4.1. The initial regime weights assume that the probability of the positive phase of the North Atlantic Oscillation (NAO+) occurring after two days given that it transpired at the initialisation time is one, for example, whereas Table 6 suggests it is only 0.740 in reality. This, in theory, provides a more realistic probability of the regime at the forecast validation time. Such weights are called 'conditional regime weights'.

The ensemble members are themselves simulated trajectories of the atmosphere, and hence regimes can also be estimated from each ensemble member. As in the previous chapter, the proportion of ensemble members that are assigned to a regime constitutes a probability of residing in that state at the forecast validation time. This third choice is called 'ensemble regime weights'. Results in Chapter 3 indicate post-processing using ensemble regime weights outperforms the initial regime weights.

Since the weights define a probabilistic forecast of the future atmospheric state, they can be assessed by their ability to capture the regime that materialises. Figure 11 shows the Brier skill score (Brier, 1950) for the three different choices of weight, averaged across the four identified regimes in the reforecast data. The climatological frequencies of the different regimes are used as a reference forecast. Although useful at very short lead times, the initial regime weights become detrimental to forecasts relative to climatology (skill score less than zero) after only a few days. Unsurprisingly, they are particularly poor at predicting the less persistent weather types. The conditional regime weights, on the other hand, are designed to be at least as good as climatology and hence always result in a positive skill score. However, they rely on information from the initial regime, and the skill score therefore tends to zero as the lead time increases. Output from the NWP model will also contain progressively less information as forecast horizon increases, with studies highlighting model deficiencies in capturing the onset and decay of atmospheric blocking events (Tibaldi and Molteni, 1990; Matsueda and Palmer, 2018). Nonetheless, the weights defined by the ensemble members offer considerably more skill than alternative approaches at all lead times considered here.

In a study such as this, where predictions are evaluated over a set of hindcasts, forecasts can be conditioned on perfect knowledge of the regime at the validation time. Although this information is not available a priori to forecasters, it is implemented here to obtain a rough upper bound on the improvements to be gained from the regime-dependent mixture model. This is henceforth referred to as the 'true regime'. Furthermore, although the conditional regime weights are found to offer better forecasts of the atmospheric state, the performance of the resulting forecasts is found to be similar to when using the initial regime. Therefore, in the subsequent analysis, results are compared only for the initial regime, ensemble regime and the true regime weights.

Regime information is incorporated into post-processing via these mixture model weights. Therefore, in order to obtain forecast distributions that utilise the regime information, the weights are estimated first, prior to fitting the regime-specific predictive distributions. Coefficients for these component distributions are then estimated conditional on the regime weights. Furthermore, since the regime-dependent weights considered here are functions of the atmospheric flow rather than constant parameters, they can adapt to the current atmospheric conditions. This allows forecasts to be post-processed differently to one another, even when trained using the same data.

This is in contrast to alternative approaches that have been introduced to combine predictive distributions (Gneiting et al., 2013). Baran and Lerch (2016), for example, estimate the mixture model weights simultaneously to the post-processing parameters. Although the resulting wind speed forecasts are found to exhibit significantly better calibration than the individual component distributions, the corresponding parameter estimation step can result in optimisation problems that are complex and unstable, and thus computationally expensive. Baran and Lerch (2018) therefore investigate the use of forecast combination approaches that use a two-step procedure to estimate post-processing parameters. The approaches discussed therein first fit two or more distinct EMOS methods individually to all training data, and then find the optimal weights to combine the resulting forecast distributions. The method presented in this chapter similarly divides the parameter estimation into two stages, but distributional coefficients are instead estimated after having obtained the mixture model weights. Doing so allows the component distributions to capture separate features of the training data that arise due to the occurrence of each weather regime.

4.2.3 Parameter estimation

Gneiting and Raftery (2007) introduce the notion of optimum score estimation, which identifies the parameter values that optimise a proper score over the available training data. Maximum likelihood estimation fits into this framework since it is analogous to minimising the logarithmic score. Another popular score in the forecasting literature is the continuous ranked probability score (CRPS), defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{u \geq y\}]^2 du, \tag{38}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, F is the forecast distribution, and y the verification (Matheson and Winkler, 1976).

However, Baran and Lerch (2016) note that the CRPS for mixture models such as that in Equation 37 cannot be evaluated analytically and hence must be approximated numerically. As a result, parameter estimation with the CRPS becomes computationally expensive. Although minimum CRPS estimation is often regarded as a more robust choice than maximum likelihood for forecast recalibration, Gebetsberger et al. (2018) suggest the resulting estimates should be similar, provided the distributional assumptions are valid. Maximum likelihood is therefore chosen to estimate parameters in this study.

The probability density function of a normal distribution truncated below at zero, $N_0(\mu, \sigma^2)$, is

$$f_{TN}(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-1}, \tag{39}$$

where $\phi(\cdot)$ is the probability density function, and $\Phi(\cdot)$ the cumulative distribution function, of the standard normal distribution. The density for a mixture of TN distributions, $\sum_{r=1}^{R} w_r N_0(\mu_r, \sigma_r^2)$, is then simply a weighted sum of the component densities:

$$f_{MM}(y) = \sum_{r=1}^{R} \frac{w_r}{\sigma_r}\phi\left(\frac{y-\mu_r}{\sigma_r}\right)\left[\Phi\left(\frac{\mu_r}{\sigma_r}\right)\right]^{-1}. \tag{40}$$

The regime-dependent approaches estimate a set of parameters for each of the R identified regimes ($\alpha_r, \beta_r, \gamma_r, \delta_r$ for $r = 1, ..., R$) by maximising the likelihood of the mixture model over the training data, conditional on each choice of the regime weights.

If the weight takes the form of an indicator function, as is the case for the initial and true regime weights, then the mixture model forecast at a given time is equivalent to a truncated normal distribution with post-processing parameters that correspond, respectively, to the initial or true regime. The CRPS thus reduces to that for a truncated normal distribution, which is given in closed form in Equation 41. Nonetheless, to retain correspondence between the different methods, all statistical models are fit using maximum likelihood. Moreover, in the case of indicator weights, each forecast-observation pair in the training data is assigned to exactly one regime, and the training data can be partitioned into R mutually exclusive, collectively exhaustive training subsets. Post-processing parameters for the truncated normal distribution associated to a regime are then estimated by maximising the likelihood only over the subset of data containing forecast-observation pairs allocated to that regime.

Regime-dependent methods with indicator weights can thus also be interpreted as analogue-based post-processing approaches (see Section 2.3.3.2), whereby a training data set is constructed from forecast-observation pairs that are believed to exhibit similar behaviour to the current forecast (Junk et al., 2015). In this case, the assumption is that the forecast biases depend on the synoptic-scale behaviour of the atmosphere, which aligns with the motivation for using regime analogues in Barnes et al. (2019).

If the probabilities of residing in each regime are not strictly zero or one then the training data consists of all available forecasts and observations. Therefore, all post-processing parameters are estimated simultaneously by numerically maximising the likelihood associated with the mixture density in Equation 40. In this case, when estimating the post-processing parameters corresponding to a regime, the probability of each historical forecast-observation pair belonging to that regime determines the leverage it has in estimating the coefficients. Although this can be considerably more time consuming than parameter estimation for conventional methods, it is not found to be prohibitively expensive. Furthermore, Thorarinsdottir and Gneiting (2010) find a local EMOS approach, in which forecast recalibration occurs separately for each individual location, to perform better than aggregating training data across several spatial locations. Despite being more computationally demanding, this approach is implemented here, allowing the post-processing to account for local biases.

4.2.4 Forecast verification

A forecast distribution is said to be calibrated if events materialise with the same frequency with which they are forecast, while sharpness refers to the concentration of the distribution. Forecasters have come to seek predictive distributions that are sharp subject to being calibrated (Gneiting et al., 2007). The evaluation of forecasts must thus account for these two qualities, something that is achieved through the use of proper scoring rules (Gneiting and Raftery, 2007). The CRPS is used to verify forecasts in the following sections of this thesis, though similar conclusions are drawn from the logarithmic score.

The CRPS for a truncated normal predictive distribution is

$$\begin{aligned}\mathrm{CRPS}\big(N_0(\mu,\sigma^2),y\big) = \\ \sigma\left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-2}\bigg\{\frac{y-\mu}{\sigma}\Phi\left(\frac{\mu}{\sigma}\right)\left[2\Phi\left(\frac{y-\mu}{\sigma}\right)+\Phi\left(\frac{\mu}{\sigma}\right)-2\right]+ \\ 2\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\Phi\left(\frac{\sqrt{2}\mu}{\sigma}\right)\bigg\} \qquad (41)\end{aligned}$$

(Thorarinsdottir and Gneiting, 2010). The CRPS (Equation 38) for a mixture model forecast, on the other hand, is evaluated using Gauss-Laguerre quadrature, which is a commonly used numerical integration scheme for non-negative data.

The CRPS is negatively oriented and hence larger values indicate poorer performance. To compare the ability of the TN and regime-dependent (RD) TN frameworks, the continuous ranked probability skill score (CRPSS) is also applied, with the conventional TN approach as the reference forecast. The CRPSS is defined as

$$\mathrm{CRPSS}(H,F,y)=\frac{\langle\mathrm{CRPS}(F,y)\rangle-\langle\mathrm{CRPS}(H,y)\rangle}{\langle\mathrm{CRPS}(F,y)\rangle}=1-\frac{\langle\mathrm{CRPS}(H,y)\rangle}{\langle\mathrm{CRPS}(F,y)\rangle}. \qquad (42)$$

F denotes the predictive distribution obtained from TN, H denotes that obtained from RDTN, y is the corresponding verification and $\langle\mathrm{CRPS}(\cdot\,,y)\rangle$ is the average CRPS over forecasts in the test data set (Wilks, 2019). As mentioned in the previous chapter, the skill score can be interpreted as the relative improvement in score upon current post-processing methods, gained from regime-dependent post-processing. Skill scores are bounded above by one and values below zero indicate the RDTN method is performing worse than its reference. Therefore, unlike the CRPS, high values of the CRPSS are desired.

4.3 Quasigeostrophic model

4.3.1 Data

In this section we present results for the regime-dependent post-processing methods applied in a simulation study involving wind speed forecasts from a spectral quasi-geostrophic three-level model of the atmosphere in the Northern Hemisphere, triangularly truncated at wavenumber 21. The vertical levels are located at 250 hPa, 500 hPa and 750 hPa. The governing equations are:

$$\frac{\partial q_i}{\partial t}+J(\Psi_i,q_i)=D_i+S_i, \qquad i=1,2,3. \qquad (43)$$

Here, Ψ_i and q_i are the streamfunction and the potential vorticity at level i, respectively, and J denotes the Jacobian operator on the sphere. The dissipative terms D_i comprise Newtonian temperature relaxation at all levels, Ekman damping at the lowest level and hyperviscosity on the time-dependent part of the potential vorticity at all levels. The time-independent but spatially varying forcing terms S_i are diabatic sources of potential vorticity.

The model parameters and forcing are tuned in a way that the model in a long-term integration exhibits a remarkably realistic mean state and variability pattern of streamfunction and potential vorticity. The model is integrated forward in time using the third-order Adams-Bashforth scheme with a constant time step of one hour. Details of the model configuration, parameter setting, parameter tuning procedure, and performance versus reanalysis data can be found in Kwasniok (2007) and Kwasniok (2019). The model configuration used here is exactly the same as described in Kwasniok (2019). The streamfunction, Ψ, represents the trajectory of particles in this model and hence the circulation of the atmosphere in the Northern Hemisphere on each of the vertical levels can be represented instantaneously by the streamfunction in 1024-dimensional space, comprised of grid point values at 64 equally-spaced longitudes and 16 Gaussian latitudes. Regimes are therefore located by searching for quasi-stationary equilibria in the streamfunction.

The system described above was first integrated forward in time for 50 years and the atmospheric regimes were identified using the resulting time series of daily streamfunction fields. To construct the training and test data sets, the QG model was then run for a further 30 years, with both the streamfunction and wind speed at all locations recorded daily. Since this systems acts as a surrogate for the atmosphere, the recorded wind speed fields are treated as observations, while the streamfunction field provides a best guess of the atmospheric state at that time. These ‘observed’ states are then used as forecast analyses. An ensemble forecast comprised of ten exchangeable members was constructed by adding random perturbations from a $N(0, 0.00025^2)$ distribution to the streamfunction at every location on the domain, expressed in spherical harmonics, and propagating the resulting initial conditions through time for seven days, using a version of the quasigeostrophic model truncated at wavenumber 19. Perturbing the analyses reflects uncertainty in the initial forecast state, while a more severely truncated model is used to replicate an imperfect NWP model. The results were not dependent on the ensemble size, and post-processing separately at each location means perturbations that are not necessarily spatially independent should not have an adverse effect on the results.

The resulting data therefore includes 30 years worth of daily forecast-observation pairs, for daily lead times up to one week ahead. Half of this data is used to train the post-processing methods, while the remaining data is used to assess the resulting predictions. For each locations, both the training and test data set thus consist of 5,475 ensemble forecasts of wind speed and their corresponding observations.

Quasigeostrophic models have previously been employed to investigate the behaviour of planetary-scale flow regimes (Marshall and Molteni, 1993; Majda et al., 2006; Franzke et al., 2008). Kondrashov et al. (2004), for example, used a similar model to study transitions between phases of the North Atlantic Oscillation and the Arctic Oscillation (or Northern Annular Mode), two dominant flow regimes in the Northern Hemisphere. One particular feature of the QG model is that it exhibits no seasonal cycle, residing perpetually in winter. This is the season in which the regime behaviour of the atmosphere is most pronounced and therefore this system has the added benefit that it could produce more robust atmospheric states (Hannachi et al., 2017).

To reduce the dimension of the data, principal component analysis (PCA) is applied to the grid of streamfunction anomaly values at 500 hPa. PCA works by finding orthonormal variables, $\boldsymbol{z}$, that are themselves linear combinations of the original variables, allowing a large proportion of the variation in the data to be represented by a comparatively small number of the transformed variables (Wilks, 2019). That is, rather than representing the atmospheric circulation using a vector of streamfunction values at each grid point

$$\boldsymbol{\Psi} = (\Psi_1, \Psi_2, ..., \Psi_{1024}), \tag{44}$$

PCA allows the flow to be described by just a few of the uncorrelated, transformed variables

$$\boldsymbol{z} = (z_1, z_2, ..., z_p), \tag{45}$$

that explain a relatively large proportion of the low-frequency variability in the atmosphere. In this study, the norm streamfunction metric is used in the PCA. The number of principal components is $p \ll 1024$; the leading three principal components are retained here, which explain 22.0% of the variation in the hemispherical streamfunction. The synoptic-scale atmospheric state is projected onto the leading principal components, and regimes are identified by performing an additional clustering step in this reduced space. This nonlinear approach allows opposite phases of a mode of atmospheric variability to exhibit spatial asymmetries, as is the case, for example, for the North Atlantic Oscillation (NAO; Cassou et al., 2004).

In particular, the time series consisting of 50 consecutive years worth of daily streamfunction anomalies is projected onto its leading three principal components, and it is from this sequence of 18,250 materialisations of $\boldsymbol{z}$ that the regimes are detected. This archived data is sequential, and so a hidden Markov model (HMM) is used to discern the regimes. Majda et al. (2006) first proposed the use of hidden Markov models in detect-

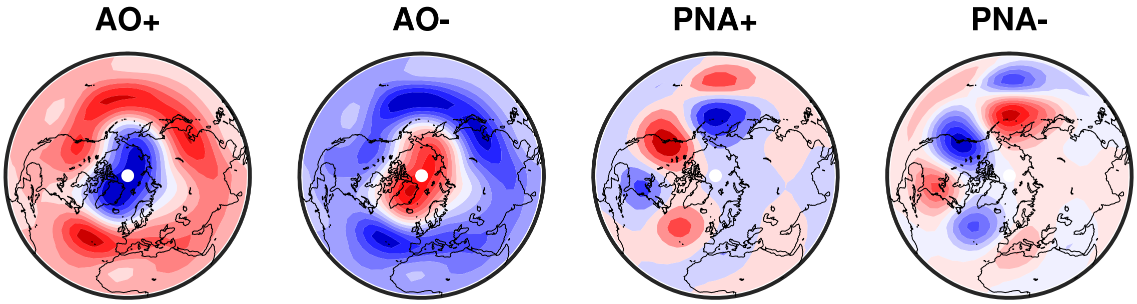

Figure 12: Regime centres produced by fitting a hidden Markov model to the transformed streamfunction anomalies from the quasigeostrophic model.
Blue regions represent negative contours, while positive anomalies are shown in red. Contours are separated by intervals of $5 \times 10^5 \mathrm{m}^2\mathrm{s}^{-1}$.

ing atmospheric regimes, highlighting their ability to distinguish between distributions despite the leading principal components exhibiting nearly Gaussian statistics: HMMs are designed to detect more persistent regimes by exploiting the system's underlying dynamics.

A HMM assumes that the transformed variables in each regime follow a multivariate normal distribution, $\boldsymbol{z} \sim N(\boldsymbol{\mu_r}, \boldsymbol{\Sigma_r})$, and hence a mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, corresponding to each state must be estimated. A transition matrix, documenting the probability of transitioning between regimes, is also derived. The mean vectors, covariance matrices and the transition matrix are all estimated over the 50 years of streamfunction fields, after transforming these fields to the leading 3 principal components, and estimation is performed using maximum likelihood via the Baum-Welch algorithm. The resulting hidden Markov model emits probabilities that the atmosphere resides in each regime on every day throughout the 50 year period. By taking the regime on each day with the highest probability of occurrence, it is straightforward to determine the most likely regime sequence over the data set, often known as the Viterbi path.

The four regime centres identified by fitting a hidden Markov model to the archived data are depicted in Figure 12. The number of clusters is chosen to be four due to the similarity of the resulting patterns to recognised atmospheric regimes: the positive and negative phases of both the Arctic Oscillation (AO+, AO-) and Pacific-North American (PNA+, PNA-) patterns. The positive (negative) phase of the AO is synonymous with a strong (weak) polar vortex over the Arctic Circle, surrounded by a band of above (below) average streamfunction anomalies in the midlatitudes. The AO thus represents a zonally symmetric seesaw in streamfunction, or pressure anomalies between the Arctic basin and the extratropics (Thompson and Wallace, 1998). The positive (negative) PNA pattern, on the other hand, consists of below (above) average streamfunction anomalies over the Aleutian Islands, and areas of high (low) anomalies over the Pacific basin and the northwestern US (Wallace and Gutzler, 1981; Leathers et al., 1991). Mean persistence times can be calculated from the Viterbi path. The AO- regime (which occurs 24.9% of the time) has the longest mean persistence time, 10.7 days, followed by the AO+ regime (26.0%) which lasts for 9.6 days on average. The PNA patterns are comparatively less persistent, with the positive mode (29.2%) lasting for 6.1 days on average and the negative mode (19.9%) only 5.6 days, making it the least stable.

4.3.2 Assigning forecasts to regimes

As mentioned above, HMMs assume a statistical distribution for the transformed streamfunction variables conditional on each underlying regime. As a result, having projected the streamfunction anomaly field onto the leading three principal components, Bayes' theorem can be used to calculate posterior probabilities of the atmosphere residing in each regime given the streamfunction values. The probability of residing in regime r given the reduced circulation, $\boldsymbol{z}$, is

$$p(r|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|r)p(r)}{p(\boldsymbol{z})} = \frac{p(\boldsymbol{z}|r)p(r)}{\sum_{j=1}^{R} p(\boldsymbol{z}|j)p(j)}. \tag{46}$$

Here, $p(\boldsymbol{z}|r)$ is the likelihood of seeing the observed or predicted streamfunction values given that the atmosphere resides in regime r, and can be calculated from the multivariate normal density with mean vector and covariance matrix associated with that regime, $\boldsymbol{\mu_r}$ and $\boldsymbol{\Sigma_r}$. The climatology frequency of regime r is denoted by $p(r)$. When the forecast must be assigned to exactly one regime, that with the highest posterior probability is chosen. Therefore, the initial and true regimes can be determined by finding the regime that maximises the posterior probability given the observed streamfunction anomaly field at the forecast initialisation time and validation time, respectively. Similarly, the predicted streamfunction fields from the ensemble members can be used to allocate each member to a regime.

Obtaining a probabilistic distribution for the regime accounts for some of the inherent uncertainty present when identifying the latent atmospheric state. However, Bayes' theorem as given in Equation 46 does not make use of the estimated transition matrix and hence does not perfectly utilise the HMM's dependence on the system dynamics. A HMM produces a time series of posterior probabilities for each state given all data in the sequence. Calculating the Viterbi path, the most probable sequence of hidden states, then allows exactly one regime to be identified at each point in time. In a forecast setting, if a window of recent values were available prior to the current forecast, then the initial regime could be determined from the Viterbi path over this window, rather than the static posterior probability. This would exploit the dynamics of the underlying states, and would therefore be particularly useful when the spatial structures of the regimes were similar, so that the temporal behaviour was more important when distinguishing between states. The regimes here differ considerably in space, and hence the estimated regime is not sensitive to the choice of method (not shown). Bayes' theorem, however, can more easily be applied to determine the future regime from each ensemble member, when the preceding states are also unknown. Therefore, for ease of implementation, Bayes' theorem is used here to ascertain the regime given a streamfunction anomaly field.

4.3.3 Results

The spatial domain of the QG model consists of 64 longitudes and 16 latitudes in the Northern Hemisphere and statistical post-processing is implemented at every grid point, yielding calibrated forecasts at 1024 locations. No spatial aggregation is performed and hence forecasts at each site are recalibrated using only previous forecast-observation pairs at the same location.

Figure 13 displays the CRPS for the TN approach, plotted on a map of the Northern Hemisphere at a lead time of six days. Forecast accuracy is worst towards the centres of the Pacific and Atlantic oceans, areas which correspond to well-known storm tracks. Maps of the CRPSS for the three regime-dependent methods, assessed using TN as reference, are also shown in Figure 13 for the same lead time. RDTN-init denotes the regime-dependent truncated normal approach conditioned on the initial regime, RDTN-ens is that using the ensemble member weights to predict the regime, and RDTN-true is dependent on the true weather regime at the forecast validation time. At locations far removed from the centres of the weather regimes, the improvements unsurprisingly fluctuate around zero. However, when the regimes strongly affect the local wind speeds, the RDTN-true method produces noticeable improvements in forecast skill. Both the RDTN-init and RDTN-ens methods appear considerably less effective than using the true regime.

In the Northern Hemisphere, wind travels counterclockwise around large-scale low pressure systems and clockwise around high pressure, with the strengths of the winds related to the north-south pressure gradient (Hurrell and Deser, 2009). The improvements gained from the RDTN-true approach in Figure 13 are therefore concentrated in the North Atlantic and Pacific basins, and over northwest Canada: these regions surround the regime centers of action, so that the wind direction and intensity vary substantially depending on the prevailing regime. The wind speeds at these locations

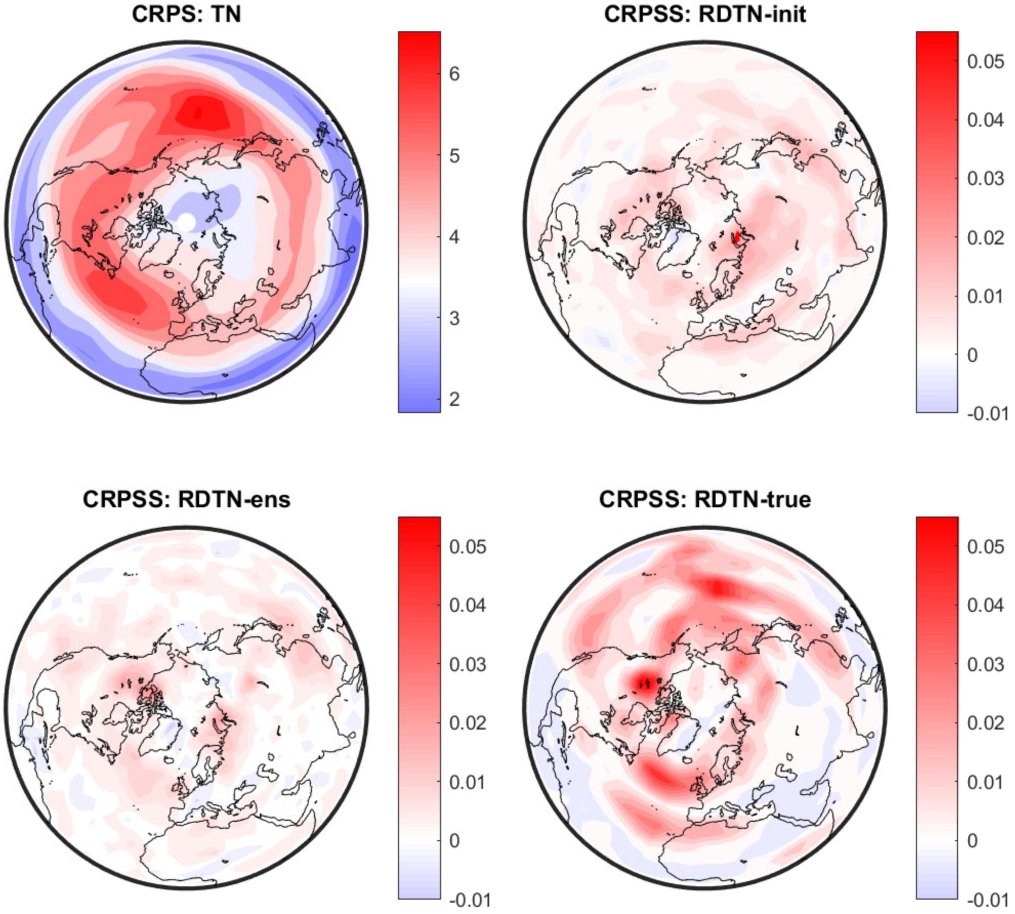

Figure 13: Map of the CRPS for the TN approach and the CRPSS for the RDTN methods using TN as the reference forecast, at a lead time of six days.

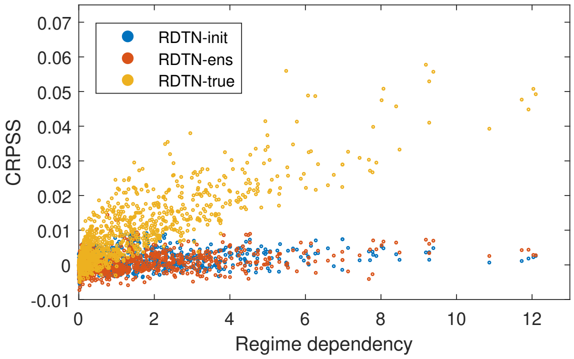

Figure 14: Scatter plot showing the measure of regime-dependency against the CRPSS at all grid points on the QG model domain at a lead time of seven days.

	Lead time (days)						
	1	2	3	4	5	6	7
RDTN-init	0.053	0.271	0.383	0.382	0.378	0.361	0.305
RDTN-ens	0.033	0.220	0.447	0.497	0.513	0.442	0.309
RDTN-true	0.064	0.390	0.651	0.770	0.820	0.842	0.855

Table 7: The correlation between the measure of regime-dependency and the CRPSS, calculated over grid points on the QG model domain.
Results are shown for the three regime-dependent methods at lead times up to one week.

are thus more heavily influenced by the different regimes, resulting in an increased need for regime-dependent post-processing methods.

Historical observations at any given location can be grouped depending on the coinciding regime. To describe the extent to which the regimes under consideration affect the wind speeds at this location, it is helpful to introduce a measure of regime-dependency. This measure is taken to be the between-group component of the empirical law of total variance:

$$\sum_{r=1}^{R} \frac{n_r}{n} (\bar{y}_r - \bar{y})^2, \tag{47}$$

where n_r/n is the proportion of days on which regime r occurs in the training data, $\bar{y}_r$ is the average wind speed given regime r, and $\bar{y}$ is the overall temporal mean wind speed. This metric quantifies the effect the regimes have on the wind speed at a particular location. A scatter plot is displayed in Figure 14, showing the regime-dependency against the CRPSS for all locations at a lead time of one week. Table 7 records the associated correlation between this metric and the improvements gained from regime-dependent post-processing at lead times up to seven days. Although the correlation is initially fairly low, the improvements gained when the true regime is used in post-

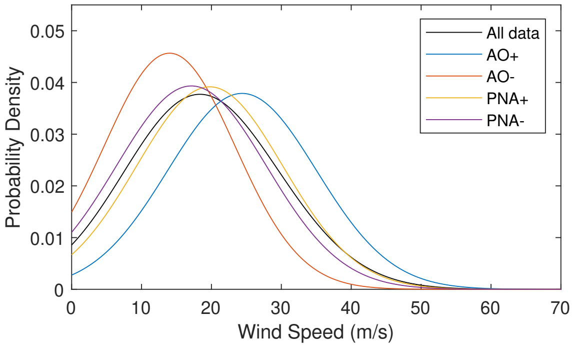

Figure 15: Empirical distribution of wind speed observations at one location on the QG model domain when the atmosphere resides in each regime.

processing become highly correlated with the measure of regime-dependency at longer lead times. This suggests that there is more potential for improvement upon current post-processing approaches for forecasts further in advance. Neither the initial regime nor the ensemble member weights capture this behaviour.

To highlight the potential improvements, we focus now on results at one location in the west of the Atlantic Ocean. The marginal distribution of the wind speed, shown in Figure 15 when the atmosphere resides in each regime, indicates the local wind speed is dependent on the prevailing state. At this location, the AO+ corresponds to a strong negative meridional gradient in streamfunction anomalies, in turn producing high zonal wind speeds. Conversely, the negative phase of the Arctic Oscillation is synonymous with low wind speeds in this area. The PNA patterns have less influence at this location, though wind speeds that are slightly lower than average occur in the negative phase. The shape of the empirical wind speed distributions also undergo noticeable changes between the regimes: in the AO- regime the wind speeds are far more positively skewed

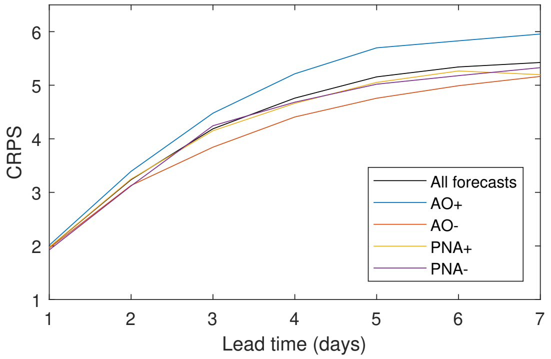

Figure 16: CRPS for the TN method against lead time at one location on the QG model domain when the atmosphere resides in each regime at the forecast validation time.

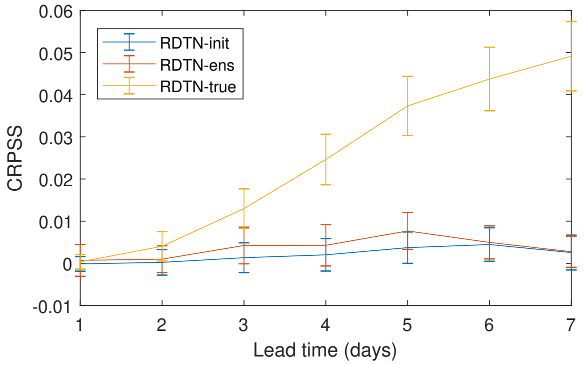

Figure 17: CRPSS against lead time for all three regime-dependent post-processing models at one location on the QG model domain, with TN as the baseline. Error bars indicate 95% confidence intervals for the skill score.

than in the AO+ pattern. Although the formulation of the mixture model in Equation 35 allows separate forecast distributions to be issued depending on the regime, the truncated normal distribution is able to adapt for such changes.

The skill of the TN post-processing approach can be evaluated using the CRPS. Figure 16 displays the breakdown of the CRPS depending on the weather regime that occurs at the forecast validation time. There is a clear difference in forecast performance depending on the prevailing weather type. Forecast accuracy is lowest when the AO+ regime materialises, in which extremely high wind speeds occur more frequently, while the lower wind speeds in the AO- are more predictable.

Figure 17 illustrates the skill of the regime-dependent TN predictive distributions relative to the conventional TN approach, assessed using the CRPSS. The uncertainty in the skill score is described by errors bars representing 95% confidence intervals, obtained via nonparametric bootstrap resampling. Although the improvements for all methods are initially negligible, RDTN-true forecasts become substantially more skilful at longer lead times: wind speed forecasts at this location improve by almost 5% by including the synoptic-scale information. The RDTN-init approach, on the other hand, fails to make any meaningful contribution to the forecast. The CRPSS for RDTN-ens is significantly larger than zero for forecasts five and six days in advance, though the magnitude of the improvement in both cases is small. Figure 16 suggests that forecast biases are initially relatively insensitive to the underlying regime and hence incorporating regimes would only be expected to benefit forecasts at longer lead times. However, by the time the biases become dependent on the weather regime, the ability of the mixture model weights to recognise the true regime deteriorates. The skill score for the RDTN-init and RDTN-ens methods therefore consistently remains close to zero.

It is possible to decompose the CRPSS for the RDTN-true approach into the constituent regimes, as displayed in Figure 18. Although wind speeds are initially most predictable in the AO-, improvements are also largest in this regime, reaching 12% for forecasts one week ahead. Predictions of the higher wind speeds in the AO+ also improve, becoming up to 6% more skilful than when regime information is neglected. The PNA patterns have much less influence on the wind speed here and hence there is little benefit to including information in these states. Nonetheless, the improvements in forecasts in the AO regimes indicate that regime-dependent approaches may be more capable of forecasting events that deviate substantially from the local climatology, including extreme weather events.

For sufficiently large lead times, the raw forecast becomes uninformative, containing no information about the predictand. In this case, the statistical post-processing methods should issue as a forecast the marginal distribution of the weather variable of interest. It is believed that if the atmospheric regime could be forecast perfectly then the improvement gained from regime-dependent post-processing would be present even this far in advance, since the regime-dependent post-processing will issue the marginal distribution of the wind speed in each regime. The additional flexibility of the mixture model therefore allows it to capture more complex features that arise due to the different regimes, such as multimodality of the marginal distribution.

Figure 19 displays the relative frequency with which the observation assumes each rank when pooled with the ensemble members, at a lead time of five days. Rank histograms are a commonly used tool for assessing the calibration of ensemble forecasts, with uniform histograms denoting reliable predictions (Anderson, 1996; Talagrand, 1997b; Hamill and Colucci, 1997). Clearly, however, the raw ensemble forecast is underdispersed, with observations falling outside the range of ensemble members more frequently than would be expected if the forecast were calibrated, regardless of

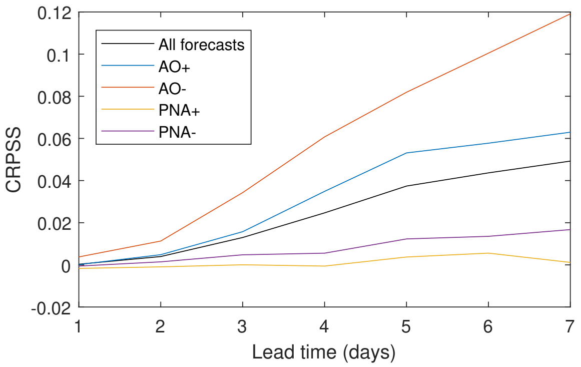

Figure 18: CRPSS for RDTN-true forecasts against lead time at one location on the QG model domain when the atmosphere resides in each regime at the forecast validation time, with TN as the reference prediction scheme.

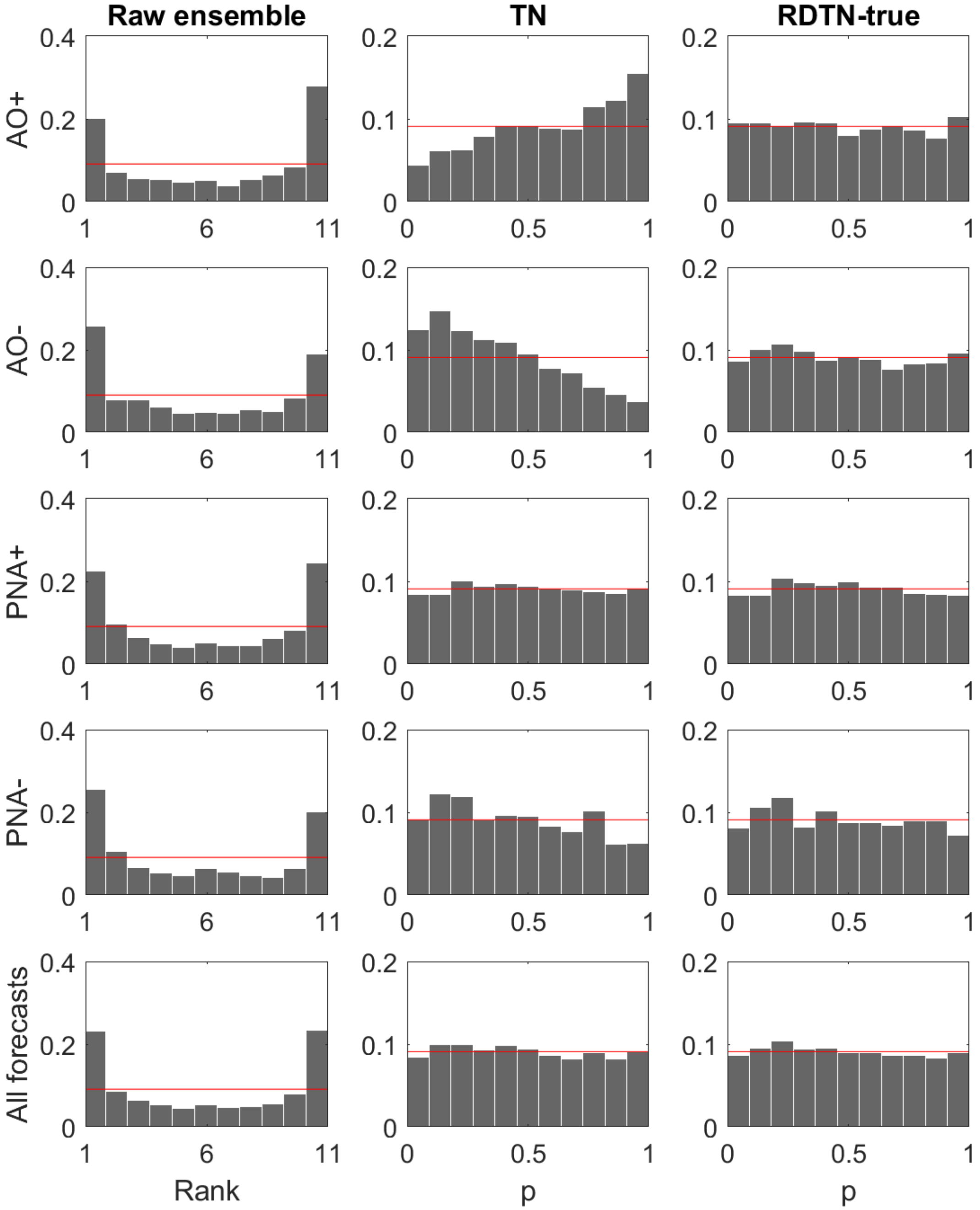

Figure 19: Rank (left) and PIT (middle and right) histograms showing the relative frequency of each bin for the raw ensemble forecasts, the TN and RDTN-true post-processing methods at a lead time of five days.
Histograms are shown for forecasts grouped by the atmospheric regime at the forecast validation time. A horizontal line is added in red at 0.091 to indicate perfect uniformity across the bins.

the underlying regime. Also shown in Figure 19 are Probability Integral Transform (PIT) histograms, the continuous analogue of the rank histogram. The PIT histograms evaluate the TN and RDTN-true forecast distributions at the verification, and record the rate at which the resulting probabilities fall into each of a number of equally-sized bins. There are eleven possible positions of the verification when pooled with the ten ensemble members and hence the number of bins is also chosen to be eleven.

Although the PIT histogram for the TN approach over all forecast-observation pairs is suitably uniform at this location, the predictive distributions estimated over the entire training set do not fit the data well when the system resides in the Arctic Oscillation patterns. In particular, oppositely skewed PIT histograms indicate that the observed wind speed falls into the upper tail of the forecast distribution when the AO+ regime occurs, and the lower tail when the AO- pattern materialises. The TN approach is thus not calibrated conditional on the regime. The RDTN-true approach, on the other hand, accounts for the varying model biases in the regimes, and the corresponding PIT histograms are close to uniform in all of the four regimes.

4.4 GEFS reforecasts

4.4.1 Data

Previous occasions in which similar atmospheric behaviour has occurred will likely lead to similar model biases. Therefore, regime-dependent statistical post-processing would be particularly well-suited to use with reforecast data, where a large set of hindcasts from a frozen operational model are available. These hindcasts, spanning several years or decades, can be used to train statistical post-processing methods (Hamill et al., 2004). In this section, the regime-dependent approaches are implemented on wind speed data from version 2 of the National Oceanic and Atmospheric Administration's Reforecast project (Hamill et al., 2013). Forecasts are taken from a recent version of the National Centers for Environmental Prediction's (NCEP) global ensemble forecasting system, and, although Hamill et al. (2013) note that they may also be prone to model biases, the reanalyses are used as a best guess for the observed wind speed values. To lessen these biases, the control member initialised from the reanalysis is omitted, resulting in a forecast of ten statistically interchangeable ensemble members.

Since regime behaviour is most prominent during winter, the data set covers the 34 cold seasons (November to March inclusive) between 1985 and 2019. Results in the previous section established that locations heavily affected by the identified weather regimes are likely to improve as a result of regime-dependent post-processing. Therefore, it is hoped that defining localised regimes over a smaller spatial domain will have a larger effect on the recalibration. Following Ferranti et al. (2015), the atmospheric

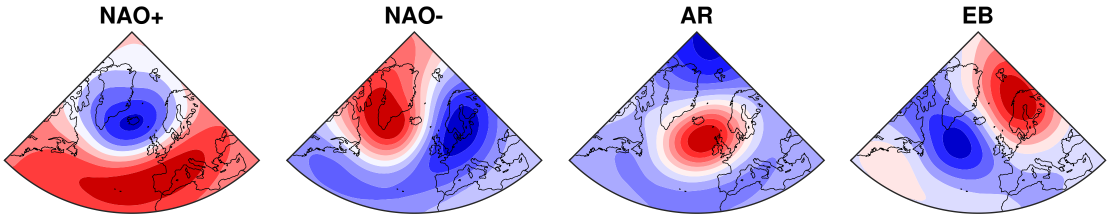

Figure 20: Regime centres identified by applying k-means clustering to the transformed reforecast geopotential height anomalies.
Anomalies are displayed at 25hPa intervals, with blue regions indicating negative contours and red regions representing positive anomalies.

regimes here are detected using k-means clustering over the Euro-Atlantic sector $(80W-40E, 30-90N)$. k-means clustering partitions data into a prespecified number of groups by assigning data points to clusters such that the distance between the point and the allocated cluster centroid is minimised (Michelangeli et al., 1995; Wilks, 2019). The number of clusters must be chosen prior to implementing the algorithm, and four regimes are again used here due to the similarity of the resulting patterns to those identified in numerous studies of regime-behaviour over this domain (e.g. Michelangeli et al., 1995; Cassou et al., 2004; Dawson et al., 2012; Ferranti et al., 2015; Matsueda and Palmer, 2018).

Reanalyses of 500hPa geopotential height anomaly fields are used to represent the atmosphere's circulation in this domain. PCA is then applied, using the Euclidean metric in grid point space, to these anomaly fields, and the clustering is performed in this reduced space. The leading three principal components are chosen, which explain 48% of the variation in the flow. Figure 20 shows the geopotential height anomalies that correspond to the regime, or cluster, centres identified using k-means clustering. Despite fewer principal components being used than by Ferranti et al. (2015) (3 rather than 10), there is similar evidence to support the use of four regimes which resemble the positive and negative phases of the NAO, as well as European Blocking (EB) and an Atlantic Ridge (AR). It has been proposed that the NAO corresponds to the same mode of circulation variability as the Arctic Oscillation described in Section 4.3 (Hurrell and Deser, 2009). The NAO thus constitutes a north-south dipole, characterised by negatively correlated geopotential height anomalies between Iceland and the Azores, though opposite phases of the NAO do not exhibit identical spatial structures. The AR pattern represents an anticyclonic regime over the eastern North Atlantic Ocean, while European Blocking consists of a dipole with positive geopotential height anomalies over Scandinavia, and negative anomalies to the south of Greenland.

The atmosphere resides in the NAO+ regime 30.8% of the time, making it the most frequently occurring regime, the NAO- and EB regimes occur similarly often (24.2% and 24.3% respectively) while AR materialises least often (20.7%). There are unavoidable gaps in the data between the winter periods in successive years, and hence a hidden Markov model cannot readily be applied. The main benefit of k-means clustering, on the other hand, is that forecasts can be assigned to a regime with ease. Franzke et al. (2008) remark that, although clustering approaches find the states that have the highest probability of occurring, the resulting regimes do not necessarily exhibit persistence. As a result, the mean persistence times are much lower for these regimes than for the patterns found using a HMM in the QG framework: AR events persist for only 3.8 days on average, the EB for 4.9 days, and the NAO- and NAO+ regimes for 5.4 and 6.1 days on average, respectively.

Forecasts are assigned to one of these four patterns by finding the regime for which the Euclidean distance between the associated cluster centroid and the transformed geopotential height anomaly field is minimised. As before, the initial and true regimes make use of the observed geopotential height anomaly field at the forecast reference time and validation time, respectively, while output from the ensemble forecast is used to allocate each member to a regime. This approach fails to account for the inherent uncertainty when assigning a forecast to a regime: every point allotted to a cluster is assumed to exhibit the same biases and systematic errors, regardless of its distance to the cluster centre, and hence a method that provides the probability of residing in the different regimes, or a degree of membership, would be more informative in this respect.

Hamill et al. (2013) remark that the method for constructing the forecast analyses in the GEFS changes in February 2011, and the forecast skill consequently improves. Therefore, to maintain the homogeneity of the model biases in this study, only forecasts in the 25 cold seasons from November 1985 to March 2010 are considered. Although this model change also affects the observed geopotential height anomalies, it provides a more informed estimate of the atmospheric state and hence data after this change are still utilised when detecting the regimes. The resulting regimes are found to be more robust when this additional data is included.

Although techniques have recently been proposed that include cyclic functions to remove seasonal model errors (Messner et al., 2017; Dabernig et al., 2017; Lang et al., 2019a), parameter estimation is typically performed operationally using a training window that consists only of the most recently available forecast-observation pairs. These rolling windows account for the recent behaviour of forecast errors, and alleviate biases owing to changes in the NWP model. The size of this window is clearly a compromise between having enough data to obtain reliable parameter estimates without using too much, so as to capture the recent behaviour of the atmosphere. The CRPS is found

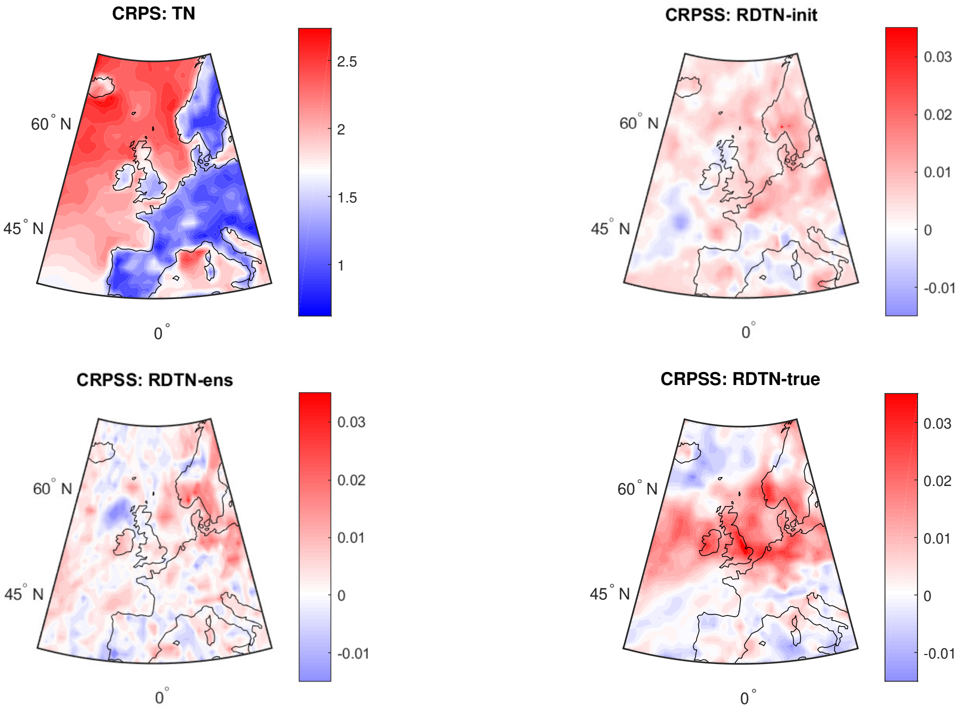

Figure 21: CRPS for the TN approach and CRPSS for the RDTN-init, RDTN-ens and RDTN-true extensions relative to TN, plotted on a map of the spatial domain under consideration at a lead time of seven days.

here to decrease as the amount of training data available increases, but is generally insensitive to the window length (not shown). In fact, at the majority of locations tested, more skilful forecasts are issued when a fixed training window is used, containing several years of past data. Therefore, for both the standard and regime-dependent post-processing methods, the first 15 cold seasons (those beginning in 1985-1999) are used as a fixed training window, while the remaining ten (2000-2009) are used as test data. The advantages of using a rolling window diminish in this case since there are no changes in the prediction system, and investigating only cold-seasons accounts for some of the seasonality in the biases.

4.4.2 Results

Post-processing is performed here on a subset of the spatial domain under consideration, which consists of 1353 locations over western Europe and the east of the North Atlantic ocean ($21W - 19E, 37 - 69N$). Locations are separated by 1 degree of longitude and latitude. The CRPS for the TN post-processing approach is displayed in Figure 21. Wind speed forecasts are significantly more skilful over land than sea, and forecasts at locations close to Iceland are particularly poor, since this corresponds to a mode of North Atlantic storm track variability (Serreze et al., 1997). The CRPSS for RDTN-

	Lead time (days)						
	1	2	3	4	5	6	7
RDTN-init	0.025	0.067	0.038	0.069	0.044	-0.046	0.004
RDTN-ens	-0.037	0.009	-0.042	0.069	0.052	0.010	0.013
RDTN-true	-0.041	0.021	-0.000	0.153	0.225	0.455	0.585

Table 8: Correlation between the measure of regime-dependency and the CRPSS, calculated over all grid points under consideration on the GEFS model domain. Results are shown for the three regime-dependent methods at lead times up to one week.

init, RDTN-ens and RDTN-true are also displayed in Figure 21 at a lead time of one week. At this longer lead time, the CRPSS for RDTN-init and RDTN-ens remains close to zero, though large improvements are seen when the true regime is used, particularly at locations surrounding the North Sea. We postulate that the spatial structure of the improvements in Figure 21 is again linked to how air flows around large-scale pressure systems. The regime centres in Figure 20 suggest that the regions between the modes of high and low pressure often intersect the area surrounding the North Sea, and therefore the wind speeds at neighbouring grid points are more dependent on the prevailing weather type. Calibrating forecasts separately in each regime thus produces larger improvements at these locations.

Table 8 shows that the improvements are again correlated with the measure of regime-dependency introduced in Equation 47. The magnitude of the metric tends to be much smaller than in the QG study, suggesting the sectorial regimes have less effect on the local wind speed than the hemispherical regimes. This is consistent with results in Tibaldi and Molteni (1990), which show that intense blocking events occur more over the Pacific than the Atlantic.

More detailed results are provided for one location close to Bergen, on the west coast of Norway. The quality of the raw ensemble forecast can be assessed using the CRPS to understand how the raw model errors change with the regime. Table 9 displays the accuracy of the raw ensemble forecasts initialised in each regime. Forecasts initialised in the NAO+ regime, which coincides with more extreme wind speeds, are considerably less accurate than those in the other regimes. Differences in the skill of forecasts among the regimes indicate conditioning the statistical post-processing on the prevailing regime may therefore be expected to yield more skilful forecasts.

	NAO+	NAO-	AR	EB	Total
Raw ensemble	1.44	1.18	1.06	1.24	1.23

Table 9: CRPS for raw GEFS ensemble forecasts initialised in each regime, at a lead time of three days, at a location close to Bergen, Norway.

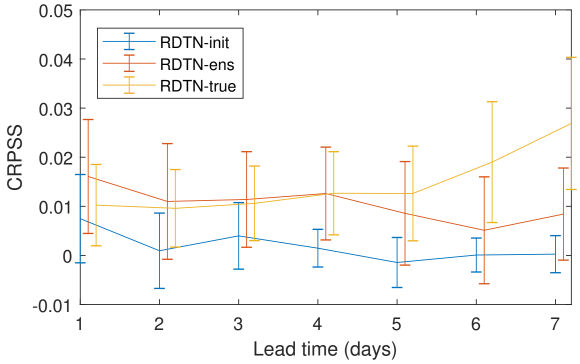

Figure 22: CRPSS against lead time for each of the three regime-dependent methods relative to TN, applied to GEFS reforecasts at a location close to Bergen, Norway. Scores for RDTN-ens have been offset by 0.1 days and RDTN-true by 0.2 days to visualise 95% confidence intervals around the skill scores.

Figure 21 suggests that using the true regime at this location provides relatively large improvements that are not present when conditioning forecasts on the regime at the initialisation time. The CRPSS is shown for all lead times in Figure 22, with 95% confidence intervals at each lead time estimated using nonparametric bootstrap resampling; in particular, forecast-observation pairs are randomly sampled with replacement from the test data set to construct a modified test set, and the CRPSS is then calculated over this new set of forecasts and observations - this is repeated 1000 times, and uncertainty in the skill score can be quantified using the variation in the resulting distribution of CRPSS values. A similar pattern emerges to that seen previously: scores for the initial regime recede to zero as lead time increases, while there appears to be more room for improvement at longer lead times, which is exploited when using the true regime. The RDTN-ens approach performs significantly better than the TN method, and is comparable to RDTN-true, for forecasts up to four days ahead, but

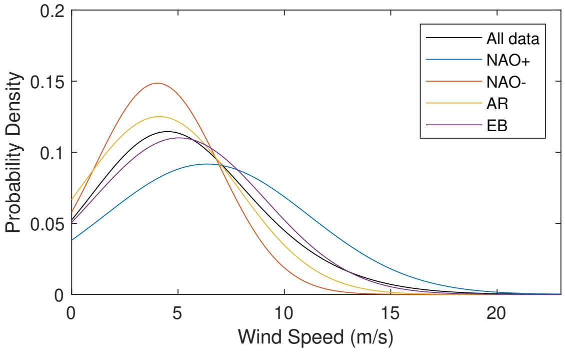

Figure 23: Empirical distributions of wind speed observations at a location close to Bergen, Norway, when each identified regime occurs in the reforecast setting.

its skill declines as lead time increases. The fact that RDTN-ens produces a larger CRPSS than RDTN-true at early lead times could indicate that the ensemble regime may be exploiting a feature of the data that is not picked up by the true regime, though it is more likely a result of sampling variation. Due to the large amounts of training data available from reforecasts, no methods perform worse than when regimes are not included in the post-processing, despite the increased complexity of the approach.

Figure 23 shows the empirical distributions of the wind speed at this location when the atmosphere resides in each regime. Positive NAO indices are linked to more intense and frequent storms in the Norwegian Sea (Serreze et al., 1997), and hence wind speeds here are largest in the NAO+ regime. Comparatively low wind speeds are associated with the NAO- regime, while the EB and AR regimes do not have much effect on the wind speed at this location. As a result, the improvements gained from regime-dependent post-processing are dominated by improvements in the two phases of the NAO. Figure 24 shows the CRPSS for the RDTN-true approach for forecasts corresponding to each regime at the forecast validation time. In particular, Figure 24 suggests that improvements at short lead times primarily occur in the NAO- regime, while at longer lead times forecasts in the NAO+ regime improve by as much as 4% upon the conventional TN approach. Since the positive phase of the NAO is associated with particularly high wind speeds at this location and the negative phase with low wind speeds, these results reinforce the idea that if the regime at the forecast validation time is correctly identified then regime-dependent post-processing can provide better forecasts of extremely high and low wind speeds. The uncertainties in these skill scores are non-negligible, as indicated by the error bounds for the overall improvement in Figure 22, though the associated confidence intervals have been omitted for ease of interpretation. Rank and PIT histograms for the various post-processing methods display similar features to those shown in Figure 19: the observation falls into the upper tail of the TN forecast distribution more frequently than expected during NAO+ events, and less frequently during NAO- events, while the RDTN-true method appears calibrated conditional on the regimes. However, since the improvements are smaller here, deviations from uniformity are less pronounced (not shown).

4.5 Discussion

This chapter builds upon the work on the regime-dependent statistical post-processing of ensemble forecasts presented in the previous chapter. It is suggested that NWP models exhibit biases that change depending on the concurrent atmospheric regime, and hence conditioning current statistical calibration methods on these regimes can enhance forecasts. Wind speed is closely connected to the movement of air in the atmosphere

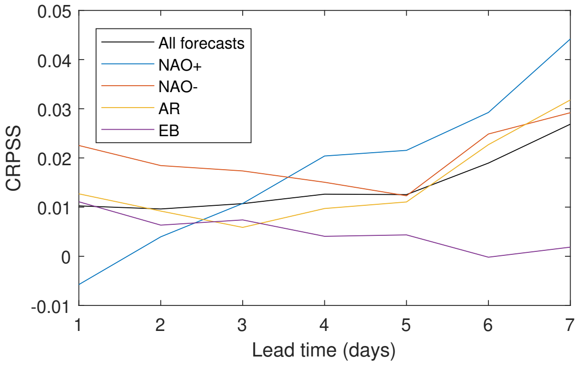

Figure 24: Skill score for RDTN-true forecasts at a location close to Bergen, Norway, partitioned by the true regime at the forecast validation time, shown at all lead times with the TN method as the reference.

and is therefore dependent on the prevailing regime behaviour. Regime-dependent extensions of Non-homogeneous Regression are proposed that utilise a weighted mixture of truncated normal predictive distributions. Mixture models of this form provide a more flexible forecast distribution than the individual component distributions, and can thus account for biases owing to changes in the atmosphere's synoptic-scale circulation. The mixture model weights represent the probabilities of residing in the identified weather regimes, and results are presented here for three ways of defining them. The first is an indicator function that depends on the regime at the forecast initialisation time, the second is the proportion of ensemble members predicting each regime at the validation time, and the regime that actually materialises at the validation time is also implemented. Although the latter approach is not applicable in practice, it is regarded here as an upper bound for the improvement and hence provides a useful comparison. It could also be argued that if the true regime were known, then it might be more useful to condition on both the true regime and the forecast regime: if the forecast predicted one regime yet the actual regime was known to be different, then the biases would be larger than if the forecast and atmosphere agreed on the regime.

The regime-dependent approaches are implemented on wind speed forecasts in two scenarios: a quasigeostrophic model of the Northern Hemisphere, and on GEFS retrospective forecasts over the Euro-Atlantic sector. Regimes are identified by projecting the large-scale flow, represented by a synoptic-scale variable at all spatial locations on the domain, onto the leading three principal components, before detecting patterns in the resulting variables. A hidden Markov model is fitted to streamfunction anomalies in the QG setting, while k-means clustering is applied to geopotential height anomalies in the reforecast data set. The retrospective forecasts are generated from a higher resolution NWP model than that studied in the QG framework, but a data-rich simulation study is also helpful when trialling a new method since conclusions can be made that are more resistant to sampling variation. Nonetheless, the results found in the reforecast setting corroborate those in the QG study.

If a probabilistic approach is used to define the regimes at the initialisation time, as is the case in the QG study, then it is possible to use these posterior probabilities as weights in Equation 35. Results indicate that accounting for this uncertainty slightly improves the performance of the RDTN-init approach, though these results were not included to maintain comparison between the results in the different settings: such a probabilistic approach is not immediately possible when using k-means clustering to identify regimes without making additional assumptions regarding the regime behaviour. It would also be possible to use the training data to obtain conditional probabilities of each regime occurring given the ensemble member regime weights. This could itself be thought of as statistical post-processing applied to the forecast of the regime.

Forecasters have noted that knowing the prevailing synoptic behaviour of the atmosphere at the initialisation time can help to predict the forecast accuracy. It is found here that in order to benefit post-processing at longer lead times, it is not enough to know the behaviour at the initialisation time, instead a good estimate of the behaviour at the validation time is required. As is discussed in the previous chapter, this is due to the spread of the ensemble members accounting for the flow-dependent uncertainty, so that the raw ensemble forecast already contains the regime information provided by the initial state. As a result, using the regime defined at the forecast initialisation time contains little information regarding the true regime at longer lead times, and therefore, although there were small improvements for forecasts at short lead times, they were significantly less pronounced for longer forecast horizons. Using the ensemble members to predict the regime offered more skill than using the initial regime, though the benefit of using this approach again reduced to zero as lead time increased.

The skill of regime-dependent methods that used the true regime, on the other hand, appeared to increase with lead time, suggesting larger relative improvements upon conventional post-processing methods are available for forecasts further in advance. This is because, as the forecast lead time increases, the accuracy of the numerical model output naturally decreases. To account for this loss of information at later lead times, statistical post-processing methods typically downweight the influence of the numerical model output on the resulting predictive distribution (i.e. by decreasing the magnitude of β and δ in Equation 36). As a result, conventional methods issue forecast distributions that do not capture the variation in the observations due to changes in the regime. Regime-dependent methods that utilise the true regime, on the other hand, are capable of achieving this, and hence, as the information present in the ensemble member diminishes, they contain progressively more information than standard post-processing approaches.

If the numerical model output is not downweighted when post-processing then it may be more informative to know both the regime that occurs at the forecast validation time, as well as the regime that is predicted. This is illustrated in Figure 25, which displays rank histograms for the raw ensemble forecast in the quasigeostrophic setting of Section 4.3, for every possible combination of the forecast and observed weather regime, at a lead time of one week. The forecast regime in this case is taken to be that which is predicted by the largest number of ensemble members, and in the few instances where two or more regimes are equally popular, one is selected at random. The rank histograms along the leading diagonal correspond to when the forecasts accurately predict the observed regime at the forecast validation time. In this case, the forecasts are underdispersed, with a slight negative bias in the PNA+ regime, and a positive bias in the PNA-. The fact that these biases change between the regimes indicates that regime-dependent post-processing should be beneficial when the forecast accurately predicts the regime. The biases are much larger, however, when the regime is not correctly predicted. There is a considerable negative bias in the forecasts when the positive phase of the Arctic Oscillation occurs, or when the negative phase of the AO is predicted. Conversely, lower wind speeds than expected occur when the AO- regime occurs, or when the AO+ regime is forecast. There are also biases related to the phases of the PNA, though these are comparatively weaker. If the information present in the ensemble were not downweighted when post-processing, then both the forecast and observed regimes should be incorporated during the recalibration.

The fact that the weather regime that materialises at the forecast validation time (i.e. the true regime) can be used successfully within post-processing methods may seem trivial: the future atmospheric regime is not known in practice, and if it were predicted correctly by the numerical weather model then the forecast biases may depend less on the prevailing regime. However, the synoptic-scale atmosphere is considerably more predictable than the more turbulent surface weather, and hence it may be comparatively straightforward to obtain a more accurate forecast of the future weather regime. To do so, one might utilise, for example, relationships between certain weather regimes and larger-scale (and yet more predictable) circulation patterns, such as the Madden-Julian Oscillation (Cassou, 2008) and the El-Nino Southern Oscillation. The improvements gained from regime-dependent post-processing methods at longer lead times may be attainable in practice if such relationships could effectively be used to obtain an accurate forecast of the true regime.

A more accurate NWP model would likely be more adept at correctly identifying the regime at the forecast validation time. However, if the NWP model is used to

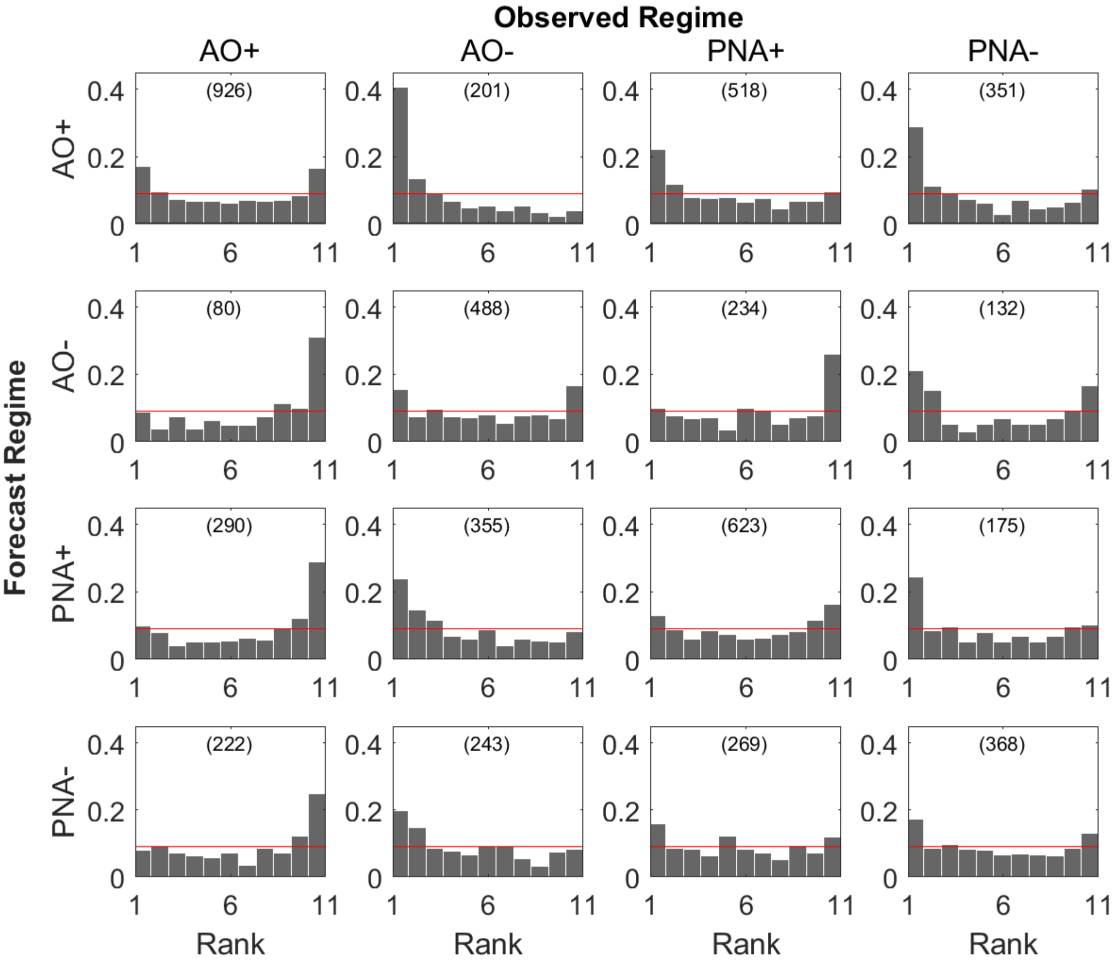

Figure 25: Rank histograms for the raw ensemble forecast in the Quasigeostrophic setting at a single grid point.
The location is the same as that for which results are presented in Section 4.3. Rank histograms are displayed for every combination of forecast and observed weather regime at a lead time of one week. The forecast regime in this case is that which is predicted by the most ensemble members, with ties between several regimes being decided at random. The number of instances from which the histogram is constructed is given in brackets on each plot.

identify the regime, then as the model produces more skilful forecasts of the large-scale circulation (from which the regimes can be identified), it may also provide better forecasts of other, smaller-scale variables, such as wind speed or temperature. The available improvements upon standard post-processing methods would therefore decrease as the biases in the model become smaller and less varied between the different regimes. This intuition also explains why the potential improvements of regime-dependent post-processing are small at short lead times: the magnitude of model biases are generally smaller and hence the differences between the regimes become insignificant. Nonetheless, Ferranti et al. (2015) show that high resolution ensemble prediction systems still exhibit biases that depend on atmospheric regimes, and hence there is reason to believe regime-dependent approaches will be useful when calibrating these more accurate forecasts. The GEFS reforecasts here were verified against model analyses, which may be subject to the same limitations as the prediction system. Since the NWP model may not correctly simulate the spatial and temporal characteristics of the observed weather regimes, evaluating forecasts against station observations may result in larger regime-dependent biases.

In addition, it may be the case that the choice of predictive distribution should vary with the regime and hence future work could investigate which distributions are most appropriate for certain weather types or situations. The numbers of regimes used in this chapter were chosen subjectively, using results from previous studies as guidance. Whether there exists a statistical procedure to estimate these regimes such that they are optimal for use in post-processing is also a topic for further research. Furthermore, it may be the case that the optimal regimes, or number of regimes, changes depending on the location or predictand under consideration. The extent to which a regime affects the wind speed at a certain location was found here to depend on its proximity to the regime centers of action. Each regime thus provided valuable information at some locations, but not others. If interest lies only in one location, then it may be preferable to estimate more localised, or even site-specific regimes, which could also vary for each predictand being forecast. This definition of a regime, however, differs slightly from synoptic-scale patterns in the atmosphere. The weather regimes considered here are advantageous because they are physically meaningful, which may not be the case for regimes estimated separately at every location for each variable. As a result, considerable work has been devoted to studying their dynamical and statistical properties, and such studies can be used to identify situations where the inclusion of regimes may be most beneficial. Previous work, for example, has noted the impact they have on local weather systems, and how they can account for the dependence between meteorological variables and multiple locations. They thus naturally lend themselves for use with post-processing in a spatial or multivariate context.

We therefore argue that the appropriate regimes, and number of regimes, should be investigated prior to post-processing, utilising previous studies of low-frequency variability in the domain under consideration. The number of regimes to use also depends on the amount of data available. Using a large number of weather types can result in overfitting of the training data, leading to less informative out-of-sample predictions. In the study reported in this chapter, estimating four times as many parameters as the original truncated normal approach did not induce any problems of this sort. Alternatively, atmospheric circulation could be incorporated into post-processing approaches without the discretisation into a finite number of regimes. It is found that improvements are only likely to be seen for regimes in which wind speeds differ severely from the average wind speed at a location. If the local weather is believed to be strongly dependent on one or two known regimes then using the continuous index for these patterns, if such indices exist, could be incorporated as additional covariates in the post-processing model. Since this requires fewer parameters to be estimated, it would be more feasible to implement with a rolling training window, or when reforecast data were not available.

We expect improvements from regime-dependent post-processing to be largest in winter, since this is when the regime-behaviour of the atmosphere is most pronounced. The regime-dependent model biases may themselves be dependent on the season. For example, blocking episodes are associated with heat waves in summer and cold snaps in winter and therefore temperature biases may be inconsistent between separate occurrences of the same regime. If all seasons are considered at once then these could be treated as separate regimes, despite corresponding to the same large-scale mode of variability. If a large number of regimes, or even smaller-scale weather patterns, were used then other latent variable methods, such as hierarchical models, may be more appropriate.

Lastly, as in Chapter 3, results here suggest that regime-dependent post-processing is particularly adept at calibrating forecasts corresponding to regimes in which the weather differs greatly from the local climatology. Further investigation into the use of regime-dependent approaches when forecasting extreme events would therefore complement previous comparison studies (Williams et al., 2014).

5 Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts

5.1 Introduction

Thus far in this thesis, regime-dependent post-processing methods have been proposed and trialled in a variety of situations in an attempt to understand when and where such methods would be beneficial. To do so, the post-processing methods have been applied to ensemble forecasts in large, potentially simulated archives of data. In practice, however, regular updates of the numerical model configuration or the data assimilation scheme are common, limiting the amount of data on which to train post-processing methods. The goal of this chapter is thus to apply regime-dependent post-processing approaches in an operational framework, where such a restriction on the available data is in place, utilising results in the previous two chapters to do so. Moreover, we demonstrate that the regime-dependent approaches can improve forecasts of extreme weather events, which, despite their obvious significance, are commonly overlooked when implementing conventional post-processing methods (Friederichs, 2010; Pantillon et al., 2018).

Established post-processing methods tend to use only the ensemble members, or ensemble mean and spread, as predictors in the statistical models. However, recent studies have highlighted the potential benefits of utilising several sources of information when post-processing (Taillardat et al., 2016; Messner et al., 2017; Rasp and Lerch, 2018). In Chapter 3, it is suggested that additional predictors could be selected using meteorological intuition. In particular, it is hypothesised that errors in numerical weather prediction (NWP) model output are dependent on synoptic-scale structures in the atmosphere, and hence this weather regime information should be utilised when post-processing. In doing so, the recalibration can remedy conditional biases owing to changes in the regime, which might otherwise be ignored.

Indeed, weather regimes are closely tied to the atmosphere's predictability (Thompson, 1957), and the effect they have on local weather systems is well-documented (Hannachi et al., 2017). They are thus, in themselves, of considerable interest to meteorological services. As such, it is common for weather forecasting centres to monitor the occurrence and behaviour of circulation patterns over the local domain. The UK Met Office, for example, have identified 30 synoptic patterns that influence the UK weather, which are now used operationally in their Decider program (Neal et al., 2016). It should therefore be straightforward to utilise this regime information in oper-

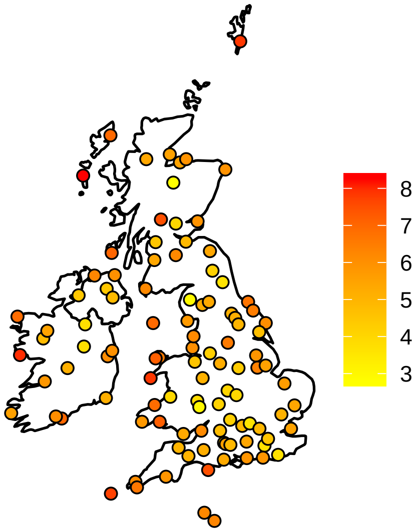

Figure 26: Station locations over the UK at which post-processing is performed. The average day-time wind speed (in m/s) is shown by the colour at each station, calculated across all days in 2018 and 2019.

ational post-processing, and this chapter demonstrates how this could be achieved. To do so, wind speed forecasts from the Met Office's global ensemble prediction system, MOGREPS-G, are subjected to regime-based post-processing approaches. The following section introduces the data available and describes the weather regimes considered throughout this chapter. The statistical post-processing methods are outlined in Section 5.3, and are analysed in Section 5.4. Finally, a summary is available in Section 5.5, along with a discussion of potential avenues for future work on regime-dependent post-processing.

5.2 Data

5.2.1 Forecasts and observations

This study utilises 10m wind speed forecasts extracted from the Met Office's global ensemble prediction system, MOGREPS-G (Walters et al., 2017; Porson et al., 2020), issued in the two year period between the 1st January 2018 and the 31st December 2019. No major model upgrades have been performed since this period and hence the model configuration is very similar to that currently in operation. The model employs a 20km horizontal resolution over the globe, with ensembles comprised of 18 constituent members. We consider here lead times at 12 hour intervals up to six days ahead. The forecasts are initialised at 12 UTC and thus validate at either midnight or midday, allowing the evaluation of both day- and night-time wind speed predictions.

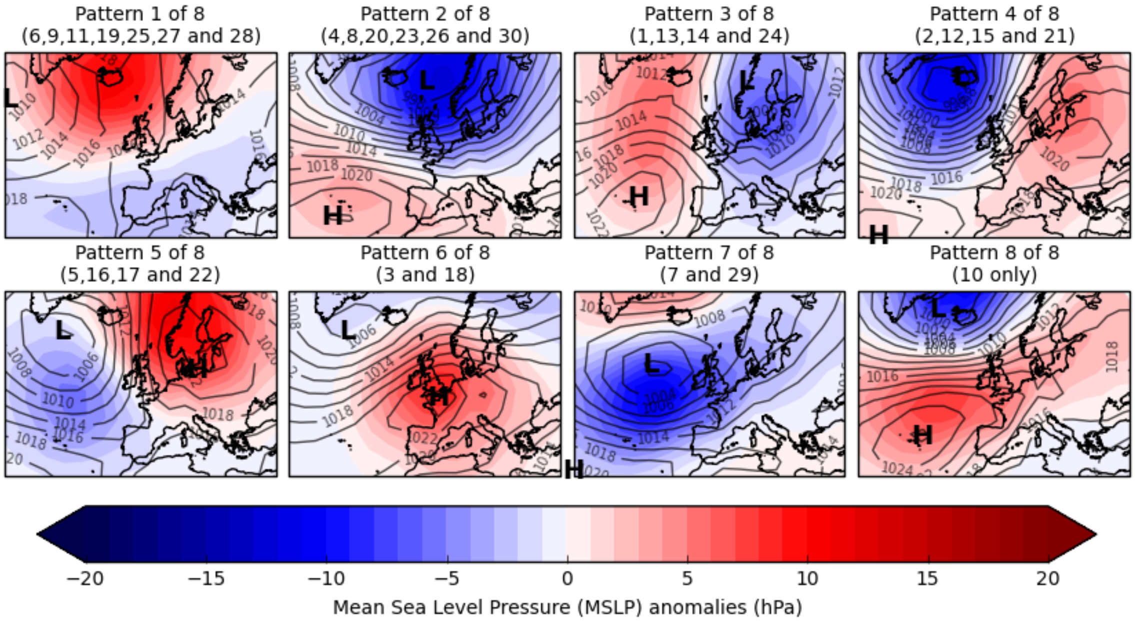

Figure 27: Mean sea level pressure anomaly fields for the 8 weather regimes in the Decider tool.
The component weather patterns are listed above each anomaly field. Taken from Neal et al. (2016) with permission.

Forecasts are evaluated at 106 stations over the UK and Ireland, pictured in Figure 26. Output from the grid-based MOGREPS-G prediction system is bilinearly interpolated to the stations prior to post-processing; the merits of post-processing using station data rather than weather model analyses are discussed in Hamill (2018) and Feldmann et al. (2019). The colour of the points in Figure 26 reflects the average wind speed at each location, from which it can be seen that larger average wind speeds tend to occur at coastal locations, particularly on the west coast, whereas inland stations record comparatively weaker winds.

5.2.2 Weather regimes

As discussed in Section 3.2, the synoptic-scale atmospheric flow can often be characterised by just a few circulation patterns, such that the continuous evolution of the atmosphere is represented by transitions between these distinct states (Franzke et al., 2011). These regimes recur at the same geographical locations and persist beyond the time scales of individual weather events, and can thus be thought of dynamically as quasi-stationary equilibria in the atmosphere's phase space (Charney and DeVore, 1979). Although a dynamical phenomenon, atmospheric regimes are regularly identified using statistical methods, which search for recurring patterns in archives of large-scale atmospheric variables (Hannachi et al., 2017).

The circulation patterns used here are those implemented in the Met Office's Decider tool, described in detail in Neal et al. (2016). That is, a set of 30 recurring weather patterns over the Euro-Atlantic region, detected using a simulated annealing clustering technique applied to 154 years of mean sea-level pressure (MSLP) anomaly fields (Philipp et al., 2007). Eight larger-scale circulation types, hereafter referred to as regimes, are then constructed by objectively grouping similar weather patterns in such a way that patterns are matched with those that exhibit similar structural properties, so that zonal flow patterns, for example, across different seasons may be clustered together. Figure 27 depicts the eight MSLP anomaly fields corresponding to the centres of the weather regimes under consideration. The regime at any time is defined as that which minimises the distance between the regime centre and the instantaneous MSLP anomaly field. Descriptions of the eight regimes in relation to the flow over the UK are available in Neal et al. (2016).

The regimes in Figure 27 are displayed in order of decreasing frequency, and their mean persistence times during the two year period considered here range from 1.2 to 2.8 days, indicating fairly transient behaviour that corroborates findings in Neal et al. (2016). The distributions of wind speeds across all locations are displayed for each regime in Figure 28. There is some clear dependence on the regimes, particularly between Regimes 1 and 2, which represent, respectively, the negative and positive phase of the North Atlantic Oscillation (NAO). The NAO is known to have a profound effect on the UK (and European) weather, with more extreme wind speeds related to the occurrence of positive NAO events (Hurrell, 1995; Hurrell and Deser, 2009).

Furthermore, if there is insufficient data available on which to reliably train the post-processing methods, then they may overfit the training data, resulting in poor out-of-sample predictions. By including additional predictors, regime-dependent post-processing methods are less parsimonious than conventional approaches, and hence more susceptible to overfitting. This is especially pertinent in operational settings, where the amount of available data is limited due to continual model upgrades. To circumvent this, the eight regimes are condensed into just three. The opposite phases of the NAO each have a significant effect on the observed wind speeds, whereas the variation between the remaining six regimes is comparatively weak. Moreover, some of the latter regimes occur fairly infrequently during the time period under examination. Therefore, the remaining six regimes are grouped together; the three regimes considered here are thus the NAO- and NAO+ regimes, as well as a third regime combining Regimes 3-8 of Figure 27. In the two year time period of interest, the NAO- regime occurs on 22.1% of the 730 days, the NAO+ regime occurs on 16.3%, and the third regime on the remaining 61.6%. It is possible to utilise all eight regimes when post-processing, but such an approach is found here to induce parameter uncertainty in the recalibration

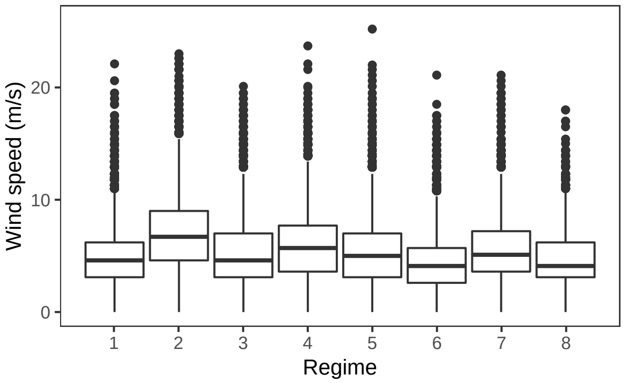

Figure 28: Boxplots of the wind speed distribution across all locations at 12 UTC when the atmosphere resides in each regime.
The boxes display the lower quartile, median, and upper quartile of the observed wind speeds. Values that exceed the upper quartile plus 1.5 times the interquartile range are plotted as points.

methods that outweighs the benefits of including the regime information (not shown).

Although Figure 28 demonstrates that the wind speed depends on the prevailing phase of the NAO, it may be the case that the MOGREPS-G ensemble forecasts capture this dependence, rendering regime-dependent post-processing methods superfluous. To assess whether or not this is the case, we evaluate the calibration of the forecasts in each regime. For this purpose, we employ the reliability index, defined as

$$\Delta = \sum_{m=1}^{M+1} \left| \rho_m - \frac{1}{M+1} \right| \tag{48}$$

(Delle Monache et al., 2006), where M is the size of the ensemble, and ρ_m denotes the relative frequency with which the observed wind speed exceeds $m-1$ of the ensemble members. If the ensemble prediction system is calibrated, then the verifying wind speed should be equally likely to fall between any two members of the ensemble, meaning $\rho_m = 1/(M+1)$, for all $m = 1, ..., M+1$. The reliability index thus measures the divergence between the relative frequencies and their ideal value, with a more reliable prediction system attaining a lower index.

To evaluate whether or not biases exist in the MOGREPS-G ensemble forecasts owing to the occurrence of certain weather regimes, we calculate this reliability index separately for forecasts associated with each regime under consideration. Figure 29 illustrates how this index varies for forecasts corresponding to the two phases of the NAO. If the prevailing regime does not influence the errors in the prediction system, then the reliability index should not change depending on the phase of the NAO, meaning points would lie along the dotted line of equality in Figure 29. The index is

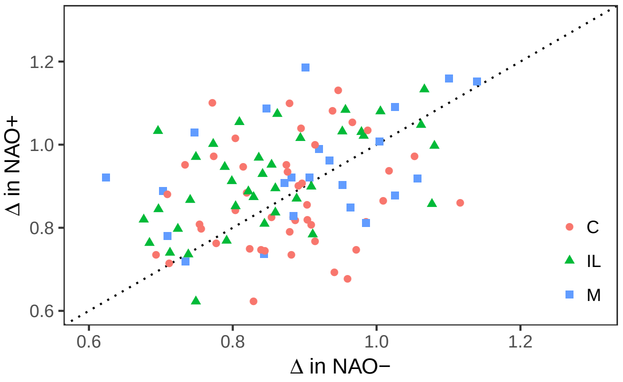

Figure 29: Reliability index (Equation 48) calculated over forecasts defined to be in the positive phase of the NAO, plotted against that calculated over forecasts defined to be in the negative phase of the NAO, and shown for each of the 106 weather stations at a lead time of 24 hours. The stations have been classified into coastal (C), inland (IL), and mountainous (M) locations.

substantially greater than zero in both the NAO- and NAO+, reflecting the underdispersive nature of the forecasts, and, although there is weak positive correlation between Δ in the two regimes (0.369), substantial deviation from the diagonal advocates the implementation of regime-dependent post-processing methods.

Moreover, the reliability index is not distributed evenly about the diagonal in Figure 29, with larger forecast errors typically occurring in the positive phase of the NAO than in the NAO-. Regime-dependent post-processing may therefore be more beneficial for forecasts associated with this cyclonic regime. The miscalibration also varies considerably with the spatial location, suggesting localised post-processing methods are desirable, while the variation in the reliability index between the two regimes also depends substantially on the station, even over the relatively small spatial domain considered here. Hence, although there is evidence to suggest regime-dependent post-processing is advantageous in this instance, it may not be necessary at all locations. To better understand the spatial properties of the forecast errors, the weather stations in Figure 28 have been classified into coastal, inland, and mountainous locations, though it appears that regime-dependent biases in the MOGREPS-G output occur for all types of locations at this lead time.

5.3 Statistical post-processing

5.3.1 Training and testing

Statistical post-processing seeks to exploit systematic errors in previous forecasts to address those that might occur in the current prediction. To do so, a training set of historical forecasts and observations is required, where the choice of training data should reflect the biases expected to manifest in the current forecast. Training windows that adapt in the presence of new data assume that the error in the current forecast will likely behave similarly to that of recent forecasts. Rolling, or sliding, windows, for example, use only the most recent forecast-observation pairs available to train the post-processing model. The length of the window is a compromise between using enough data to obtain reliable parameter estimates, and not using too much data, so as to capture the recent behaviour of the model biases.

The limited amount of data used in a rolling window can often result in unstable parameter estimates (Scheuerer, 2014; Lang et al., 2019a). Although the window length can be extended to avoid this, continually increasing the length contradicts the motivation for using a rolling window. Once the window can no longer account for the seasonal cycle in the data, it may be more sensible to use a larger amount of data, spanning several seasons or years. This data set could be fixed to obviate the need to update parameter estimates for every new forecast. In this study, since the observed wind speeds do not exhibit a strong annual cycle, the additional training data afforded by a fixed window is found to outweigh the adaptive nature of a rolling window (not shown). Therefore, the coefficients of all post-processing models described in this section are estimated over a fixed window between the 1st January 2018 and the 31st December 2018. The remaining year of data is used as a test data set, on which to verify the resulting forecasts.

Although there is little seasonal variability in the model biases, Figure 29 illustrates that there is a considerable spatial dependency. Therefore, to account for locally-varying biases, post-processing is performed separately at each station under consideration. Despite being more computationally expensive, this method performs significantly better than a global post-processing approach, in which one set of model parameters is estimated for every location, and the site-specific post-processing approach is thus favoured here.

Just as the post-processing methods are trained using station observations, the resulting forecasts are also verified against wind speed measurements at the locations of interest. The resulting forecast distributions are assessed using the continuous ranked probability score (CRPS). For a forecast with predictive cumulative distribution function (CDF) F and corresponding observation y, the CRPS is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{u \geq y\}]^2 du \tag{49}$$

(Matheson and Winkler, 1976). The CRPS is negatively oriented, so that a lower score indicates a more accurate forecast, and, as a proper score, it assesses both the reliability and sharpness of the forecast (Gneiting and Raftery, 2007). The total CRPS is then taken to be the average CRPS over all forecasts in the test data. The continuous ranked probability skill score (CRPSS) is also used to measure the skill of the regime-dependent approaches relative to conventional post-processing. The CRPSS can be interpreted in terms of the percentage improvement in the forecast accuracy relative to a baseline forecast, and hence larger values are desired (Wilks, 2019).

The CRPS and its skill score are commonplace in weather forecasting. In the following section, rank and Probability Integral Transform (PIT) histograms (Hamill and Colucci, 1997; Dawid, 1984; Gneiting et al., 2007), as well as the coverage of 90% prediction intervals, are also used to assess the reliability of forecasts. The sharpness, or resolution of the predictive distribution is then considered using the width of the 90% prediction intervals. A smaller width indicates a sharper, or more refined forecast, but is desirable only subject to calibration (Gneiting et al., 2007).

In Chapter 4, it is suggested that if extreme wind speeds, or wind gusts, can be linked to the occurrence of certain weather patterns, then regime-dependent post-processing methods may be capable of improving forecasts of these high-impact weather events. Therefore, predictive performance in the upper tail of the forecast distribution is evaluated using the threshold-weighted continuous ranked probability score (twCRPS). The twCRPS is defined as

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{u \geq y\}]^2 \omega(u) du \tag{50}$$

for some non-negative weight function $\omega(\cdot)$ (Gneiting and Ranjan, 2011). As in Lerch and Thorarinsdottir (2013), interest lies in the upper tail of the forecast distribution, and hence the weight function used here is $\omega(u) = I(u \geq t)$, for some threshold t. When such a weight function is used, the twCRPS is a strictly locally proper scoring rule (Holzmann et al., 2017) that considers only the performance of the forecast distribution above this threshold, and it thus concerns predictions of the upper tail behaviour.

5.3.2 Non-homogeneous Regression

Non-homogeneous Regression (NR), or Ensemble Model Output Statistics (EMOS), is possibly the most frequently implemented post-processing approach, owing largely to its simplicity and the ease with which it can be modified for use in different situations. NR assumes that the weather variable to be forecast, referred to as the predictand or response variable, follows a statistical distribution that depends on the raw ensemble output. The choice of predictive distribution is determined by the weather variable of interest: wind speed, for example, is non-negative, and hence a sensible choice would be a forecast distribution with a positive support, such as a gamma (Sloughter et al., 2010), truncated normal (Thorarinsdottir and Gneiting, 2010), or truncated logistic (Messner et al., 2014; Scheuerer and Möller, 2015) distribution. The latter is found here to outperform alternative options.

We therefore assume that the future wind speed is a random variable, Y, that follows a truncated logistic distribution with location and scale that depend on the ensemble mean, $\bar{x}$, and variance, s^2, respectively:

$$Y|\boldsymbol{x} \sim L_0(\alpha + \beta\bar{x}, \sqrt{\gamma + \delta s^2}), \tag{51}$$

where $\boldsymbol{x}$ denotes the vector of ensemble members, and $L_0(\mu, \sigma)$ denotes the logistic distribution that has been truncated below at zero, with location parameter μ and scale parameter σ. Hence, the square of the scale is expressed as a linear function of the ensemble variance, while truncation at zero ensures the resulting forecast distributions assign mass only to non-negative wind speed values. The post-processing parameters (α, β, γ and δ) are estimated by finding those that minimise the total CRPS over the training data, and, to ensure positive variance components, optimisation is performed using $\xi = \sqrt{\gamma}$ and $\kappa = \sqrt{\delta}$ rather than γ and δ directly.

The CRPS for a truncated logistic distribution with location μ and scale σ can be given in closed form:

$$\mathrm{CRPS}(L_0(\mu, \sigma), y) = \sigma \left(\log \frac{1 - p_0}{1 - p_y} - \frac{p_0^2}{(1 - p_0)^2} \log p_0 - \frac{1 + p_0}{1 - p_0} \log p_y - \frac{1}{1 - p_0} \right). \tag{52}$$

This expression is identical to, but more compact than, that derived by Scheuerer and Möller (2015). In keeping with the notation therein, p_0 and p_y are the values of the logistic, $L(\mu, \sigma)$, CDF evaluated at 0 and y, respectively. The total CRPS is then the average CRPS across all forecast-observation pairs.

The twCRPS can similarly be derived in closed form. Our result, verified against numerical integration, is that

$$\mathrm{twCRPS}(L_0(\mu, \sigma), y) = \frac{\sigma(p_t - 1 - \log p_t)}{(1 - p_0)^2} \tag{53}$$

if $y \leq t$, and

$$\mathrm{twCRPS}(L_0(\mu, \sigma), y) = \sigma \left(\log \frac{1 - p_t}{1 - p_y} - \frac{p_0^2}{(1 - p_0)^2} \log p_t - \frac{1 + p_0}{1 - p_0} \log p_y - \frac{1 - p_t}{(1 - p_0)^2} \right) \tag{54}$$

if $y > t$. Here, p_t is the logistic CDF evaluated at the threshold t. Again, the twCRPS values of the individual observations are averaged to give the total twCRPS of the data set.

Although interest here is on the logistic distribution truncated below at zero, we remark that Equations 52, 53 and 54 can be generalised to any truncation point l of the logistic distribution by replacing p_0 with p_l, the CDF of the logistic distribution at l. As $l \to -\infty$ and $p_l \to 0$, the logistic distribution $L(\mu, \sigma)$ is recovered from the truncated logistic distribution. Correspondingly, Equation 52 tends to the well-known CRPS for the logistic distribution,

$$\mathrm{CRPS}(L(\mu, \sigma), y) = -\sigma[1 + \log p_y + \log(1 - p_y)] = y - \mu - \sigma - 2\sigma \log p_y, \tag{55}$$

given by Taillardat et al. (2016) and Jordan et al. (2017). Similarly, the twCRPS of the logistic distribution can be recovered from Equations 53 and 54:

$$\mathrm{twCRPS}(L(\mu, \sigma), y) = \sigma(p_t - 1 - \log p_t) \tag{56}$$

if $y \leq t$, and

$$\mathrm{twCRPS}(L(\mu, \sigma), y) = \sigma \left(p_t - 1 + \log \frac{1 - p_t}{1 - p_y} - \log p_y \right) \tag{57}$$

if $y > t$, a result that, to our knowledge, has not previously been given in the literature.

5.3.3 Regime-dependent mixture model

Mixture models have previously been shown to be an effective way of combining information from several sources in a weather forecasting context (Wilks, 2002; Gneiting et al., 2013; Baran and Lerch, 2016, 2018). In Chapter 3, we therefore propose a mixture model to include regime information when post-processing. Mathematically, this involves extending Equation 51 to

$$Y|\boldsymbol{x} \sim \sum_{r=1}^{R} w_r L_0(\alpha_r + \beta_r \bar{x}, \sqrt{\gamma_r + \delta_r s^2}), \tag{58}$$

where R is the number of regimes under consideration: in this case, $R = 3$. A separate forecast distribution is associated with each identified regime, and, making the same distributional assumptions about the predictive distribution in each regime, there are $4R$ parameters to estimate, one set for each regime. The weights associated with each regime, w_r, allow the model to account for uncertainty present when attributing the forecast to a regime. It is important to note that the weights in this case are functions of time, rather than parameters as such, highlighting that the weights will change depending on the prevailing behaviour of the atmosphere. Potential approaches to calculate the mixture model weights are discussed in Chapter 4, where it is found

		Observed regime			
		NAO-	NAO+	Other	Total
Forecast regime	NAO-	111	1	57	169
	NAO+	0	72	47	119
	Other	49	45	348	442
	Total	160	118	452	730

Table 10: Contingency table for the forecast and observed regimes at a lead time of six days.

that, in comparison with alternative choices, the regimes predicted by high resolution numerical weather models provide a reasonable estimate of the future synoptic-scale state.

As well as recording the regime that manifests, the Met Office stores the daily regime that is forecast by their global deterministic model, up to six days in advance. Therefore, the mixture model weight used here is an indicator function that takes the value one only for the regime that is forecast by this deterministic model. In this case, the mixture model reverts to a truncated logistic distribution with the appropriate coefficients based on the regime that is forecast. Parameter estimation and forecast verification can thus again be performed using the CRPS and twCRPS as given in Equations 52, 53 and 54.

Although the large-scale atmospheric state is considerably more predictable than the more turbulent weather, this forecast of the regime may itself exhibit biases. Table 10 presents the number of instances that each regime is predicted and observed in the 730 days under consideration, shown at a lead time of six days. Even this far in advance, the numerical weather model only once mistakes the phase of the NAO, with the deterministic global forecast correctly estimating the future regime on roughly 70% of the 730 days. It is possible to address this error by applying a simple form of post-processing to these regime forecasts: the mixture model weight corresponding to a particular forecast could be given by the relative frequency of each regime occurring given the regime that is forecast. For example, if the NAO- regime is forecast to occur in six days time, then Table 10 suggests that the mixture model weight corresponding to this regime should be 0.657 (111/169), whereas those for the NAO+ and ‘Other’ regimes should be 0.006 (1/169) and 0.337 (57/169), respectively. Since this weight is not an indicator function, the post-processing coefficients associated with each regime would be estimated simultaneously.

The ability of these two weight functions to forecast the future weather regime is assessed using the Brier score (Brier, 1950) in Figure 30. Results are also shown for a persistence forecast constructed from the regime at the forecast initialisation time, and the climatological regime frequencies. From Figure 30, we see that the initial regime

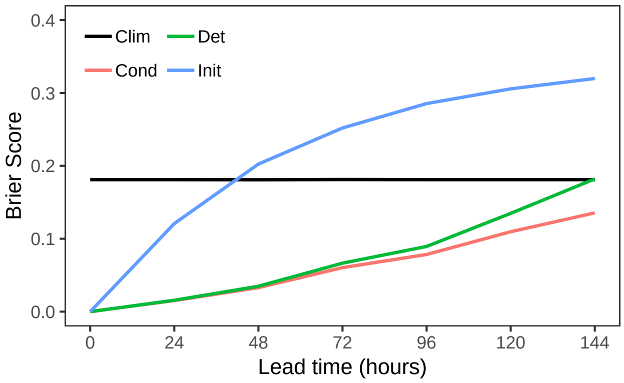

Figure 30: Brier score against lead time for the climatological frequencies of the regimes (Clim), the initial regime (Init), the deterministic forecast regime (Det) and the conditional regime probability (Cond), calculated between 2010 and 2016.

becomes no better than climatology after just two days, highlighting the transient behaviour of the regimes considered here. The deterministic regime forecast, however, maintains skill until six days in advance, at which point the recalibration becomes beneficial. The recalibrated regime forecast is constrained to be always as good as climatology, and hence will, in theory, always have non-negative skill, even after the deterministic forecast on which it is based is completely uninformative.

Nonetheless, the forecasts resulting from the two choices of weight perform comparably (not shown), and hence results in the following section are for the simpler, deterministic regime forecast. Since we are post-processing wind speed forecasts issued by a global ensemble prediction system, it would also be possible to derive a forecast of the future regime from each of the ensemble members, and use the proportion of members predicting each regime to define the mixture model weights. However, these regime forecasts have not been stored for the prediction system and date range under consideration, and hence are not readily available. Therefore, although the distribution of the ensemble members would capture some of the flow-dependent uncertainty in the future regime, making them a sensible option when post-processing in real time (see Chapter 4), the regime predicted by the deterministic global model is used to define the weights in this study. Using a weight that is an indicator function means parameters can be estimated separately by stratifying the training data into regime-dependent subsets depending on the regime that is forecast, as described in Chapter 3. This is more numerically stable than estimating all parameters simultaneously, and hence also less computationally expensive (Baran and Lerch, 2018).

5.3.4 A hybrid approach

The regime-dependent mixture model is more complex than the truncated logistic forecast distribution, and is thus more susceptible to overfitting in the presence of limited training data. Therefore, the mixture model, although offering more flexibility, could potentially hinder forecast performance when the model biases do not depend on the regimes under consideration. A more robust approach might involve identifying situations in which regime-dependent post-processing is expected to be most beneficial, and implementing it only on such occasions. This emulates ideas in Lerch and Thorarinsdottir (2013), whereby a separate post-processing model is applied depending on the prevailing circumstances. Lerch and Thorarinsdottir (2013), and later Baran and Lerch (2015), implement a post-processing model that switches between two predictive distributions depending on whether the ensemble median exceeds a predefined threshold. In doing so, the forecast can better adapt to the biases expected of the current model output.

It remains to identify circumstances in which regime-dependent post-processing should be beneficial. In the previous chapter, we consider the spread of the average wind speeds between regimes at a variety of locations, and show that improvements gained by including the regime information are correlated with this between-regime spread. That is, locations whose wind speeds are more heavily influenced by changes in the weather regime tend to benefit most from regime-dependent post-processing. Therefore, we calculate here a similar measure of regime-dependency from the observations in the training data, and record whether or not it exceeds a certain threshold. The measure of regime-dependency is taken to be the between-group component of the empirical law of total variance applied to the wind speed observations in the training data:

$$\sum_{r=1}^{R} \frac{n_r}{n} (\bar{y}_r - \bar{y})^2, \tag{59}$$

where n_r/n is the proportion of days on which regime r occurs in the training data, $\bar{y}_r$ is the average wind speed given regime r, and $\bar{y}$ is the overall mean wind speed. The measure is shown for each location in Figure 31. The stations have been divided into coastal, inland, and mountainous locations, from which it can be seen that, just as locations on the west coast of the UK and Ireland were earlier identified as being associated with higher average wind speeds, they are also particularly affected by the phase of the North Atlantic Oscillation.

The measure of regime-dependency introduced here is similar to that used in Chapter 4, but can better account for the different climatological frequencies of the regimes. Larger values suggest a stronger regime influence and hence advocate the use of regime-

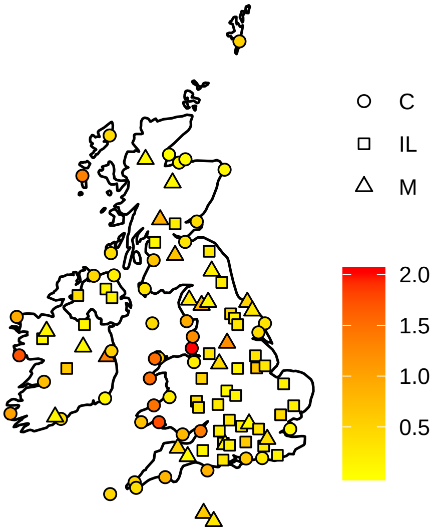

Figure 31: The regime-dependency (m^2s^{-2}; Equation 59) of the day-time wind speeds at each station on the domain under consideration. The points have been classified into coastal (C), inland (IL), and mountainous (M) locations.

dependent post-processing, whereas if this quantity lies below the chosen threshold, then the truncated logistic distribution is issued as the forecast, and the regime information is ignored. Of course, there are several alternative ways to measure the regime-dependency that may be more indicative of situations where regime-dependent post-processing is desirable. A brief comparison of possible measures is performed in the following section.

Lerch and Thorarinsdottir (2013) estimate the threshold for their combination-based method by finding that which minimises the CRPS in the training data. This approach fails here since the motivation for pooling together the two post-processing methods is to avoid situations in which the mixture model overfits the training data; in the training data, the mixture model almost always performs better than the truncated logistic method, and hence the optimal threshold in terms of the CRPS is zero, suggesting regime-dependent post-processing should always be implemented. However, this does not necessarily translate to better out-of-sample forecasts, and an alternative method of choosing the threshold is thus required. The threshold used here is the same for all lead times, and is chosen subjectively such that the regime-dependent mixture model is applied to roughly 50% of forecasts.

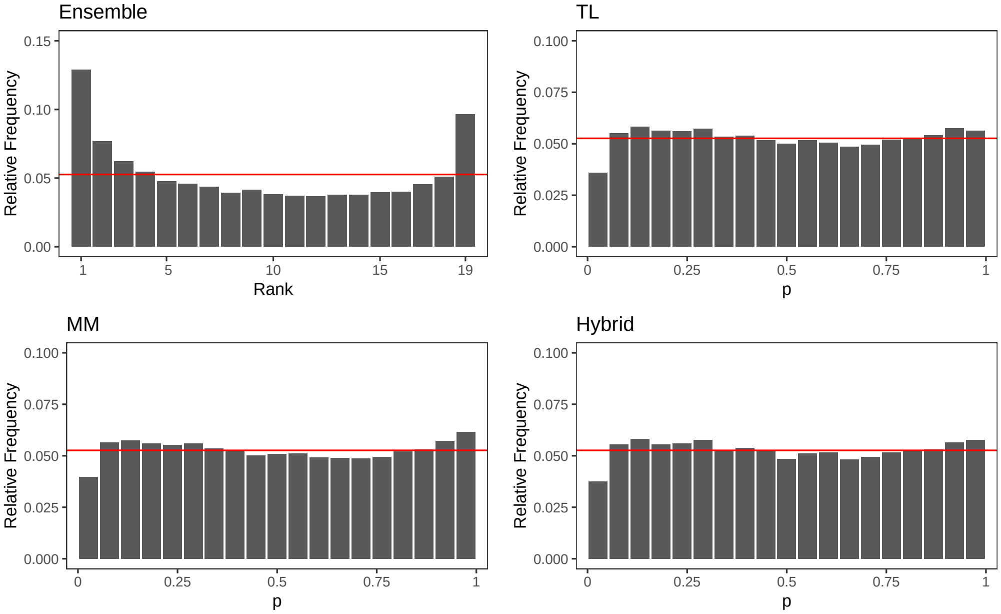

Figure 32: Rank and Probability Integral Transform (PIT) histograms for the raw ensemble forecast, along with the three post-processing methods, shown at lead time of 4 days. A horizontal red line is drawn at 1/19, indicating perfect calibration.

5.4 Results

Firstly, consider the calibration of the post-processed forecasts. Figure 32 shows rank and probability integral transform (PIT) histograms for the ensemble forecast, and for the various post-processing methods. The raw ensemble is again found to be underdispersed, with the observation falling either above or below all ensemble members with a disproportionately high frequency, whereas the post-processing methods yield PIT histograms that are approximately uniform, indicating well-calibrated forecasts. There appears, however, to be some systematic deviation from uniformity in the tails of the post-processed predictive distributions, though this is not remedied when using alternative families of parametric distributions.

Table 11 displays the average coverage and width of the 90% prediction intervals obtained from the various post-processing methods. The truncated logistic distribution generates a coverage that is close to, but slightly larger than, the optimal 90% coverage, and the regime-dependent approaches improve on this slightly. We see, however, that the standard post-processing method tends to be overdispersed in the more predictable NAO- regime, but underdispersed in the NAO+, suggesting it does not fully capture the changes in predictability arising due to the regimes. The regime-dependent approaches decrease the spread of the forecast distribution in the NAO- regime and increase the spread in the NAO+ regime, thus yielding a coverage closer to the optimal value. Being

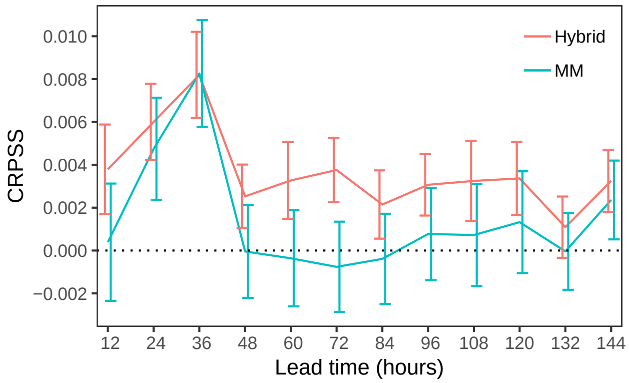

Figure 33: CRPSS for the regime-dependent mixture model forecast (MM), as well as the hybrid approach, relative to the conventional truncated logistic forecast distribution, shown against lead time.
Error bars indicate 95% confidence intervals for the skill score, obtained via non-parametric bootstrap resampling.

a combination of the two, the hybrid method exhibits prediction intervals that are a compromise between those of the truncated logistic distribution and the mixture model.

Looking now at the improvement in skill gained by including regime information, Figure 33 shows the continuous ranked probability skill score (CRPSS) against forecast horizon for the two regime-dependent methods, with the truncated logistic forecast used as a baseline. The score is averaged over all locations and forecast instances in the test data set. Neither method performs consistently worse than the standard approach, though the improvements are always relatively small. This is to a lesser extent when the hybrid method is used, despite the CRPSS being constrained to equal zero at a large proportion of locations. The improvements are largest at short lead times, before receding to zero as the forecast horizon increases. As is apparent in Figure 30, the quality of the regime forecasts deteriorates with lead time, and hence the biases in the forecast become less dependent on the regime that is predicted by the deterministic

		NAO-	NAO+	Other	Overall
Coverage	TL	92.03	87.69	91.56	91.01
	MM	90.26	89.86	90.17	90.14
	Hybrid	91.02	88.47	91.16	90.67
Width	TL	5.61	6.09	5.76	5.78
	MM	5.37	6.66	5.50	5.67
	Hybrid	5.43	6.28	5.66	5.71

Table 11: Average coverage (%) and width (m/s) of 90% prediction intervals derived from the truncated logistic (TL), mixture model (MM), and hybrid forecasts at a lead time of 24 hours.

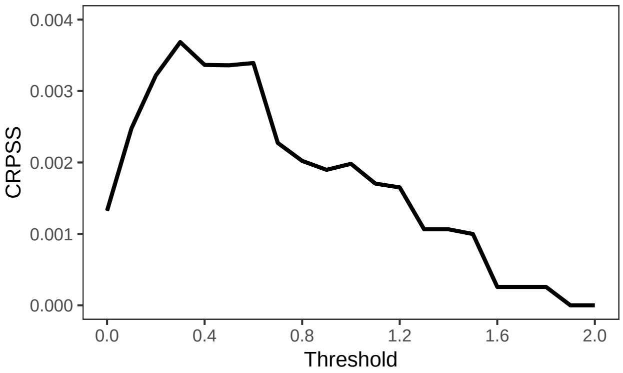

Figure 34: CRPSS of the hybrid approach relative to conventional post-processing for different choices of threshold, at a lead time of 5 days.
The threshold relates to the value of the regime-dependency defined in Equation 59, above which the regime-dependent mixture model will be implemented.

weather model, meaning there is less benefit to post-processing using this definition of the regime. As was the case in Chapter 4, at these longer lead times it is found that significant improvements are available if post-processing is conditioned on the true regime at the forecast validation time (not shown); the forecast accuracy across all locations improves by almost 1% in this case. Although this future regime is unknown in practice, this result highlights that a more accurate prediction of the future regime would be informative when post-processing medium- and long-range weather forecasts.

Despite its simplicity, the threshold for switching between the truncated logistic distribution and the mixture model appears to distinguish well between situations when regime-dependent post-processing is, and is not, desirable. The skill score for the hybrid approach is shown in Figure 34 for a range of possible threshold choices at a lead time of five days. A threshold of zero corresponds to always implementing the regime-dependent mixture model, and, conversely, as the threshold increases, this is equivalent to always enforcing the truncated logistic forecast, and hence never utilising regime information. As a result, the CRPSS tends to zero for higher thresholds. A threshold of 0.4 was used here, which appears suitable at this lead time.

The effect of changing the threshold can also be seen in Figure 35. Since the post-processing methods are trained using a fixed, site-specific window, each location is associated with only one value of the regime-dependency (Equation 59). The top panel of Figure 35 shows the forecast skill for the mixture model at each location under consideration, plotted against this measure of regime-dependency. A vertical line is drawn at the chosen threshold. As anticipated, locations whose wind speeds are highly dependent on the prevailing NAO phase tend to exhibit larger skill scores, with improvements at individual locations reaching 4% at this lead time. A distinction has also been made

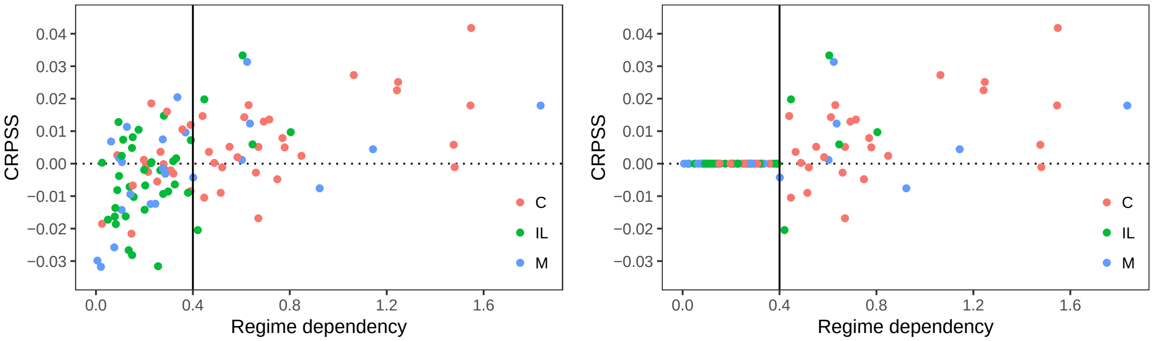

Figure 35: CRPSS of the regime-dependent mixture model (left) and hybrid (right) approaches relative to conventional post-processing, plotted for each of the 106 locations against the climatological, between-regime wind speed variation (Equation 59). Shown at a lead time of 5 days. Points are coloured depending on their classification into Coastal (C), Inland (IL) and Mountainous (M) locations. A solid vertical line is drawn at the chosen threshold used for switching between the truncated logistic and mixture model forecast distributions.

between coastal, inland, and mountainous sites, from which it can be seen that regime-dependent post-processing is typically most beneficial at locations on the coast of the UK and Ireland for this lead time. However, there are several negative skill scores to the left of the chosen threshold in Figure 35, suggesting the mixture model tends to do more harm than benefit at sites where the regime-dependence is weak.

The lower panel of Figure 35 shows the analogous plot for the hybrid approach. Below the threshold, the skill scores are zero, since the same forecast is being issued as the reference. This removes the largely negative effect the more complex model has at these locations. Conversely, the larger, positive skill scores that tend to occur at stations heavily influenced by the regimes are still present. This approach thus combines the benefits of the established post-processing with the more complex, yet more flexible, regime mixture model. Note, however, that some negative skill scores still occur for the hybrid approach, and some positive skill scores below the threshold are nullified. This could be avoided if an alternative way to choose the threshold exists that can better recognise when regime-dependent post-processing is desirable.

With this in mind, we look to compare various choices of the measure of regime-dependency. Six different measures are evaluated, all of which assume the following form:

$$\sum_{r=1}^{R} \frac{n_r}{n} (\bar{z}_r - \bar{z})^2, \tag{60}$$

where $\bar{z}$ denotes the mean of a quantity z calculated over the entire data set, while $\bar{z}_r$ is the average of the same quantity estimated only from forecast-observation pairs

	WS	B	SE	AE	EV	RI
z	y	$\bar{x} - y$	$(\bar{x} - y)^2$	$\lvert\bar{x} - y\rvert$	s^2	$\sum_{m=1}^{M} \lvert\rho_m - \frac{1}{M+1}\rvert$
SRCC	0.503	0.321	0.305	0.243	0.305	0.072

Table 12: Spearman's rank correlation coefficient (SRCC) between various measures of regime-dependency and the relative improvement gained by regime-dependent post-processing, as measured using the CRPSS.
Results are shown at a lead time of five days. The measures considered are the variations in the following quantities arising due to changes in the regime (Equation 60): the wind speed observations (WS), the bias of the ensemble mean forecast (B), the squared error of the ensemble mean forecast (SE), the absolute error of the ensemble mean forecast (AE), the ensemble variance (EV) and the reliability index of the raw ensemble forecasts (RI).

associated with regime r. Hence, all of these measures of regime-dependency represent the amount of variation in a metric that can be explained by changes in the regime. Clearly, Equation 59 is a particular example of this with z equal to the observed wind speed. In addition to the regime-dependent variation of the wind speed itself, we also consider the variation in the bias of the ensemble mean forecast, the squared error of the ensemble mean forecast, the absolute error of the ensemble mean forecast, the ensemble variance, and the reliability index of the raw ensemble forecast (Equation 48). These are summarised in Table 12.

For use within the hybrid method proposed here, a measure is desired that generates a monotonic relationship between the regime-dependency and the improvement gained by the regime-dependent mixture model relative to the conventional truncated logistic EMOS approach. Hence, to assess the utility of the various measures, we report Spearman's rank correlation (Wilks, 2019) between each measure and the improvement, as quantified using the CRPSS; a higher absolute correlation indicates a more monotonic relationship. To maintain comparison with Figure 35, Table 12 displays the results at a lead time of five days. The measures of regime-dependency have been calculated over the training data, since this information would be available to practitioners at the time of forecasting. Although post-processing is concerned with forecast errors rather than the observations themselves, the improvements are most strongly correlated with variations in the observed wind speed owing to the regimes. This highlights the extent to which climatological information is used by the post-processing model to recalibrate the ensemble forecasts. This is particularly pertinent at longer forecast horizons, where less information is contained in the ensemble prediction system, and hence post-processing methods should decrease the influence that the numerical forecasts assert on the resulting predictive distribution and use instead more information from the climatological distribution of the wind speeds.

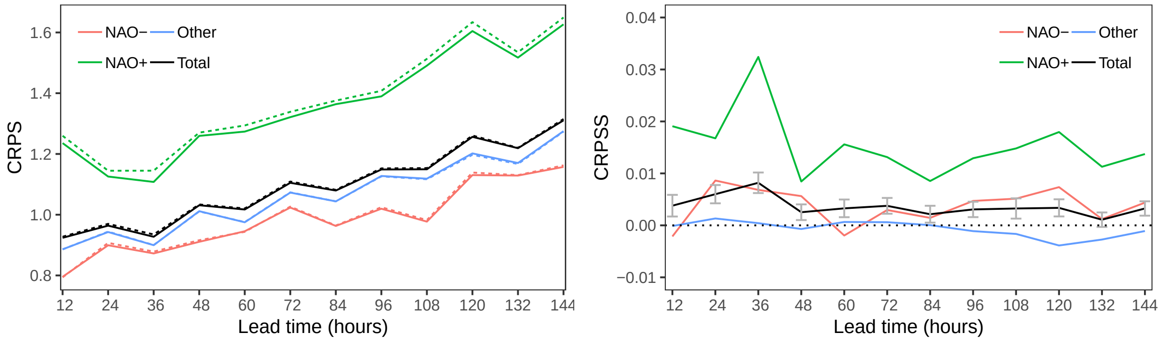

Figure 36: Left: CRPS for the hybrid approach (solid) and the truncated logistic forecast (dashed) when each regime occurs. Right: CRPSS for the hybrid approach relative to the truncated logistic forecast when each regime occurs.
Both metrics have been averaged over all stations, and 95% confidence intervals for the total skill score are shown as grey error bars.

Moreover, since Equation 59 is not dependent on the forecasts, it would be possible to calculate this quantity over a larger set of historical observations, rather than only the training data. This is not possible if forecast biases were used to calculate this measure of regime-dependency instead. In any case, the bias in the ensemble mean forecast exhibits a lower correlation with the CRPSS, and this decreases further when considering the squared or absolute forecast error. Similar conclusions are drawn for other lead times. Surprisingly, the correlation is lowest between the improvements and the reliability index of the raw ensembles. One reason for this is that the index does not distinguish between the direction of any errors. For example, the reliability index generated by a negatively biased prediction system would be the same as that produced by forecasts exhibiting a positive bias of the same magnitude, behaviour that was observed in the previous chapter for wind speed forecasts associated with different regimes; particularly when the regimes correspond to the same mode of synoptic-scale variation, as is the case here. More complex measures of regime-dependency could be derived to exploit this, but this is not considered here.

We now focus further on the behaviour of the hybrid approach. Figure 36 displays the CRPS for the truncated logistic EMOS and the hybrid methods, evaluated separately for days associated with each regime. The corresponding skill score of the hybrid approach relative to the truncated logistic forecast is also shown. Note first that the fluctuations between lead times indicate forecasts verifying during the night are generally more accurate than those made for midday. There is also a noticeable structure to the improvements. Although there is little difference in the CRPS between the approaches in the NAO- and 'Other' regimes, in the NAO+ the hybrid approach improves upon the conventional method by as much as 3%, even when averaged over all locations.

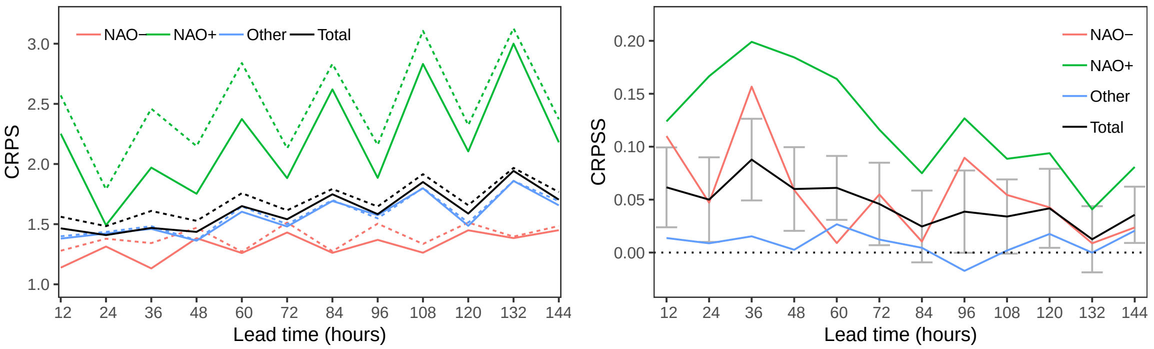

Figure 37: Same as in Figure 36 but for one location on the southwest coast of Wales.

This agrees with results in the previous section, whereby forecasts assigned to regimes that differ most from climatology are those most likely to improve. As seen in Figure 28, the NAO+ is associated with higher wind speeds, suggesting the regime-dependent approaches may help to forecast the occurrence of high-impact wind speed events.

The improvements in the NAO+ reach 3% across all locations, but the CRPSS at particular stations can be much larger than this. Figure 37 displays the CRPS and its skill score for the hybrid approach relative to the truncated logistic forecast at one location on the south-west coast of Wales. When neither the NAO- nor NAO+ regime occurs, the improvements are negligible and tend to fluctuate around zero for all lead times. During NAO+ events, on the other hand, the regime-dependent approach greatly outperforms the standard post-processing method, with improvements reaching almost 20% at a lead time of 36 hours. Large CRPSS values are also observed in the NAO- regime, though the improvement in this regime depends strongly on the lead time of interest. Overall, the improvement gained from regime-dependent post-processing at this location remains above 5% until 3 days in advance, before decreasing somewhat for longer lead times. As noted in Chapters 3 and 4, this decrease in skill is due to the mixture model weight becoming less adept at predicting the future weather regime as the forecast horizon increases.

These results suggest that the hybrid approach better captures the upper tail behaviour of the observed wind speeds. This is confirmed in Figure 38, where the skill score of the twCRPS is displayed for the two regime-dependent approaches relative to the truncated logistic EMOS forecasts. The twCRPS in Figure 38 is calculated using a threshold equal to 8m/s, roughly corresponding to the 90th percentile of the wind speed observations across all locations. The improvements now reach almost 3% when averaged over all locations and all regimes at a lead time of 36 hours, indicating that the upper tail of the forecast distribution displays a more pronounced improvement than that observed when considering the entire predictive distribution. Similar results

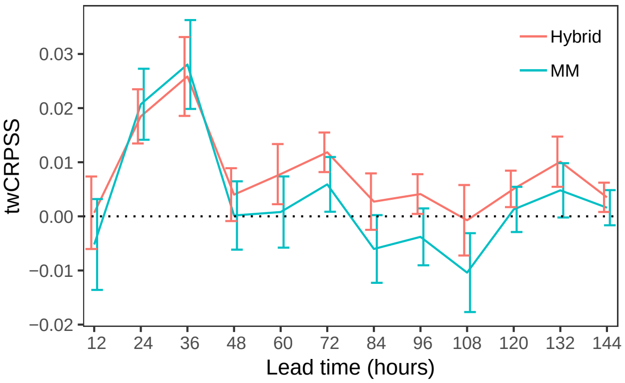

Figure 38: Threshold-weighted CRPSS for the regime-dependent mixture model forecast (MM), as well as the hybrid approach, relative to the conventional truncated logistic forecast distribution, shown against lead time.
Error bars indicate 95% confidence intervals for the skill score, obtained via non-parametric bootstrap resampling. The threshold used in the twCRPS is 8 m/s, roughly corresponding to the 90th percentile of the marginal distribution of the observed wind speeds.

are also found using thresholds of 10 and 12m/s, which are respectively close to the 95th and 98th percentiles of the distribution of observations

5.5 Discussion

This work investigates how weather types can be used to calibrate ensembles of weather forecasts, focusing in particular on how such approaches can be applied in an operational setting. A mixture model approach to include regime information into statistical post-processing methods has previously been proposed in Chapters 3 and 4, though these studies have utilised large sets of data, which are not always readily available to forecasting centres. To circumvent this lack of data, this work combines a conventional ensemble model output statistics approach with a regime-dependent method, producing a prediction system that can adapt to issue the most relevant forecast given the current circumstances.

For example, always implementing a regime-dependent mixture model is found here to outperform a truncated logistic post-processing method when averaged over all locations, though at several locations at which the regime-dependence is small, the mixture model results in a less accurate forecast. The reason for this is that the small benefits that would be obtained by regime-dependent post-processing at these locations are outweighed by the additional parameter uncertainty induced by the more complex approach. To alleviate this issue, the study presented here has looked at implementing regime-dependent post-processing models only when it is expected to be beneficial.

A measure of regime-dependency is calculated over the training data, and a regime-based approach is applied only if this measure exceeds a certain threshold. If not, then standard post-processing is implemented.

This hybrid approach offers consistent improvements over a conventional truncated logistic-based EMOS approach, with the forecast accuracy at particular locations, measured using the continuous ranked probability score, improving by over 5%. These locations tend to be stations on the west coast of the UK and Ireland, which are heavily affected by the movement of air masses across the Atlantic Ocean. This improvement becomes larger yet when considering forecasts linked to particular regimes. The regimes considered here are the opposite phases of the North Atlantic Oscillation, along with a third group corresponding to when neither of these NAO regimes occur. If the positive phase of the NAO, generally associated with wind storms and more extreme wind speeds, is predicted, then the regime-based methods yield forecasts that are 3% more accurate than those generated using the conventional EMOS approach when averaged over all locations, with improvements at individual stations reaching 20%. Similar results are also observed when the threshold-weighted CRPS is used to assess the upper tail of the forecast distributions, demonstrating that regime information can benefit predictions of high-impact wind speed events, which, despite their obvious significance, are commonly overlooked when implementing conventional post-processing methods (Williams et al., 2014; Friederichs et al., 2018; Pantillon et al., 2018). This result also has implications for multivariate post-processing approaches, since compound weather events - combinations of multiple weather hazards - are often associated with the occurrence of certain atmospheric regimes.

The NAO regimes considered in this study are utilised operationally in the Met Office's Decider tool, described in detail in Neal et al. (2016). As discussed in Section 5.2.2, this product in fact consists of eight weather regimes, which are themselves constructed by clustering together a further 30 weather patterns. However, the regime-dependent mixture model estimates a separate predictive distribution corresponding to each regime. Therefore, for the relatively small amount of data used here, post-processing with the set of eight regimes induces parameter uncertainty that outweighs the benefits of including these additional regimes. As a result, if the mixture model of Section 5.3.3 is applied using the eight regimes, then the resulting forecasts are found to perform worse than those discussed here (not shown).

Alternatively, more parsimonious approaches could be applied to add the regime information into post-processing that do not involve specifying an entirely new forecast distribution for each regime. Since the focus here is on the North Atlantic Oscillation, one example is applied (not shown) that employs a NAO index as an additional predictor in the EMOS model (along with an interaction with the ensemble mean).

Such an approach requires fewer model parameters and is thus yet more suitable for use with limited data sets, but is found to be less informative than the mixture model described herein. Instead, data-driven post-processing methods may be an effective way to model more complex relationships between the weather regimes and the existing predictors (e.g. Rasp and Lerch, 2018). For example, the influence that weather regimes assert on the biases of ensemble prediction systems may itself depend on the time of the year, and regime-dependent post-processing methods should be capable of addressing this. We anticipate that this would be more prevalent for weather variables that exhibit a particularly pronounced seasonal cycle, such as temperature, than for wind speed or precipitation, though more data-rich studies may wish to investigate this. The seasonal dependence of the weather regimes could also be incorporated into post-processing using mixture model-based approaches by utilising regimes that are more closely linked to particular seasons (Grams et al., 2017), thereby assuming that seasonality in the weather variable of interest can be attributed to the occurrence of particular atmospheric regimes (Scheuerer, 2014).

Nonetheless, in this study, it is found that wind speed forecasts benefit more from larger training data sets than from the inclusion of seasonal information. Hence, the regime-dependent approach proposed here uses weather regimes as a basis from which to select the training data for the post-processing method, in place of more contemporary approaches that rely on the season. It is demonstrated that operational weather forecasts may exhibit biases that depend on the prevailing weather regime, and hence it is necessary to investigate the presence of conditional biases prior to post-processing. If such biases are identified, as is the case for the MOGREPS-G prediction system considered here, then regime-dependent approaches are necessary to remove them. As such, work is now ongoing to integrate the regime-dependent hybrid approach proposed here into IMPROVER (Evans et al., 2020), a library of algorithms implemented by the Met Office that utilise Rose and Cylc suites (Oliver et al., 2018, 2019) to post-process and verify weather forecasts (https://github.com/metoppv/improver).

6 Accounting for skew when post-processing temperature forecasts

6.1 Introduction

The previous chapters have considered regime-dependent statistical post-processing methods, focusing in particular on applications to wind speed forecasts. This chapter, on the other hand, concerns gridded temperature forecasts over the UK, issued by the Met Office's high resolution, convection permitting MOGREPS-UK ensemble prediction system. MOGREPS-UK ensemble forecasts can be post-processed using the IMPROVER system described in Section 5.5 (https://github.com/metoppv/improver), and the work presented in this chapter therefore builds upon the existing functionality within IMPROVER.

Results in Chapters 3 - 5 indicate that the way in which regimes affect the local weather depends heavily on the location under consideration, and hence the influence of the regimes can vary drastically even within a relatively small spatial region. Therefore, although temperature is expected to exhibit a strong dependence on the prevailing weather regime, this may not be the case for temperature observations amalgamated over a large number of locations. However, in generating a calibrated forecast field, IMPROVER utilises simultaneously information regarding all spatial locations on the domain under consideration. Hence, due to the paucity of methods to address spatial biases when post-processing gridded forecasts (Dabernig et al., 2020), applying the regime-dependent statistical post-processing framework to entire temperature fields is not expected to be beneficial in this instance. Therefore, we consider in this chapter alternative methods to extend the capabilities of IMPROVER when post-processing gridded temperature forecasts.

These surface temperature forecasts are of high demand to several industries, and also to the general public. It is therefore imperative that these forecasts are accurate and reliable, something that is typically not true for forecasts (either point forecasts or ensembles) generated from operational prediction systems. An a posteriori adjustment of the forecast is therefore necessary to alleviate systematic errors, whilst simultaneously quantifying the predictive uncertainty. To achieve this, state-of-the-art statistical post-processing methods issue probabilistic forecasts in the form of predictive distributions. Although non- and semi-parametric post-processing approaches have recently received increased attention in the literature (e.g. Van Schaeybroeck and Vannitsem, 2015; Taillardat et al., 2016; Henzi et al., 2019; Bremnes, 2020), established post-processing methods usually make distributional assumptions regarding the weather variable being forecast.

Several studies have therefore considered the appropriate statistical distributions to employ when post-processing a range of weather variables. For non-negative variables such as wind speed, the distribution should have a non-negative support (Thorarinsdottir and Gneiting, 2010; Messner et al., 2014; Scheuerer and Möller, 2015), while that for precipitation should be non-negative, but also contain a positive probability of being exactly zero (Sloughter et al., 2007; Scheuerer, 2014; Scheuerer and Hamill, 2015a). For temperature, on the other hand, the normal distribution is almost invariably employed, both in a post-processing context and throughout the wider field of atmospheric science (Von Storch and Zwiers, 2001).

This is in part due to the appealing properties possessed by the normal distribution, which has led to its implementation in various branches of statistical modelling. For example, the normal distribution is widely used in linear regression (Klein et al., 1959; Glahn and Lowry, 1972; Gneiting et al., 2005), time series models (Möller and Groß, 2020), and spatial statistics (Scheuerer and König, 2014; Scheuerer and Büermann, 2014; Feldmann et al., 2015); it is conjugate to itself, and is thus convenient for Bayesian approaches (Stephenson et al., 2005; Siegert et al., 2016b; Barnes et al., 2019); while it also generalises easily to multiple dimensions, making it the canonical choice for multivariate analysis (Schuhen et al., 2012; Feldmann et al., 2015; Barnes et al., 2019).

Because the normal distribution is so widely applied in studies concerning temperature, theoretical developments in statistical post-processing models are often trialled first on temperature forecasts. The studies listed above are numerous examples of this, as are more recent approaches to ameliorate conventional post-processing methods (Messner et al., 2017; Rasp and Lerch, 2018; Schuhen et al., 2020). Nonetheless, Gebetsberger et al. (2018) have recently questioned the uninhibited use of the normal distribution as a means for post-processing temperature forecasts. In particular, the authors suggest instead that the logistic and Student's t distributions are at times more appropriate, potentially because their heavier tails can account for the additional uncertainty that arises when estimating post-processing parameters (Siegert et al., 2016a).

Moreover, the empirical distribution of temperature observations is regularly found to exhibit skew, often in particular seasons (Von Storch and Zwiers, 2001). Although post-processing is concerned with the conditional distribution of temperature given the numerical weather model output (and potentially other predictors) rather than its unconditional, or climatological distribution, it is common to post-process using a parametric distribution that resembles the climatological distribution of the response variable. In doing so, the forecast avoids assigning non-zero probabilities to weather events that cannot occur, while also capturing the limiting case where the outcome is independent of any predictors, in which instance the conditional distribution reverts to the variable's climatological distribution.

Therefore, Gebetsberger et al. (2019) propose recalibrating temperature forecasts using a Type-I generalised logistic distribution within the Non-homogeneous Regression (NR), or Ensemble Model Output Statistics (EMOS), framework (Gneiting et al., 2005). The Type-I generalised logistic distribution extends the ordinary logistic distribution by including an additional shape parameter, thereby permitting skewed predictive distributions. Alternatively, asymmetric predictive distributions could be obtained by transforming the temperature forecasts prior to post-processing, so that they conform to the assumptions made by more recognisable and convenient statistical methods, before applying the inverse transformation to the recalibrated forecasts. Hemri et al. (2015), for example, apply the well-known Box-Cox transformation to rainfall runoff before implementing Non-homogeneous Gaussian Regression. There has previously been little interest in applying such transformations to temperature.

In this chapter, we post-process short-range MOGREPS-UK temperature forecasts using both the generalised logistic approach of Gebetsberger et al. (2019), and the Non-homogeneous Gaussian Regression method after having applied a suitable transformation to the temperature forecasts. In particular, we post-process MOGREPS-UK temperature forecast fields using EMOS with a normal, logistic and Type-I generalised logistic distribution, and compare the resulting forecasts to those generated using Non-homogeneous Gaussian Regression after having first applied a non-linear transformation to the MOGREPS-UK ensemble output. It is demonstrated that relaxing the assumption of symmetry in the predictive distribution when post-processing temperature ensemble forecasts can enhance the performance of the resulting forecast fields, particularly during summer and winter and in mountainous regions.

The model and data used to illustrate this are discussed in the following section. In Section 6.3, we briefly discuss asymmetric variants of the normal and logistic distributions, as well as transformations that can be applied to address skew within samples of data. These approaches are then extended for use within the Non-homogeneous Regression framework in Section 6.4. Methods for parameter estimation and forecast verification are also discussed in Section 6.4. Section 6.5 presents the performance of forecasts post-processed using these variants of NR and comments on the choice of data to use when evaluating the performance of the gridded forecast fields. Finally, Section 6.6 concludes.

6.2 Data

This study utilises daily 2m temperature forecasts extracted from the Met Office's MOGREPS-UK ensemble forecasting system (Hagelin et al., 2017) at lead times of 12, 24 and 36 hours. The forecasts were issued in the one year period between the 1st January and the 31st December 2018, during which time the model employed a 2.2km horizontal resolution over the UK, and generated ensemble forecasts comprised of 12 members. The forecasts are initialised at 03 UTC and thus validate at 15 or 03 UTC, allowing both day- and night-time temperature predictions to be assessed.

Commonly, the aim of operational forecasting centres is to obtain a calibrated forecast field, from which predictions can be made for any location of interest. To do so, statistical post-processing methods rely on an archive of historical forecasts and observations that adequately span the spatial domain under consideration, from which to learn previous errors of the prediction system. The spatial coverage afforded by weather recordings at synoptic stations, however, is typically inadequate: recordings over seas and oceans are generally particularly sparse (Hamill, 2018). Post-processing methods that utilise only the observations provided by this irregular network of stations are therefore unsuited to address the forecast biases at all locations on the domain, meaning systematic errors remain present in the post-processed forecast field.

Instead, it would be desirable if the observations were available on a grid, similar to that of the forecasts. For this purpose, it is common to treat the analysis fields of a high resolution numerical weather prediction model as the observations when post-processing, rather than the recordings available at synoptic stations. The model analysis is the 'best guess' of the atmospheric state at a particular time given the meteorological data to hand, as identified using data assimilation (Kalnay, 2003). Although recent advances in data assimilation have made a significant contribution to the improved performance of numerical weather models (Alley et al., 2019), the analysis field is still prone to errors. Therefore, although using model analyses to train post-processing methods accounts for data scarcity, the resulting forecasts typically underestimate the uncertainty present in reality (Feldmann et al., 2019).

An approach to combine both the model analyses and the weather observations at synoptic stations might be desirable when training post-processing methods, but given the lack of such an approach, the post-processing methods discussed herein are trained using model analyses. The analyses used are from the Met Office's deterministic, convective-scale UKV model, which operates on a domain with varying resolution, comprised of an inner domain with a resolution of 1.5km, and a surrounding 4km resolution area (Tang et al., 2013). Bilinear interpolation is used to map the MOGREPS-UK ensemble forecasts onto the smoother, inner domain of the UKV grid-space prior to post-processing, and any further references to the UKV model domain relate to its inner domain. The result is a forecast grid consisting of roughly half a million grid points (810 latitudes, 621 longitudes).

The UKV model domain is displayed in Figure 39, along with the mean observed

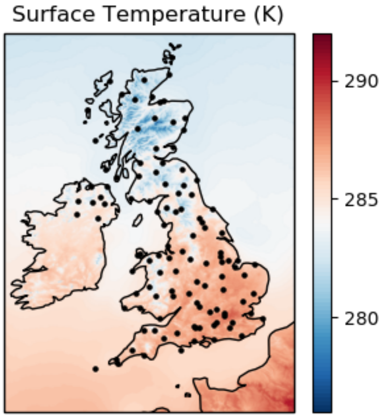

Figure 39: Average observed temperature (K) across 2018 at 15 UTC on the UKV model domain.
The black points represent the 116 weather stations at which temperatures are considered in Section 6.5.2.

temperature field estimated over 2018. The average temperature generally decreases as latitude increases, with highest temperatures in the southeast of the UK and northern France, as expected. To demonstrate that the annual climatological temperature distribution is skewed, Figure 40 displays the sample skewness of the temperature observations at each grid point, estimated over the same period. The sample skewness is defined as the Fisher-Pearson coefficient of skew, equal to

$$\frac{\sqrt{n(n-1)}}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2]^{\frac{3}{2}}}, \tag{61}$$

where the temperature observations are denoted by y, with local mean $\bar{y}$, and n represents the number of observations from which the skewness is estimated. The empirical temperature distribution across the entire year is generally negatively skewed in the northwest region of the domain, whereas inland temperatures tend to be slightly positively skewed. The ensemble member forecasts tend to capture this general behaviour well (not shown). Figure 40 also illustrates that the skew varies further in particular seasons, with the temperature more negatively skewed in winter and more positively skewed in summer. Since the negative skew in winter and the positive skew in summer are a result of the occurrence of more extreme low or high temperatures in these seasons, post-processing methods that account for skew may be better suited to capture these more extreme weather events (Williams et al., 2014).

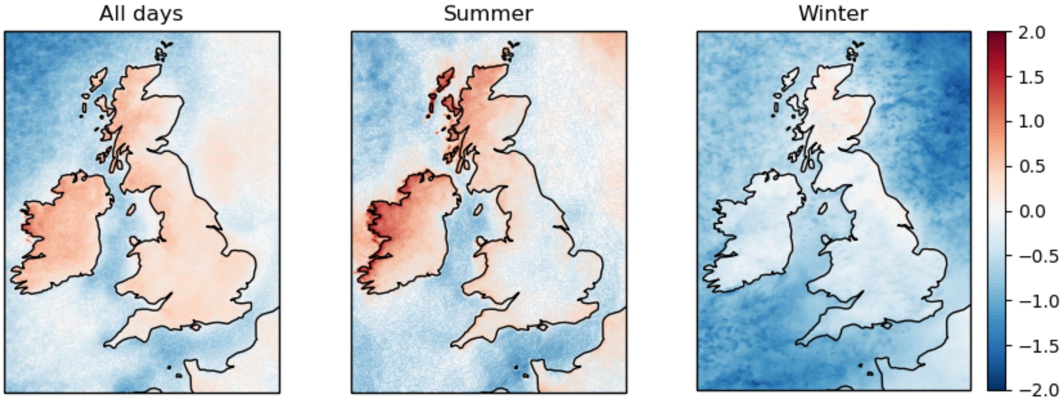

Figure 40: Average sample skewness at 15 UTC on the UKV model domain, shown for all days (left), all summer days (middle), and all winter days (right) during 2018. The sample skewness is defined as the Fisher-Pearson coefficient of skew, as given in the text.

6.3 Accounting for skew

6.3.1 Skewed distributions

Although skewed distributions are not at all uncommon, recognised extensions of symmetric distributions, such as the normal and logistic distributions, to account for possible skewness are comparatively sparse. Azzalini (1985) introduced a very general class of skewed distributions whose probability density function (PDF) is of the form

$$g(y;\lambda) = 2f(y)F(\lambda y), \tag{62}$$

where f denotes a PDF that is symmetric about zero (e.g. the normal or logistic PDF), with corresponding cumulative distribution function (CDF) F. The shape parameter λ controls the skew of the distribution and, since f is symmetric, g encompasses f when this shape parameter is zero. Although distributions of this type are theoretically appealing, the resulting distribution functions are typically complex and difficult to manipulate (Gupta and Kundu, 2010). Instead, the Type-I generalised logistic distribution (Johnson et al., 1995) provides a convenient, more accessible alternative that Gupta and Kundu (2010) argue is considerably more appropriate for practical studies. As such, Gebetsberger et al. (2019) propose employing this distribution to post-process temperature forecasts.

The probability density function of the Type-I generalised logistic distribution is

$$f_{GL}(y;\mu,\sigma,\lambda) = \frac{\lambda \exp\big(-\frac{y-\mu}{\sigma}\big)}{\sigma\big[1+\exp\big(-\frac{y-\mu}{\sigma}\big)\big]^{\lambda+1}}, \tag{63}$$

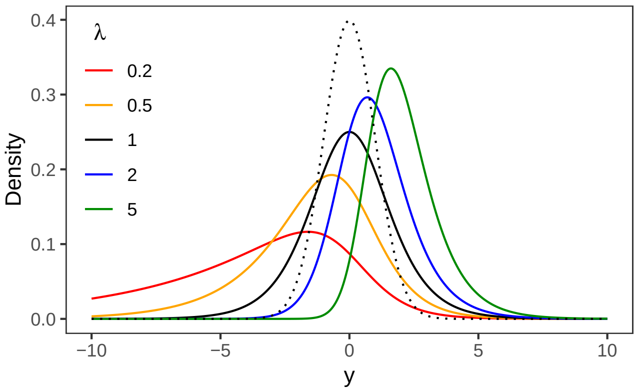

Figure 41: Examples of the probability density function of the standard (i.e. location equal to zero, scale equal to one) Type-I generalised logistic distribution for various values of the shape parameter.
The density function of the standard normal distribution is also displayed (dotted black line).

and its CDF is

$$F_{GL}(y;\mu,\sigma,\lambda) = \frac{1}{\left[1+\exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{\lambda}}. \tag{64}$$

The distribution is governed by a location parameter μ and positive scale σ and shape λ parameters. Unlike the skewed logistic distribution in the form of Equation 62, the first four central moments of this generalised logistic distribution can all be expressed in closed-form, in terms of the polygamma function (Gupta and Kundu, 2010).

However, examples of this distribution's PDF in Figure 41 suggest that it is unsuited to model heavily positive skew. Indeed, using the equation for the skew of the Type-I generalised logistic distribution presented in Gebetsberger et al. (2019) and properties of polygamma functions, it is possible to prove that the skewness of this distribution increases monotonically with the shape parameter λ, is bounded below by -2 and is bounded above by about 1.14. Yet further extensions of the logistic distribution exist - Johnson et al. (1995), for example, outlined four types of generalised logistic distributions, the simplest of which is the Type-I generalised logistic distribution characterised by Equations 63 and 64 - though, as with Equation 62, increasing the complexity of the parametric distribution makes inference increasingly difficult.

6.3.2 Transformations

Alternatively, rather than changing the distribution with which to represent temperature, transformations could be applied to temperature values so that they more readily conform to the assumptions made by particular, more desirable distributions. For example, it is often convenient to transform variables so that they appear more symmetric, allowing the implementation of more familiar statistical methods (Wilks, 2019). As such, general purpose transformations have been developed to transform data sets so that they more closely resemble a sample from a Gaussian distribution. Arguably the most well-known example of this is the Box-Cox transformation (Box and Cox, 1964).

However, the Box-Cox transformation is only suitable for non-negative quantities. Although it is possible to include an additional shift parameter in the Box-Cox transformation that ensures all data are positive, Yeo and Johnson (2000) introduced a more unified approach to transform quantities defined on the entire real line $\mathbb{R}$:

$$\psi(z;\tau) = \begin{cases} [(z+1)^{\tau} - 1]/\tau, & z \geq 0, \tau \neq 0, \\ \log(z+1), & z \geq 0, \tau = 0, \\ -[(-z+1)^{2-\tau} - 1]/(2-\tau), & z < 0, \tau \neq 2, \\ -\log(-z+1), & z < 0, \tau = 2, \end{cases} \tag{65}$$

where $\tau \in \mathbb{R}$ is a parameter that controls the shape of the resulting distribution. When τ is equal to one, we recover the identity transformation. For values of τ smaller than one, on the other hand, the upper tail of the support is contracted, while the lower tail is extended, suggesting the variable at hand is positively skewed, whereas the opposite is true when $\tau > 1$. In the following sections, we compare this transformation with the Type-I generalised logistic distribution as a means of generating skewed predictive distributions when statistically post-processing temperature forecasts. To maintain consistency with Gebetsberger et al. (2019), the Type-I generalised logistic distribution is hereafter referred to as the skew-logistic distribution.

6.4 Statistical post-processing

6.4.1 Non-homogeneous Regression

The Non-homogeneous Gaussian Regression (NGR) approach introduced by Gneiting et al. (2005) assumes that the future temperature is a random variable Y that follows a normal distribution with a mean that depends linearly on the mean of the ensemble member temperature forecasts, $\bar{x}$, and a variance that depends linearly on their variance, s^2:

$$Y|\boldsymbol{x} \sim N(\alpha_N + \beta_N \bar{x}, \gamma_N + \delta_N s^2), \tag{66}$$

where $\boldsymbol{x}$ denotes the vector of M ensemble members $(x_1, ..., x_M)$, and $(\alpha_N, \beta_N, \gamma_N, \delta_N)$ are parameters to be estimated. We discuss the nature of the parameter estimation in the following section. The two regression parameters for the distribution's location (α_N and β_N) address the biases in the ensemble mean forecast, while the two regression parameters controlling the spread of the forecast distribution (γ_N and δ_N) account for dispersion errors in the ensemble.

This approach, which falls into the broad class of distributional regression methods (Klein et al., 2015), can be extended to employ alternative parametric distributions, in which case it is more generally referred to as Non-homogeneous Regression (NR) or Ensemble Model Output Statistics (EMOS). Gebetsberger et al. (2018), for example, propose utilising a logistic distribution within this framework:

$$Y|\boldsymbol{x} \sim L\big(\alpha_L + \beta_L \bar{x}, \sqrt{\gamma_L + \delta_L s^2}\big). \tag{67}$$

As in Equation 66, the location parameter and the square of the scale parameter depend linearly on the ensemble mean and variance, respectively.

The skew-logistic distribution then extends Equation 67 through the inclusion of an additional shape parameter:

$$Y|\boldsymbol{x} \sim L\big(\alpha_S + \beta_S \bar{x}, \sqrt{\gamma_S + \delta_S s^2}, \lambda_S\big), \tag{68}$$

where $L(\mu, \sigma, \lambda)$ represents the skew-logistic (i.e. Type-I generalised logistic) distribution with location μ, scale σ, and shape λ, whereas $L(\mu, \sigma)$ denotes the ordinary logistic distribution with location μ and scale σ (and shape equal to one).

Lastly, rather than changing the distribution to be used within NR, we consider a suitable transformation of the temperature forecasts and observations prior to post-processing. Hemri et al. (2015) implement a similar approach whereby the Box-Cox transformation is applied to rainfall runoff before post-processing the transformed forecasts using Non-homogeneous Gaussian Regression. Since temperature is not constrained to be positive, we instead apply the Yeo-Johnson transformation (Equation 65) to the temperature data, but similarly implement Non-homogeneous Gaussian Regression to post-process the transformed forecasts. In particular, the approach proceeds as follows. The forecasts and observations in the training data set are first standardised by removing the mean temperature observed in the training data, and dividing by the standard deviation of these temperature observations:

$$y_j^* = \frac{y_j - \bar{y}}{\sqrt{v_y}}, \qquad x_{m,j}^* = \frac{x_{m,j} - \bar{y}}{\sqrt{v_y}} \tag{69}$$

$\forall \ m = 1, ..., M, \ \ j = 1, ..., N$, where j indexes over the N forecast-observation pairs in the training data set, and $x_{m,j}$ represents the m-th ensemble member on the j-th forecast instance. The sample mean and variance of the N observations y_j in the training data are represented by $\bar{y}$ and v_y, respectively. Although this standardisation could be performed individually for each grid point under consideration as a way to incorporate localised information (Dabernig et al., 2020), the mean and standard deviation are calculated across all grid points here to ensure a fair comparison with the alternative NR approaches considered in this study. Moreover, this standardisation may not always be necessary, but it means the resulting temperature forecasts do not depend on the original unit of measurement.

Having standardised the forecasts and observations, the shape parameter of the Yeo-Johnson transformation is estimated by finding the value $\hat{\tau}$ that maximises the (profile) likelihood of a Gaussian distribution given the transformed and standardised temperature observations in the training data set, $\psi(y_j^*; \tau)$, as described in Yeo and Johnson (2000). The standardised temperature observations y_j^* and ensemble members $x_{m,j}^*$ in the training data set are then transformed according to $\psi(\cdot\,; \hat{\tau})$, before fitting a Non-homogeneous Gaussian Regression model (Equation 66) to these transformed, standardised forecasts and observations. Using the same value of τ to transform the temperature forecasts and observations ensures they remain on the same scale.

Unlike the variations of NR described above, this approach does not assume that the future temperature follows a particular parametric distribution, but rather that the standardised and Yeo-Johnson transformed temperature is normally distributed. Therefore, in order to generate a forecast for the future, un-transformed temperature, it is necessary to sample from the predictive distribution issued by NGR for the transformed temperature, before applying the inverse of the Yeo-Johnson transformation, and finally recentring and rescaling using $\bar{y}$ and v_y.

6.4.2 Parameter estimation

The normal and logistic NR methods described above require four post-processing coefficients to be estimated, whereas the skew-logistic distribution includes also a fifth. These parameters are generally estimated by minimising a loss, or penalty function over a set of past forecasts and observations, referred to as the training data set. The training data set used in this study consists of forecasts issued during a rolling time window comprised only of the 30 days directly preceding the current forecast initialisation time. In estimating a new set of coefficients for each forecast, this time-adaptive training window can account for the behaviour of recent model errors (Gneiting et al., 2005).

To address locally-varying biases when post-processing, on the other hand, it is common to fit separate post-processing models either at every location under consideration (Thorarinsdottir and Gneiting, 2010), or for groups of locations based on proximity (Scheuerer and Hamill, 2015a), local climatological properties (Hamill et al., 2017;

Friedli et al., 2020), or local characteristics of the forecast (Lerch and Baran, 2017). Due to the extensive number of grid points considered here, elaborate methods to group together locations are computationally expensive, while site-specific post-processing is infeasible. For this reason, one post-processing model is fit to temperature forecasts across all grid points, which is the current framework implemented within IMPROVER. Such an approach has the benefit that the resulting post-processing model can be applied to forecasts at all locations, including those for which past observations are not available. Moreover, because the post-processing methods are applied to temperature forecasts aggregated over a substantially large number of grid points, there is always sufficient data from which to estimate reliable post-processing coefficients: a 30 day rolling window consists of over 15 million temperature forecast-observation pairs (30 × 810 × 621). Results presented in the following section are thus insensitive to the length of the training window (not shown).

There are two common choices for the loss function to use when estimating the coefficients over the training window. The first is the logarithmic or negative log-likelihood score, the minimisation of which is equivalent to maximum likelihood estimation. Maximum likelihood estimation is used consistently throughout statistics due to its attractive theoretical properties (Gebetsberger et al., 2018; Wilks, 2019) and it can easily be implemented for the skew-logistic forecast distribution using the probability density function given in Equation 63. However, it has become routine in post-processing studies to estimate parameters by minimising the continuous ranked probability score (CRPS), since the resulting forecast distributions tend to be sharper (Gneiting et al., 2005). The CRPS is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{u \geq y\}]^2 \, du, \tag{70}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Due to its continued use as a tool for both parameter estimation and also forecast verification, analytical solutions to this integral have been derived for several parametric distributions. Gneiting et al. (2005), for example, present a closed-form expression for the CRPS of a Gaussian predictive distribution:

$$\mathrm{CRPS}(N(\mu, \sigma^2), y) = \sigma \left\{ \frac{y-\mu}{\sigma} \left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1 \right] + 2\phi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\}, \tag{71}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the PDF and CDF, respectively, of the standard normal distribution, while Taillardat et al. (2016) and Jordan et al. (2017) derive the CRPS for the logistic distribution:

$$\text{CRPS}(L(\mu,\sigma),y) = y - \mu - \sigma - 2\sigma\text{log}F_L\left(\frac{y-\mu}{\sigma}\right), \tag{72}$$

where F_L is the CDF of the standard logistic distribution, $L(0,1)$. To evaluate the CRPS for the skew-logistic distribution, Gebetsberger et al. (2019) use numerical integration techniques. In Appendix 6, we show that the CRPS for the standard skew-logistic distribution can be expressed as the following infinite series:

$$\begin{aligned}\text{CRPS}(L(0,1,\lambda),y) = -\text{log}F_L(y) &+ \sum_{k=1}^{\infty}\frac{1}{k}[1 - F_L^k(y)] \\ &- 2\sum_{k=0}^{\infty}\frac{1}{k+\lambda}[1 - F_L^{k+\lambda}(y)] + \sum_{k=0}^{\infty}\frac{1}{k+2\lambda}.\end{aligned} \tag{73}$$

The convergence of this series is fast, but becomes slower if there exist observations in the training data that lie in the extreme upper tail of the forecast distribution, that is, observations for which $F_L(y)$ approaches the radius of convergence, 1.

It is also shown in Appendix 6 that there is an analytical representation of the CRPS for all rational values of the shape parameter (i.e. $\lambda = a/b$ with $a, b \in \mathbb{N}$). For example, if $a = 1$ and $b = 2$, then the CRPS becomes

$$\text{CRPS}(L(0,1,1/2),y) = y + 4\text{log}\frac{1 + F_L^{-1/2}(y)}{2}. \tag{74}$$

Similarly to the CRPS of the logistic distribution, we have that

$$\text{CRPS}(L(\mu,\sigma,\lambda),y) = \sigma\text{CRPS}\left(L(0,1,\lambda),\frac{y-\mu}{\sigma}\right), \tag{75}$$

and Equations 73 and 74 therefore easily extend to all possible values of the location and scale parameters.

Gebetsberger et al. (2018) argue that, since both estimators are consistent, maximum likelihood and minimum CRPS estimation should both yield calibrated forecasts if a suitable parametric distribution is employed in the statistical post-processing model. However, if invalid distributional assumptions are made by the statistical model, then training the approach by minimising the CRPS results in overly sharp forecasts, whereas the logarithmic score, in penalising poorer forecasts more heavily, encourages the post-processing method to overestimate the forecast spread. Therefore, to assess the distributional assumptions made by the different post-processing methods compared in this study, parameter estimation is performed using both minimum CRPS and maximum likelihood estimation.

In all cases, to ensure the regression coefficients for the scale of the predictive distributions are positive, the loss function is minimised with respect to $\xi = \sqrt{\gamma}$ and $\kappa = \sqrt{\delta}$ rather than γ and δ directly, and the shape parameter of the skew-logistic distribution is similarly estimated using a square-root link function to ensure positiveness. This shape parameter is estimated simultaneously to the model's other regression parameters, while the coefficients of the NGR post-processing method applied to Yeo-Johnson transformed temperatures are estimated after having obtained $\hat{\tau}$, as described previously. However, due to the extensive amount of data provided by the high resolution MOGREPS-UK forecast fields, minimum CRPS estimation for the skew-logistic distribution becomes computationally infeasible here, and results are therefore only presented using maximum likelihood estimation for this approach. Nonetheless, we demonstrate in Section 6.5.2 that minimum CRPS estimation with the skew-logistic distribution is readily applicable to other settings where smaller training data sets are in place.

6.4.3 Forecast verification

Although Non-homogeneous Regression issues forecasts in the form of predictive distributions, it is often more practical, and thus more common, to deal with a finite number of ensemble members. Therefore, after having post-processed the raw forecast using Non-homogeneous Regression, an ensemble is generated from the 12 evenly spaced $(\frac{1}{13}, \frac{2}{13}, ..., \frac{12}{13})$ quantiles of the post-processed forecast distribution. In the case of the transformed temperatures, these sampled quantiles are then subjected to the inverse Yeo-Johnson transformation to generate forecasts for the (un-transformed) temperature, as is described at the end of Section 6.4.1.

Methods to verify ensemble forecasts can then be applied. The most common tool to assess ensemble forecasts is the rank histogram, which counts the frequency with which the observed temperature value assumes each rank when pooled among the ensemble members (Thorarinsdottir and Schuhen, 2018). Deviation from uniformity in the rank histogram indicates a miscalibrated ensemble forecast, and systematic structures to this deviation can be used to diagnose the nature of any deficiencies in the prediction system (Hamill, 2001).

Furthermore, the goal of probabilistic forecasting is often stated as increasing the sharpness of the forecast, subject to calibration (Gneiting et al., 2007). The coverage of the ensemble forecast is the proportion of instances in which the observed temperature falls between the lowest and highest ensemble members - more formally, this is the coverage of the forecast's $100(M-1)/(M+1)\%$ prediction interval; in our study, with $M = 12$, this corresponds to the forecast's 85% prediction interval. If the observation is equally likely to assume any rank among the ensemble members, then this coverage should be equal to $(M-1)/(M+1)$, the proportion of ranks the observation can take on while remaining between the lowest and highest ensemble member. The range, or width, of the ensemble members then provides a measure of the forecast sharpness, and Gneiting et al. (2007) suggest that this spread should be minimised, subject to achieving the optimal coverage.

To rank and compare the competing forecast distributions, it is useful to employ an objective measure of forecast performance. For this purpose, several scoring rules have been proposed. A scoring rule maps the predictive distribution and its corresponding observation to a numerical value, thereby objectively quantifying the forecast accuracy. Such a scoring rule is said to be proper if its statistical expectation is minimised when the forecast distribution is equivalent to the distribution from which the observations arose (Gneiting and Raftery, 2007). Therefore, if a forecaster has access to the generation mechanism behind the observations, then there is no incentive for them to issue anything else as their forecast.

Both the logarithmic score and the CRPS are examples of proper scoring rules. The logarithmic score is a local score and thus relies on the predictive density function of the forecast, which is not readily available for forecasts in the form of an ensemble. Therefore, although Tödter and Ahrens (2012) propose a continuous ranked extension of the logarithmic score that is applicable to ensemble forecasts, the CRPS is more commonly implemented, since it reduces conveniently to

$$\mathrm{CRPS}(\boldsymbol{x}, y) = \frac{1}{M}\sum_{j=1}^{M}|y - x_j| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|, \tag{76}$$

for an ensemble forecast $\boldsymbol{x}$ with M members. The CRPS is negatively oriented, so that a lower CRPS indicates a more skilful forecast. The CRPS thus rewards spread among the ensemble members while penalising any deviation between the ensemble members and the observation, thereby assimilating both the reliability and sharpness of the forecast (Gneiting and Raftery, 2007). The total CRPS is then taken to be the average CRPS over all forecasts.

The continuous ranked probability skill score (CRPSS) is also applied here to assess the difference in accuracy between the various prediction schemes. This skill score is calculated as the difference between the total CRPS for a reference forecast scheme and for a competing scheme, divided by the score for the reference (e.g. Wilks, 2019). The skill score is positively oriented and bounded above by one, with values below zero indicating that the forecast under consideration performs worse than the reference to which it is being compared.

Since the aim is to obtain a calibrated forecast field over the region of interest, the forecasts are evaluated using UKV model analyses. For computational efficiency, rather than assessing the forecasts at every grid point on the UKV model domain, we consider forecasts at every 8-th latitudinal coordinate and every 6-th longitudinal coordinate on the domain. Results in the following section have thus been calculated on a grid of roughly 10,000 locations over the UK. Moreover, to better understand the qualitative behaviour of the forecasts, the performance of the forecasts relative to weather observations at 116 station locations over the UK is also illustrated, having bilinearly interpolated the forecast field to these sites. The synoptic stations considered here are displayed in Figure 39.

6.5 Results

6.5.1 Gridded forecast performance

Rank histograms for the raw ensemble forecasts and those generated using the various post-processing methods are displayed in Figure 42 at a lead time of 36 hours. The MOGREPS-UK ensemble prediction system in this case exhibits a pronounced negative bias, failing to capture the higher temperature observations; this is particularly pertinent in summer. Post-processing using Non-homogeneous Gaussian Regression trained using minimum CRPS estimation addresses this bias, though the observed temperature is found to lie outside the range of ensemble members more often than would be expected if the ensemble were calibrated, indicating the forecast is underdispersed and hence overconfident. Conversely, when the NGR approach is trained by minimising the logarithmic score, the resulting forecasts become overdispersed, reflecting the higher penalty that the logarithmic score assigns to overconfident forecasts.

A similar result is presented in Gebetsberger et al. (2018). The authors therefore propose employing a similar post-processing framework featuring distributions with heavier tails, such as the logistic distribution. However, when the CRPS is chosen as the loss function, the resulting forecasts are also found here to lack dispersion. When trained by minimising the logarithmic score, on the other hand, both the logistic and skew-logistic NR approaches appear reasonably well calibrated, whereas the NGR forecasts applied to Yeo-Johnson transformed temperatures are slightly overdispersed. Conversely, this transformation-based approach appears to rectify deficiencies in the alternative methods when trained using minimum CRPS estimation, though there appears to be some remaining structure in the histogram that indicates a heavier-tailed forecast distribution may be more appropriate even after transformation. As such, given the large number of forecasts used to construct these rank histograms, a chi-squared test for uniformity (e.g. Wilks, 2019) indicates that none of the post-processing methods produce forecasts that are perfectly probabilistically calibrated.

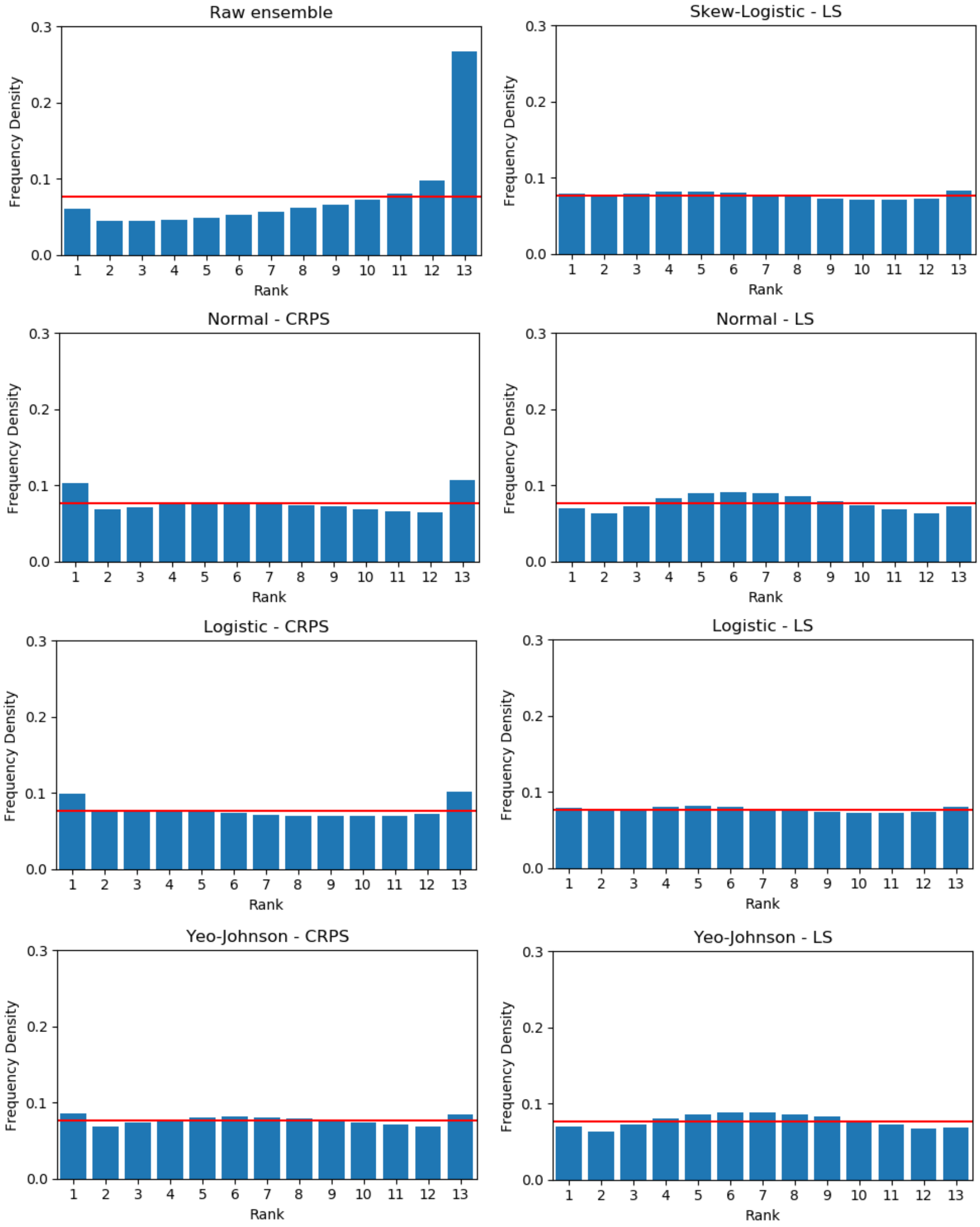

Figure 42: Rank histograms for the raw ensemble forecasts and the various post-processing methods, trained by minimising the CRPS (left column) and minimising the logarithmic score, LS (i.e. maximum likelihood; right column) at a lead time of 36 hours.
The UKV model analyses are treated as the observed values and the ranks have been aggregated over all dates and locations. The horizontal red line shown at 1/13 is indicative of perfect calibration.

The rank histograms in Figure 42 are constructed using the UKV model analyses as the observed temperature values. Figure 43 shows the analogous histograms when verifying the ensembles against temperature recordings at synoptic stations over the UK. Since the post-processing methods are trained using UKV model analyses, they are suited to address the discrepancies between the MOGREPS-UK ensemble forecasts and these analysis fields. The post-processing methods do not, however, represent the additional uncertainty that arises due to error in the analysis, and also that induced by downscaling the gridded forecasts to individual sites. As such, the rank histograms in Figure 43 suggest the ensembles, even after post-processing, are subject to considerable bias and dispersion errors, corroborating recent results in Feldmann et al. (2019). Nonetheless, the approaches based on Yeo-Johnson transformations reduce the underdispersion relative to the alternative methods.

The accuracy of the different post-processing methods can be compared more formally using the CRPSS (again calculated across all locations and dates under consideration), available in Figure 44. IMPROVER currently implements the NGR approach trained using minimum CRPS estimation, and this thus constitutes a canonical choice for the reference scheme when computing the skill scores. Results are shown for the forecasts evaluated against both the UKV model analyses and the station data. In the former case, the two post-processing methods applied to Yeo-Johnson transformed forecasts provide the largest benefit, with skill scores roughly equal to 2% at all lead times, and the improvements are yet larger when the forecasts are assessed using station data, reaching almost 5% for forecasts 12 hours in advance. Surprisingly, the approaches trained by minimising the logarithmic score marginally outperform those designed to minimise the CRPS, even though forecast accuracy is assessed here using the CRPS.

We note, however, that this is not surprising when the forecasts are evaluated relative to station observations, since Figure 42 illustrates that these approaches tend to be more overdispersed than those trained using minimum CRPS estimation. These methods thus inadvertently account for some of the additional uncertainty present when forecasting the less predictable station data, leading to an increased accuracy when assessed using observations of this type. More generally, this highlights that evaluating post-processing methods against observations that exhibit markedly different characteristics to those used to train the methods can lead to invalid inferences regarding the quality of the post-processing models. For example, even if a post-processing method were capable of determining the exact generating process underlying the analysis fields in the training data set, this method would not necessarily perform best when assessed using weather recordings at synoptic stations.

Of course, provided a suitable scoring rule is employed, the forecasts generated by the post-processing methods can reliably be compared regardless of the choice of

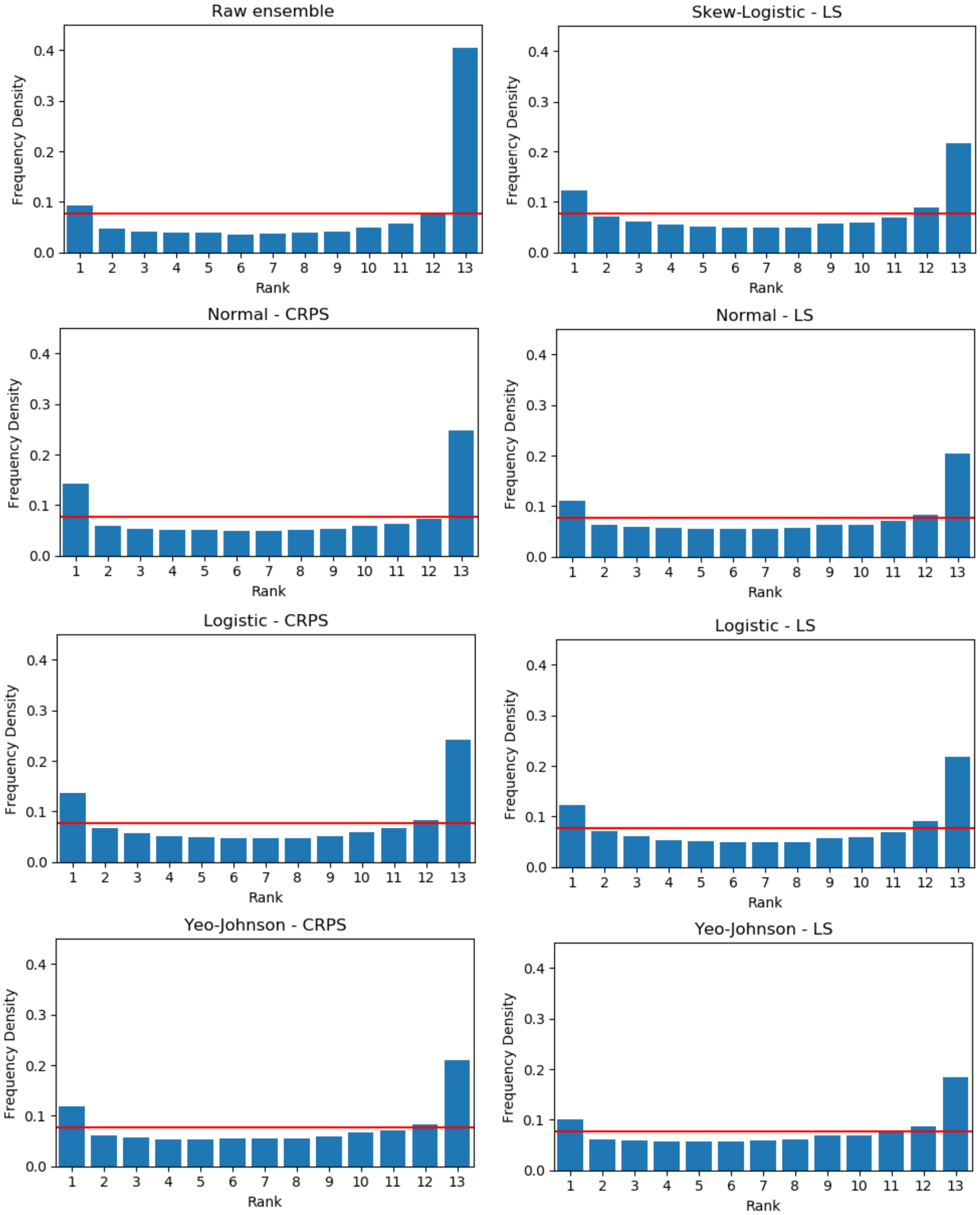

Figure 43: As in Figure 42, but with temperature recordings at synoptic stations treated as the observed values.

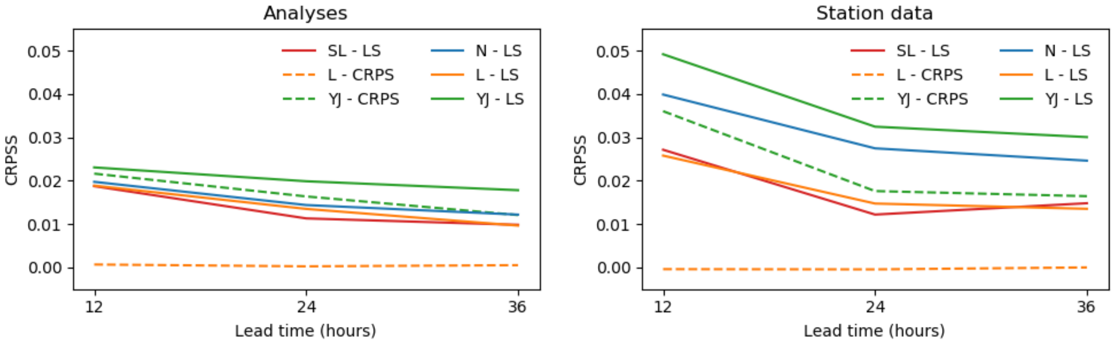

Figure 44: The CRPSS for the different post-processing methods at each lead time, relative to the NGR approach trained using minimum CRPS estimation.
The forecasts are verified against both UKV model analyses (left) and weather station observations (right). The colours distinguish between the different parametric assumptions, while the line type reflects the loss function used to train the post-processing methods. All standard errors in the first plot are negligible (see Table 14), while those in the second plot are all roughly 0.002 for all lead times. These have been omitted from the plot to aid interpretation.

observations on which the comparison is based. But if the goal of the study is to gauge the effectiveness of various post-processing methods, in terms of their ability to address the errors that they encounter in the training data set, then the different methods should be evaluated using the same type of observations as those with which they are trained. On the other hand, of course, if the principle goal is to employ a post-processing method that provides the most accurate forecasts at the synoptic weather stations, but practical constraints mean only model analyses are available to train the post-processing methods, then it might be worthwhile to choose a post-processing framework that is known to over-predict forecast uncertainty.

Since the interest here is on comparing the competing post-processing methods, all results are henceforth presented when evaluating forecast performance against the UKV model analysis fields. Table 13, for example, displays the average CRPS over all locations at a lead time of 12 hours. The mean squared error (MSE) of the ensemble mean forecast is also presented, as is the average range of the ensemble and the corresponding coverage. The CRPS for all post-processing methods improves upon that of the raw ensemble by over 20%, while the Yeo-Johnson transformed predictive distributions generate further improvements upon the other approaches. The MSEs, on the other hand, are largely indistinguishable between the different post-processing methods, suggesting the improvements are mainly due to a better representation of the shape of the predictive distribution. The coverage in this case is the proportion of instances in which the observed temperature falls within the ensemble members. Since the ensembles are each comprised of 12 members, the optimal coverage is $11/13 = 0.85$.

	CRPS	MSE	Width	Coverage
Raw ensemble	0.5096	0.7640	1.4372	0.6078
Normal - CRPS	0.3988	0.5408	1.7445	0.7780
Logistic - CRPS	0.3985	0.5408	1.7927	0.7875
Yeo-Johnson - CRPS	0.3902	0.5429	1.8821	0.8367
Normal - LS	0.3909	0.5415	2.0197	0.8654
Logistic - LS	0.3913	**0.5406**	1.8965	**0.8453**
Yeo-Johnson - LS	**0.3896**	0.5440	2.0851	0.8714
Skew-Logistic - LS	0.3913	0.5410	1.8891	0.8439

Table 13: CRPS, MSE, and the average width and coverage of 85% prediction intervals defined by the range of the ensemble members
The corresponding standard errors for the scores are available in Allen et al. (2021a). Since the ensembles comprise 12 members, an optimal coverage would be 11/13 = 0.8462. All metrics have been computed at a lead time of 12 hours using the UKV model temperature analyses as observations, and are averaged over all locations and days under consideration. The optimum CRPS, MSE and coverage among the different methods is shown in bold.

As was observed in the rank histograms, the normal and logistic distributions trained using minimum CRPS estimation are underdispersed, issuing coverages that fall below the optimal value. Estimating coefficients using maximum likelihood, or allowing the predictive distribution to exhibit skew, on the other hand, increases the spread of the ensemble members, and, in turn, produces forecasts that are well-calibrated with respect to this measure.

The extra flexibility in the skewed forecast distributions is attributable to the inclusion of an additional parameter: λ for the skew-logistic distribution and τ for the Yeo-Johnson transformation. A time series of this parameter, estimated for each forecast day in the test data set, is shown in Figure 45 for both day- and night-time predictions. The distributions of these parameters indicate that at 15 UTC, the conditional distribution of the temperature observations given the ensemble output is negatively skewed in winter ($\lambda < 1, \tau > 1$), and positively skewed in summer ($\lambda > 1, \tau < 1$). In autumn and spring, both coefficients are closer to one, suggesting the more parsimonious normal and logistic distributions are suffice during these seasons. For night-time temperature forecasts, there is less variation in the shape coefficients, and the predictive distributions appear slightly negatively skewed throughout the year. These results are reinforced by the continuous ranked probability skill score (CRPSS), which is used here to measure the improvement of the various NR approaches relative to the normal forecast distribution trained by minimising the CRPS. The CRPSS, displayed separately for each season in Table 14, indicates that improvements are largest in summer and winter, though still noticeable in autumn and spring. The large negative bias in the

	Autumn	Spring	Summer	Winter	Total
Raw ensemble	-19.24	-32.73	-42.73	-10.57	-27.81
Logistic - CRPS	-0.10	0.12	0.04	-0.01	0.07
Yeo-Johnson - CRPS	1.12	1.01	**3.29**	3.18	2.16
Normal - LS	**1.47**	**1.65**	1.94	2.92	1.97
Logistic - LS	1.39	1.15	2.38	2.67	1.89
Yeo-Johnson - LS	1.39	1.22	3.14	**3.49**	**2.31**
Skew-Logistic - LS	1.32	1.22	2.38	2.59	1.88

Table 14: CRPSS (scaled by 100) for the prediction systems relative to Non-homogeneous Gaussian Regression trained using minimum CRPS estimation, displayed separately for each season.
The skill scores have been computed at a lead time of 12 hours using the UKV model temperature analyses as observations, and are averaged over all locations and days under consideration. Standard errors corresponding to the skill scores are available in Allen et al. (2021a). The optimum skill score in each season among the different methods is shown in bold.

raw MOGREPS-UK ensemble forecasts is also apparent in Table 14, with statistical post-processing offering the most benefit in spring and summer. Analogous conclusions are drawn when verifying the forecasts against the station data (not shown).

Perhaps surprisingly, although the forecasts benefit from the increased flexibility provided by the Yeo-Johnson transformation, the skew-logistic forecast distributions perform comparatively to the logistic NR approach, both quantitatively and qualitatively. This is in part explained by the upper bound on the positive skewness of these forecast distributions, as discussed in Section 6.3, which can be seen from the fluctuating behaviour of the shape parameter during summer in Figure 45. This is further

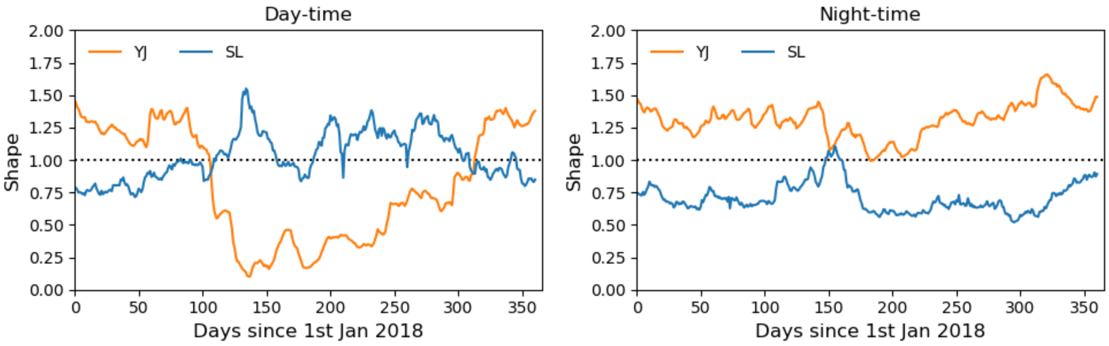

Figure 45: Time series of the shape coefficient of the Yeo-Johnson transformation, τ, and skew-logistic predictive distributions, λ as estimated over a time-adaptive 30 day training window.
The shape is displayed for both day- (15 UTC; 12 hour forecasts; left panel) and night-time (03 UTC; 24 hour forecasts; right panel) temperatures. The shapes corresponding to 36 hour forecasts are very similar to those displayed for 12 hour forecasts.

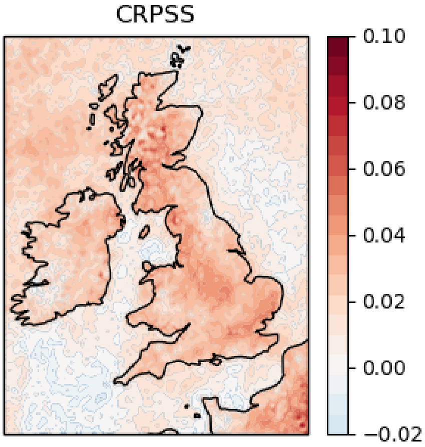

Figure 46: Map of the continuous ranked probability skill score (CRPSS) for the Yeo-Johnson transformed NGR approach, relative to the standard NGR forecasts, at each grid point.
The skill score is shown at a lead time of 12 hours and has been estimated over all of 2018. Both methods are trained using minimum CRPS estimation, and the gridded UKV model analysis fields are treated as the observations.

reinforced by the seasonal skill scores in Table 14, where the quantitative performance of the skew-logistic forecasts in summer is almost identical to that of the original logistic forecasts.

Lastly, not only does the skewness of the unconditional temperature distribution change for different seasons, but Figure 40 indicates that it varies also for different locations. Figure 46 therefore displays the CRPSS for the forecasts generated using NGR applied to Yeo-Johnson transformed temperatures, relative to those using conventional NGR, calculated separately for each grid point. Both methods are trained here by minimising the CRPS over the training data, allowing focus to be placed on the benefits gained by the more flexible distribution. The Yeo-Johnson based post-processing approach performs marginally worse than NGR at a band of locations over the North Atlantic and the North Sea, but significantly improves the performance of the resulting forecasts at inland locations across the UK. The accuracy of forecasts at individual grid points improves by as much as 10%, with the largest benefits appearing in mountainous regions in northern Scotland, which agrees with results in Gebetsberger et al. (2019). Schuhen et al. (2020) have also recently identified deficiencies in the MOGREPS-UK output and associated NGR post-processed forecasts when predicting the temperature at mountainous locations.

6.5.2 Local forecast performance

Based on Figure 46, a suitable extension of the post-processing framework implemented herein would be to calibrate grid points over land and sea separately. More generally, since the post-processing methods are trained using temperature forecasts and observations aggregated across all (over 500,000) grid points on the UKV model domain, it could be the case that the error distribution of the ensemble mean forecast becomes skewed due to the combination of (potentially symmetric) forecast error distributions across several locations. In this respect, although the previous section demonstrated that skewed predictive distributions can help to account for this behaviour, the benefits of these approaches may diminish if spatial information were incorporated into the post-processing models.

To investigate whether or not this is the case, we restrict attention to the temperature forecasts at 116 grid points over the UKV domain, which correspond to the grid points associated with the station locations displayed in Figure 39. The post-processing set up is largely similar to before: all methods considered in the previous section are also compared here, trained using both maximum likelihood and minimum CRPS estimation over a rolling 30 day window, with the UKV temperature analysis fields still treated as the observations. However, in contrast to the previous setting, a local post-processing framework is applied, whereby the coefficients of the various post-processing methods are estimated separately at each of the 116 grid points. In addition, results are also presented here for the skew-logistic NR approach trained using minimum CRPS

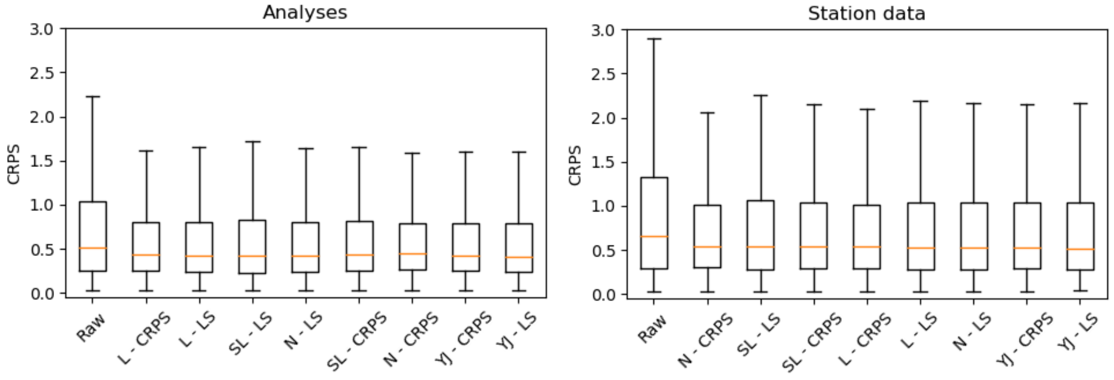

Figure 47: Boxplots of the CRPS for the raw ensemble forecast and the various local post-processing methods (as described in Section 6.5.2).
Results are shown for forecasts verified against UKV model analyses (left) and weather station observations (right) at a lead time of 36 hours. The boxes contain the median (orange line) and the lower and upper quartiles of the empirical CRPS distribution. Values of the CRPS that exceed (fall below) the upper (lower) quartile plus (minus) 1.5 times the interquartile range are defined as outliers, and have been removed from both plots. The methods have been ordered by decreasing median CRPS.

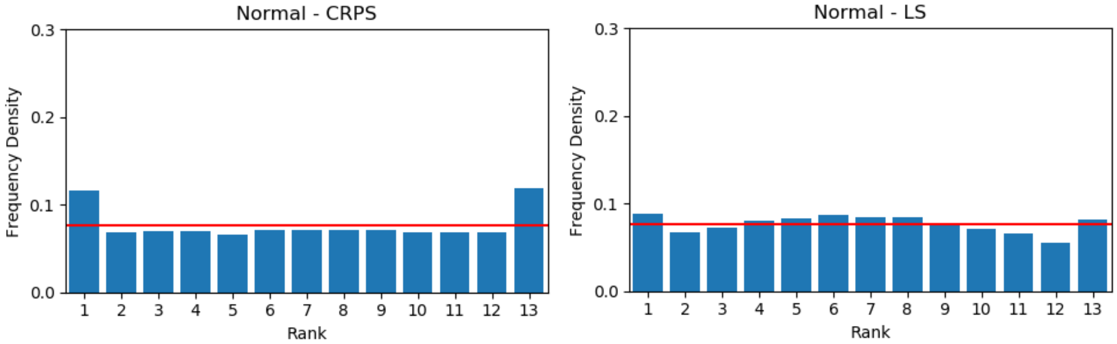

Figure 48: Rank histogram for the local NGR forecasts trained using minimum CRPS estimation (left) and maximum likelihood (right) at a lead time of 36 hours. The UKV model analyses are treated as the observed values and the ranks have been aggregated over all dates and locations. The horizontal red line shown at 1/13 is indicative of perfect calibration.

estimation, which can feasibly be implemented using the reduced amount of training data; the training data set now consists of only 30 forecast-observation pairs. Details regarding how this approach is implemented are discussed in Appendix 6.

Figure 48 displays the rank histograms for the two post-processing methods that employ a normal predictive distribution in this localised setting, at a lead time of 36 hours. A similar pattern manifests to that observed in Figure 42: the forecasts trained by minimising the CRPS are noticeably underdispersed, whereas those trained using maximum likelihood overestimate the forecast uncertainty. This is the case also for the logistic NR approaches (not shown), suggesting even when grid points are considered individually, there is still the need to make more flexible parametric assumptions when post-processing.

Boxplots of the CRPS for all post-processing methods, averaged over all locations and test days, are presented in Figure 47 for the same lead time. Firstly, we note that when applying a local post-processing approach, the discrepancy between the predictability of model analyses and station recordings is still apparent. The CRPS, which is defined on the same scale as the temperature values, is larger for forecasts assessed using the weather station data, reiterating that the post-processing methods are less adept at capturing the station-specific temperatures than they are at predicting the UKV analyses. This is also true for the raw MOGREPS-UK output. In both cases, however, all post-processing methods offer substantial improvements upon the raw, uncorrected ensemble forecast, as expected. The post-processing methods in Figure 47 have been ordered according to their median CRPS value, which is lowest for the two methods that employ a Yeo-Johnson transformation, regardless of whether UKV analyses or station recordings have been used to assess the forecasts.

	12 hours	(SE)	24 hours	(SE)	36 hours	(SE)
Raw ensemble	-42.01	(0.61)	-12.22	(0.31)	-21.72	(0.42)
Logistic - CRPS	0.30	(0.18)	0.30	(0.11)	0.01	(0.09)
Yeo-Johnson - CRPS	-0.04	(0.08)	-0.63	(0.15)	-0.80	(0.18)
Skew-Logistic - CRPS	-0.17	(0.18)	-2.31	(0.15)	-2.31	(0.14)
Normal - LS	-0.83	(0.13)	-0.79	(0.14)	-0.49	(0.11)
Logistic - LS	**0.32**	**(0.20)**	**0.47**	**(0.12)**	**0.10**	**(0.06)**
Yeo-Johnson - LS	-1.50	(0.14)	-1.00	(0.17)	-0.87	(0.18)
Skew-Logistic - LS	-1.65	(0.23)	-2.06	(0.18)	-1.78	(0.18)

Table 15: CRPSS (scaled by 100) for the prediction systems relative to Non-homogeneous Gaussian Regression trained using minimum CRPS estimation, when a localised post-processing method is implemented.
Standard errors (computed using 1000 non-parametric bootstrap resamples and scaled by 100) are displayed in brackets next to the score. The skill score is shown at all lead times using the UKV model temperature analyses as observations, and is averaged over all grid points and days under consideration. The optimum skill score among the different methods is shown in bold.

The corresponding skill scores for all methods are displayed in Table 15, with the NGR approach trained using minimum CRPS estimation again chosen as the reference scheme. Although Figure 48 suggests that the assumption of normality is invalid, the skill scores indicate that the alternative NR approaches offer little improvement upon this baseline approach. The reason for this appears to be a result of the more flexible post-processing methods becoming overfit on the reduced training data set, since they require the estimation of an additional shape parameter. As such, there are a few forecast cases in which these approaches perform particularly poorly in comparison with the reference approach, resulting in heavier tailed distributions of the CRPS values (see Figure 47). Hence, although the median CRPS value of the Yeo-Johnson based approaches is lower than that of the reference scheme, the mean is higher, leading to negative skill scores. This sensitivity to the amount of training data is particularly pertinent for the skew-logistic approach, since the shape coefficient is estimated simultaneously to the other post-processing parameters. The post-processing approach applied to Yeo-Johnson transformed temperatures, on the other hand, could more easily be adapted to account for the amount of training data by estimating the shape coefficient over an augmented data set, possibly utilising information from several locations. The remaining post-processing parameters could then be estimated locally, after obtaining a more reliable estimate for τ.

6.6 Discussion

This chapter has studied the performance of short-range temperature forecast fields over the UK, issued by the Met Office's MOGREPS-UK ensemble prediction system. The MOGREPS-UK forecasts exhibit a strong negative bias, and statistical post-processing is therefore necessary to recalibrate the numerical model output. To do so, a Non-homogeneous Regression approach is implemented here with four different choices of the parametric assumptions. Focus is particularly on the performance of skewed predictive distributions, including a variant of the logistic distribution that has recently been proposed to account for changes in the shape of empirical temperature distributions (Gebetsberger et al., 2019), as well as a novel approach that non-linearly transforms the temperature forecasts prior to post-processing, which can similarly generate asymmetric predictive distributions. Regardless of the choice of parametric family that is employed in Non-homogeneous Regression, post-processing yields forecasts that are significantly more accurate and reliable than the raw temperature ensemble forecasts, particularly in summer.

However, it is common to employ a normal distribution within the NR framework when post-processing temperature forecasts, whereas such an approach is found here to be inappropriate. In particular, the resulting forecasts are found to be either under or overdispersed, depending on the loss function used to train the forecasts, corroborating results in Gebetsberger et al. (2018). Instead, the most accurate forecasts, as measured using the continuous ranked probability score, are generated by an approach that applies Non-homogeneous Gaussian Regression after having transformed the temperature forecasts and observations in the training data set so that they appear more symmetric. This is true when using both high resolution UKV model analyses and station data to assess the resulting forecasts, though we argue that conclusions should be treated with caution in the latter case, since these observations are less predictable than the analysis fields, meaning post-processing methods that overestimate the predictive uncertainty appear more appealing. The non-linear transformation implemented here is the Yeo-Johnson transformation (Yeo and Johnson, 2000), which belongs to the more general class of power transformations frequently used in the wider field of statistical modelling (Wilks, 2019). As such, although applied here to temperature forecasts, power transformations could also easily be implemented when post-processing several alternative weather variables (Hemri et al., 2015).

In any case, as in Gebetsberger et al. (2019), the results presented herein demonstrate the potential benefit provided by more flexible parametric assumptions when post-processing temperature forecasts. We illustrate this when applying a global post-processing approach, whereby several grid points are re-calibrated simultaneously, but demonstrate that deficiencies in conventional methods exist also in more local settings. However, as when incorporating additional predictors into the post-processing model, more complex predictive distributions render the post-processing methods more dependent on the amount of available training data. Moreover, it may be the case that the more flexible parametric assumptions add information to the forecast that could alternatively be introduced via additional predictors, and future studies may wish to investigate this. Nonetheless, as long as the unconditional temperature distribution exhibits skew, we anticipate that asymmetric predictive distributions will be beneficial at longer lead times than those considered here; as the lead time increases, the observed weather variable becomes independent of the inputs to the post-processing model, meaning the conditional distribution of the weather variable reverts to its climatological, or unconditional distribution (Allen et al., 2020).

Furthermore, all forecasts in this study have been evaluated using both UKV model analyses, as well as temperature recordings at synoptic stations over the UK. Assessing the forecasts against model analyses allows the spatial characteristics of forecast performance to be better understood; as in Gebetsberger et al. (2019), the benefits of issuing skewed predictive distributions were particularly large in mountainous regions, with improvements at individual grid points reaching almost 10%. Similar improvements were also observed when verifying forecasts against temperature recordings at synoptic stations over the UK. Although there is still error in these recordings (Ferro, 2017), they generally provide a much more accurate reflection of the weather that actually materialises. However, since the post-processing methods are trained against high resolution model analyses, they are designed to correct forecast biases relative to these gridded analysis fields. As such, the post-processing methods are poorly suited to capture the additional uncertainty present when predicting the station-based temperature recordings, resulting in underdispersed forecast distributions. Rank histograms suggest that this underdispersion is less acute for the approaches that over-predict the uncertainty in the training data, and this is reflected by measures of forecast accuracy. The results presented herein therefore call for more effective approaches of combining the station data and the analysis fields when post-processing. This could be achieved, for example, by treating the post-processed predictive distributions trained using the analysis fields as prior distributions when forecasting the station data, or by suitably assimilating the two sources of information prior to fitting the post-processing model.

Finally, this study has considered forecast distributions constructed using the Nonhomogeneous Regression (NR) framework, which generally relies on specifying a unimodal predictive distribution centred around the (bias-corrected) ensemble mean, and with scale or variance that depends on the ensemble spread. One benefit of the skew-logistic forecast distribution is that the shape parameter of these forecast distributions could similarly be estimated using the ensemble sample skewness as a predictor, to incorporate flow-dependent shape information provided by the numerical model output. This was not considered here to maintain comparison with alternative approaches, though this could easily be investigated in future studies. Alternatively, it might be the case that if the ensemble members naturally reflect the skew in the temperature distributions then a post-processing approach that dresses each ensemble member individually, such as Bayesian Model Averaging (BMA; Raftery et al., 2005), might be able to utilise symmetric component distributions while also capturing this asymmetry. This reflects the additional flexibility provided by the mixture distribution used in BMA compared to NR, since it uses information independently regarding each ensemble member. In this sense, post-processing methods that make simple assumptions about the conditional distribution of the weather variable being forecast are at times inadequate. Suitably transforming the data or utilising more flexible parametric distributions (e.g. Allen et al., 2019) are potential ways of alleviating this, as might be non- or semi-parametric approaches, which have recently received increased attention in the field of post-processing (Van Schaeybroeck and Vannitsem, 2015; Taillardat et al., 2016; Henzi et al., 2019; Bremnes, 2020).

Appendix 6: Minimum CRPS estimation with the Type-I generalised logistic distribution

In this appendix we derive expressions of the continuous ranked probability score (CRPS) for forecasts in the form of Type-I generalised logistic distributions, with probability density function (PDF) and cumulative distribution function (CDF) as defined in Equations 63 and 64. Let F_λ denote the CDF of the standard skew-logistic distribution, $L(0,1,\lambda)$. The CRPS for forecasts in this form is defined as

$$\begin{aligned}\mathrm{CRPS}(L(0,1,\lambda),y) &= \int_{-\infty}^{\infty} \left[F_\lambda(u) - \mathbb{1}\{u \geq y\}\right]^2 du \\ &= \int_{-\infty}^{y} F_\lambda^2(u)\,du + \int_{y}^{\infty} \left[1 - F_\lambda(u)\right]^2 du.\end{aligned} \tag{77}$$

Without loss of generality, we restrict attention to the standard skew-logistic distribution, with $\mu = 0$ and $\sigma = 1$, since the CRPS in this case can easily be extended for other location and scale parameters using Equation 75. Note now that the CDF of the standard generalised logistic distribution, F_λ, is simply the standard logistic CDF, F_L, raised to the power of the shape parameter λ. Substituting $s = F_L(u)$ gives $s^\lambda = F_\lambda(u)$

and $ds = s(1-s)\,du$, so that Equation 77 becomes

$$\mathrm{CRPS}(L(0,1,\lambda),y) = \int_0^{F_L(y)} \frac{s^{2\lambda-1}}{1-s}\,ds + \int_{F_L(y)}^1 \frac{\left(1-s^\lambda\right)^2}{s(1-s)}\,ds. \tag{78}$$

If the shape parameter λ is a rational number, that is, $\lambda = a/b$ with $a, b \in \mathbb{N}$, then the CRPS is available in closed-form. Let $v = F_L^{1/b}(u)$ so that $v^a = F_\lambda(u)$ and $dv = [v(1-v^b)/b]\,du$. Then Equation 78 can be written as

$$\mathrm{CRPS}(L(0,1,a/b),y) = b\int_0^{F_L^{1/b}(y)} \frac{v^{2a-1}}{1-v^b}\,dv + b\int_{F_L^{1/b}(y)}^1 \frac{\left(1-v^a\right)^2}{v(1-v^b)}\,dv. \tag{79}$$

The two integrands are now rational functions and the integrals can be calculated analytically using partial fractions. For $b = 1$, we recover Equation 78 with $\lambda = a$, and for all $a \in \mathbb{N}$ we get

$$\mathrm{CRPS}(L(0,1,a),y) = y - 2\log F_L(y) + \sum_{k=1}^{a-1} \frac{1}{k}\left[1 - 2F_L^k(y)\right] - \sum_{k=a}^{2a-1} \frac{1}{k}. \tag{80}$$

For $a = 1$, this together with Equation 75 gives Equation 72 and, e.g., for $a = 2$ we get

$$\mathrm{CRPS}(L(0,1,2),y) = y + \frac{1}{6} - 2F_L(y) - 2\log F_L(y). \tag{81}$$

For $b = 2$, the expression valid for all odd $a \in \mathbb{N}$ is still rather simple:

$$\begin{aligned}\mathrm{CRPS}(L(0,1,a/2),y) = y &+ 4\log\frac{1+F_L^{-1/2}(y)}{2} \\ &+ 4\sum_{k=0}^{(a-3)/2} \frac{1}{2k+1}\left[1 - F_L^{(2k+1)/2}(y)\right] - \sum_{k=1}^{a-1}\frac{1}{k}.\end{aligned} \tag{82}$$

For $a = 3$, for example, we get

$$\mathrm{CRPS}(L(0,1,3/2),y) = y + \frac{5}{2} - 4F_L^{1/2}(y) + 4\log\frac{1+F_L^{-1/2}(y)}{2}. \tag{83}$$

Furthermore, there is an infinite series representation of the CRPS for all positive real values of λ. Going back to Equation 78 and using the power series $1/(1-s) = \sum_{k=0}^\infty s^k$ we find

$$\mathrm{CRPS}(L(0,1,\lambda),y) = \sum_{k=0}^\infty \int_0^{F_L(y)} s^{k+2\lambda-1}\,ds + \sum_{k=0}^\infty \int_{F_L(y)}^1 s^{k-1}(1-2s^\lambda+s^{2\lambda})\,ds. \tag{84}$$

Integrating the terms of these series is straightforward, and the resulting components combine to produce

$$\begin{aligned}\mathrm{CRPS}(L(0,1,\lambda),y) = -\log F_L(y) + \sum_{k=1}^{\infty}\frac{1}{k}\left[1-F_L^k(y)\right] \\ -2\sum_{k=0}^{\infty}\frac{1}{k+\lambda}\left[1-F_L^{k+\lambda}(y)\right] + \sum_{k=0}^{\infty}\frac{1}{k+2\lambda}.\end{aligned} \tag{85}$$

We remark that both the integration technique for rational λ and the infinite series technique are immediately applicable also to the CRPS in a couple of other settings. These are (i) the CRPS of a truncated skew-logistic distribution with any truncation point, (ii) the threshold-weighted CRPS (twCRPS; Gneiting and Ranjan, 2011) of the skew-logistic distribution with any threshold (when using an indicator weight function) and (iii) the twCRPS of any truncated skew-logistic distribution. This might be useful in future work when post-processing non-negative meteorological variables such as wind speed or precipitation and/or evaluating the tail of a forecast distribution. Compact expressions for the CRPS and the twCRPS of the truncated logistic distribution ($\lambda = 1$) have been used in Chapter 5 in the post-processing of ensemble wind speed forecasts.

However, it is not immediately obvious how to efficiently utilise these expressions when numerically optimising the CRPS for a skew-logistic distribution. One approach would be to employ symbolic algebra packages to evaluate the CRPS of the skew-logistic distribution analytically for a sequence of rational shape parameters, at a range of possible values of y. Interpolating this output would then provide a smooth function that approximates the CRPS at values of λ and y. Then, using Equation 75, numerical optimisation routines could be used to optimise the smooth interpolant with respect to the location, scale and shape parameters over the training data set.

Alternatively, numerical optimisation routines could use finite approximations of the infinite series in Equation 85. However, the repeated evaluation of this series is more time consuming than computing the CRPS for normal and logistic distributions in Equations 71 and 72. This is especially true when large volumes of data are considered, as is the case here, since the convergence of the series is slow when observations in the training data lie in the extreme upper tail of the forecast distribution, which is more likely to occur when considering larger archives of data. To illustrate this, Figure 49 displays the CRPS for a standard skew-logistic distribution with shape parameter equal to one half, approximated using the infinite series in Equation 85 truncated at term K. Even with $K = 500$, the series approximation is not accurate in the extreme upper tail, and a considerably larger number of terms is required to avoid this.

Increasing the number of terms in the series obviously increases the time it takes to approximate the CRPS, prohibiting its use in numerical optimisation routines. To

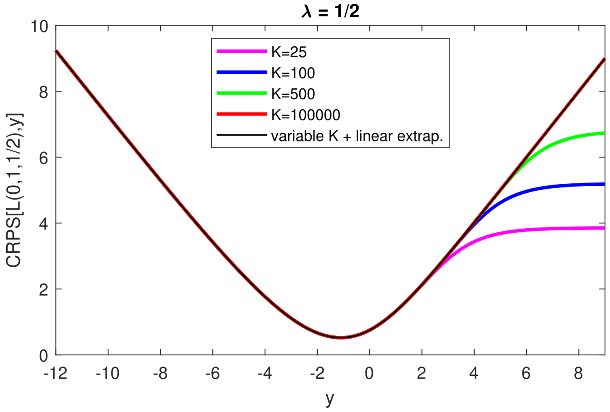

Figure 49: CRPS of a standard skew-logistic predictive distribution with shape parameter equal to one half, plotted as a function of the observation y.
The CRPS is approximated using the infinite series representation in Equation 73 truncated at K terms, for various choices of K.

circumvent this, we introduce an approximation of the infinite series in Equation 85 that uses a variable number of terms, K, depending on the value of y:

$$K = \begin{cases} 25, & y \leq 1.5, \\ 100, & y \leq 3, \\ 500, & y \leq 5. \end{cases} \tag{86}$$

If $y > 5$, then we make use of the linearity of the CRPS for large y, and approximate $\text{CRPS}(L(0,1,\lambda), y)$ using $\text{CRPS}(L(0,1,\lambda),5) + y - 5$, where $\text{CRPS}(L(0,1,\lambda),5)$ is evaluated using the series with 500 terms. Figure 49 illustrates that even though at most 500 terms are used in the series using with approach, the resulting approximation of the CRPS performs just as well as that obtained using 100,000 terms in the series without employing a linear extrapolation in the upper tail. Hence, the optimisation of the skew-logistic forecast distributions in Section 6.5.2 has been performed using this approximation to the CRPS.

7 A sequential decomposition of proper scoring rules

7.1 Introduction

Thus far, we have considered only ways of extending established methods to statistically post-process weather forecasts. Just as important as generating the forecasts themselves is the ability to evaluate their performance. The objective assessment of forecasts plays an integral role in the development of a prediction system, and numerous scoring rules have thus been devised to quantify a forecaster's ability. Scoring rules condense all information regarding forecast performance into a single numerical value, allowing an objective framework that can easily rank and compare competing schemes. However, the value of a forecast may depend on how it is to be used, and it is therefore necessary to consider several different aspects of a forecast's performance (Jolliffe and Stephenson, 2012). To fully understand the strengths and limitations of a prediction system, a variety of diagnostic tools should be employed, including graphical displays, summary statistics and numerical performance measures.

Although scoring rules provide only a single measure of forecast quality, they can often be decomposed into components that each assess a distinct aspect of the forecast. These components are typically related to the marginal and conditional distributions of the forecasts and observations, and score decompositions therefore connect scoring rules to the distributions-oriented framework for verification introduced in Murphy and Winkler (1987). A thorough history of the use of decompositions in forecast evaluation is available in Mitchell (2020).

The most commonly applied decomposition is the partitioning of a score into terms quantifying the uncertainty, resolution and reliability of the forecast. This has been studied in most detail using the Brier score (Brier, 1950; Murphy, 1973b), and is often posited as a reason for the score's popularity. The uncertainty describes the inherent variability in the forecasting scenario, while the resolution measures the extent to which this variation is captured by the forecast. The reliability, or (auto-)calibration of a forecast, on the other hand, refers to how well predictions align with their corresponding observations. Several alternative scores have been similarly decomposed, including the weighted Brier score (Young, 2010), the discrete and continuous ranked probability scores (Sanders, 1963; Murphy, 1972; Candille and Talagrand, 2005; Hersbach, 2000), the error-spread score (Christensen et al., 2015), the quantile score (Bentzien and Friederichs, 2014), and the logarithmic, or ignorance score (Weijs et al., 2010; Tödter and Ahrens, 2012). Bröcker (2009) builds on work by DeGroot and Fienberg (1983) to show that any proper scoring rule can be partitioned into terms that represent the uncertainty, resolution and reliability of the forecast, using the entropy and divergence functions associated with that score. A simple and interpretable generalisation of this result is provided by Siegert (2017).

Partitions of the score provide additional feedback to the forecaster, which can then be used to identify strengths and limitations in the prediction scheme, and, in turn, help to improve future forecasts (Murphy et al., 1989). Other partitions have therefore also been proposed that supply the forecaster with alternative information to that contained in the uncertainty, resolution and reliability terms (Yates, 1982; Mitchell, 2020), and the most suitable choice of decomposition depends on what information would be most beneficial to the forecast users. Atger (2004), for example, remarks that the reliability component provides limited information regarding forecast quality to end users, since a forecast that is reliable according to this criterion can still exhibit large conditional biases. Forecast quality could depend, for example, on the time of the year, the spatial location, or on the value of the forecast itself, and it is therefore useful to evaluate the performance of a forecast under different circumstances. If forecasters were able to identify situations in which performance is particularly poor, then they could more easily develop their forecast strategy to account for these deficiencies.

Murphy (1995) extends the distributions-oriented framework to account for the additional information provided by an auxiliary weather variable, and the aim of this chapter is to extend score decompositions in a similar manner. In the following section, we propose a novel decomposition that allows one to analyse the uncertainty, resolution and reliability terms of a proper scoring rule under different circumstances, whilst maintaining a connection to the standard definitions of the components. A similar problem has recently been considered by Ehm and Ovcharov (2017), and this is discussed in Appendix 7.1. The new decomposition allows for a more rigorous examination of the sources of information in a forecast, and thus further extends the analogy between scoring rules and the well-known analysis of variance framework. This is examined in Section 7.3. The decomposition is applied to the Brier score in Section 7.4, which is then used to assess probability of precipitation forecasts from the Met Office's convection permitting MOGREPS-UK ensemble prediction system in Section 7.5. Finally, Section 7.6 concludes.

7.2 Decompositions of proper scoring rules

7.2.1 A sequential decomposition

Suppose we are interested in assessing a probabilistic forecast F via a scoring rule S. Let $s(F, G)$ denote the expected score for the forecast under S, when the outcome Y is distributed according to G, i.e. $s(F, G) = E_{Y \sim G}[S(F, Y)]$. Bröcker (2009) shows that the forecast's expected score, for any choice of S, can be factorised into terms representing the uncertainty, resolution and reliability of the forecast. In keeping with the notation therein, the entropy of distribution G is denoted by $e(G) = s(G, G)$, and the divergence between two distributions by $d(G, H) = s(G, H) - s(H, H)$:

$$\begin{aligned} E_Y[S(F, Y)] &= e(p(y)) - E_F[d(p(y), p(y|F))] + E_F[d(F, p(y|F))], \\ &= UNC_Y - RES_F + REL_F, \end{aligned} \tag{87}$$

where $p(y)$ denotes the marginal distribution of the response variable, and $p(y|F)$ the conditional distribution of the response variable given the forecast. Throughout this work, any expectation operator will use a subscript to denote the variable(s) over which the expectation is calculated, so that E_Y signifies the expectation with respect to the distribution of Y, for example. The above decomposition holds for all scoring rules, though propriety ensures that the components are non-negative. A scoring rule is said to be proper if an optimal score is expected when the forecast issued is equivalent to the distribution from which the verifying observation is drawn. Proper scoring rules therefore encourage forecasters to issue their true beliefs, rather than hedging (DeGroot and Fienberg, 1982; Bröcker and Smith, 2007b). Scoring rules considered here are assumed to be proper and negatively oriented, so that a smaller score is preferable.

The first term of the decomposition expresses the inherent variability of the predictand, and is thus known as the uncertainty, UNC_Y. This term is independent of the forecast, and a subscript Y indicates that this is a property of the outcome variable rather than the forecasts. The uncertainty is quantified by the entropy of the marginal distribution of the outcome, and a larger uncertainty is indicative of a less predictable forecast scenario, which in turn leads to a higher (worse) score. The second component is the resolution, RES_F, which loosely measures the information content in the forecast (Bröcker, 2009). The resolution acts negatively on the score, so that more informed predictions correspond to larger resolution terms, and hence result in smaller scores. If the score attained for a perfect forecast is zero, as is often the case, then the ratio of the resolution to the uncertainty provides the proportion of uncertainty that can be explained by the forecast. Furthermore, the uncertainty and resolution are often considered together, as a measure of the sharpness or refinement in the forecast (Blattenberger and Lad, 1985; DeGroot and Fienberg, 1982; Mitchell, 2020).

The final term, REL_F, assesses the statistical consistency between the forecasts and observations. Since it acts positively on the score, this component can be thought of as the extent to which the forecast is miscalibrated, and hence a lower reliability term is desired. The component is equal to zero only if the conditional distribution of the predictand given the forecast, $p(y|F)$, is equal to the forecast itself. For example, if a forecast predicts the occurrence of an event with probability P, then the event should materialise on $100P\%$ of occasions when such a forecast is issued. The forecast in this case is said to be reliable, or calibrated. Statistical methods can often be used to calibrate predictions, and hence the resolution can be thought of as the potential information in the forecast, while the reliability represents the information that is lost via miscalibration. Although the resolution and reliability depend on both the forecasts and the outcome, a subscript F is used to highlight that these terms measure to what extent the forecasts capture the behaviour of the predictand. This notation is useful to distinguish between different decompositions of the expected score.

The factorisation of the expected score into uncertainty, resolution and reliability terms is informative, but it may be useful to investigate how these terms change under different circumstances. In doing so, forecasters can more easily identify and then rectify deficiencies in their prediction scheme. Consider a set of possible states $\{A_1, ..., A_J\}$ of the forecasting system, representing some auxiliary information on which to condition the forecast evaluation (Ehm and Ovcharov, 2017). The states are assumed to be mutually exclusive and collectively exhaustive, so that the prevailing state, A, must always manifest as one of these J possible options: $A \in \{A_1, ..., A_J\}$.

A state-dependent, or local decomposition can similarly be applied to the expected score for the forecast conditional on each state:

$$\begin{aligned} E_Y[S(F,Y)|A = A_j] =& e(p(y|A_j)) - E_F[d(p(y|A_j), p(y|F, A_j))|A_j] \\ & + E_F[d(F, p(y|F, A_j))|A_j], \\ =& UNC_{Y|j} - RES_{F|j} + REL_{F|j}. \end{aligned} \tag{88}$$

Possible choices for the auxiliary information, or states, include partitions of time (e.g. seasons or weather regime occurrences), space (e.g. spatial regions or grid points), or realisable values of a meteorological variable. Note, however, that choosing states that depend on the observations will raise the forecaster's dilemma (Lerch et al., 2017). To avoid this, the states must be chosen such that the state assigned to a forecast is identifiable prior to the forecast being issued.

The expected score is recovered from Equation 88 by taking the expectation over all possible states. This is akin to calculating the score separately for forecasts pertaining to each state, and then summing the resulting scores, each weighted by the relative frequency with which that state occurs. Likewise, uncertainty, resolution and reliability components can be calculated using a weighted sum of the terms corresponding to each state:

$$
\begin{aligned}
E_Y[S(F,Y)] &= E_A[E_Y[S(F,Y)|A = A_j]] \\
&= E_A[UNC_{Y|j}] - E_A[RES_{F|j}] + E_A[REL_{F|j}], \\
&= UNC_{Y|A} - RES_{F|A} + REL_{F|A}.
\end{aligned} \tag{89}
$$

This generates a decomposition of the expected score that contains three components representing the expected uncertainty, $UNC_{Y|A}$, resolution, $RES_{F|A}$, and reliability, $REL_{F|A}$, within the chosen states.

A similar factorisation is used, for example, to decompose the ranked probability and ranked ignorance scores, where the states correspond to the thresholds being forecast (Murphy, 1972; Candille and Talagrand, 2005; Tödter and Ahrens, 2012). However, the terms in Equation 89 are not equivalent to those in Equation 87. For example, the three terms vary depending on the choice of states, and a forecast is reliable according to this definition only if it is calibrated conditional on the occurrence of each possible state. Equation 89 therefore provides a stronger criterion for reliability than Equation 87 (Candille and Talagrand, 2005), and it follows that the conditional reliability term $REL_{F|A}$ must be at least as large as the overall reliability, REL_F.

Equations 87 and 89 thus form two distinct decompositions, each assessing slightly different characteristics. Terms that are independent of the forecast states, as in Equation 87, allow for definitions of the components that are easy to interpret, and are robust to the choice of states. Yet Equation 89 provides helpful information regarding the forecast performance in different situations that is not available from Equation 87. We therefore propose an alternative decomposition that amalgamates the two expressions, thereby possessing the benefits of both:

$$
\begin{aligned}
E_Y[S(F,Y)] = \{UNC_{Y|A} + RES_A\} &- \{RES_A + RES_{F|A} - RES_{A|F}\} \\
&+ \{REL_{F|A} - RES_{A|F}\}.
\end{aligned} \tag{90}
$$

The first, fourth and sixth components ($UNC_{Y|A}, RES_{F|A}$ and $REL_{F|A}$) are those present in Equation 89, while the first, second and third sets of curly braces are equivalent, respectively, to the classical uncertainty, resolution and reliability components in Equation 87. A proof of this is available in Appendix 7.2. This decomposition consists of five terms, rather than three, and contains two terms, RES_A and $RES_{A|F}$, that are both added and subtracted from the decomposition, and therefore have no effect on the overall score:

$$
\begin{aligned}
RES_A &= E_A[d(p(y), p(y|A))], \\
RES_{A|F} &= E_A[E_F[d(p(y|F), p(y|F, A))|A]].
\end{aligned} \tag{91}
$$

Although these extra terms contribute nothing to the score, they are themselves useful, in that they represent supplementary information regarding the behaviour of the forecasts and the response variable conditional on the possible states.

For example, RES_A is the expected divergence between the marginal distribution of the outcome and the conditional distribution of the outcome given each state. It thus describes the variation in the outcome due to changes in A, and can be thought of as the resolution of the states (note the similar form to the forecast resolution in Equation 87). The uncertainty has thus been divided into within-state ($UNC_{Y|A}$) and between-state (RES_A) contributions.

On the other hand, $RES_{A|F}$ can be thought of as the variation in the response variable that arises due to changes in the state and is not captured by the forecast. Or, equivalently, the increase in resolution that would be attained if the forecast perfectly captured the dependence of the predictand on A. Moreover, the sum of RES_A and $RES_{F|A}$ is equal to the joint information provided by both the forecast and the states. That is, $RES_{F,A} = RES_A + RES_{F|A}$, where $RES_{F,A} = E_A[E_F[d(p(y), p(y|F, A))|A]]$. The labelling of terms provides an intuitive interpretation here - the uncertainty resolved by F and A can be written as the uncertainty resolved by A, plus the additional uncertainty resolved by F after accounting for A, or vice versa. Hence, by expressing the forecast resolution as the joint information provided by the forecasts and states minus the resolution of the states given the forecast, Equation 90 allows for the sequential analysis of sources of information in the forecast.

The reliability component is now equal to the difference between $RES_{A|F}$ and the expected reliability of the forecast F with respect to A, $REL_{F|A}$. These two terms are equal, and hence the reliability is zero, when $F = p(y|F)$, which is the standard requirement for calibration. Note, however, that a reliable forecast is not necessarily calibrated with respect to the possible states. An example of this occurs when positive and negative errors cancel each other out, leading to a system that is calibrated on the whole, but is subject to conditional biases (Hamill, 2001). If the forecast is reliable with respect to the possible states, then $F = p(y|F, A)$, for all possible A, and both terms in the reliability become zero. In this case, the forecasts satisfy a stronger criterion of calibration, but the absence of conditional biases manifests in the score via the resolution term, since the negative impact of $RES_{A|F}$ is eliminated.

7.2.2 Conditioning on several states

The decomposition in Equation 90 has the desirable property that it can easily be extended further to assess the forecast simultaneously on two separate sets of states, say $\{A_1, ..., A_J\}$ and $\{B_1, ..., B_L\}$. For example, we may be interested in assessing how well the forecasts can capture variation in the predictand due to changes in both the season and weather regime. This could be achieved using separate applications of Equation 90 for each choice of state, but these states are unlikely to be independent, with some variations in weather regime likely linked to changes in the season. If the two states are considered simultaneously, then we can evaluate how well the forecasts capture the seasonality in the outcome, as well as the remaining variation due to changes in regime after having accounted for the seasonal cycle.

Another example of this is given in Ehm and Ovcharov (2017), where the score is conditioned on values of the forecast as well as some additional information - i.e. the possible values of $B \in \{B_1, ..., B_L\}$ correspond to the possible values of the forecast. However, there are other differences between Equation 90 and the decomposition presented in Ehm and Ovcharov (2017), and this is discussed further in Appendix 7.1.

For any two choices of states, the decomposition becomes

$$\begin{aligned} E_Y[S(F,Y)] =& \{UNC_{Y|A,B} + RES_{A,B}\} - \{RES_{A,B} + RES_{F|A,B} - RES_{A,B|F}\} \\ &+ \{REL_{F|A,B} - RES_{A,B|F}\}, \end{aligned} \tag{92}$$

where the terms of this decomposition are defined similarly to those in Equation 90, but with A replaced with the union of A and B. This decomposition does not separate the individual effects of the two sets of states, but rather considers their joint effect. However, as mentioned previously, the joint resolution of A and B can be decomposed further using the fact that $RES_{A,B} = RES_A + RES_{B|A} = RES_B + RES_{A|B}$. Note that if the two sets of states are independent then the joint information can be expressed as the sum of the individual resolutions. An analogous breakdown also holds for the conditional resolution, $RES_{A,B|F}$. This result can similarly be used in the presence of more than two states, though, in practice, considering more states requires further stratification of the data at hand, and hence may only be feasibly implemented when large data sets are available.

This partitioning of the joint resolution allows for the sequential analysis of the information content provided by a set of states: the uncertainty can now be written as the expected variation in the outcome for a fixed A and B, plus the variation that arises due to changes in A (or B), and that due to changes in B given a fixed A (or A given a fixed B). We can thus sequentially quantify the amount of uncertainty attributable to A and B. Similarly, in the resolution term we are sequentially quantifying the amount of uncertainty attributable to A and B that is captured by the forecasts.

7.3 Analysis of Information

The previous section discussed how the terms of Equations 90 and 92 can be used to quantify the amount of uncertainty attributable to different sources. The interpretations of these terms therefore bear resemblance to those in the well-known analysis of variance (ANOVA) framework, which seeks to quantify how much variation in the response variable can be explained by changes between different factors. A duality between decompositions of scoring rules and the analysis of variance has already been noted for quadratic scores (Blattenberger and Lad, 1985).

Consider a set of n observations and K factors, or treatment groups. The observations are written as $y_{k,j}$, where $k \in \{1, ..., K\}$ denotes the associated treatment group, and $j \in \{1, ..., n_{k\bullet}\}$ is the unit within the group of interest, with $n_{k\bullet}$ denoting the number of observations given treatment group k. The sum of the $n_{k\bullet}$ across all possible groups is equal to the total number of observations. A one-way analysis of variance then decomposes the variance in the observations into between-factor and within-factor effects. The decomposition is typically expressed as:

$$\sum_{k=1}^{K} \sum_{j=1}^{n_{k\bullet}} (y_{j,k} - \bar{y})^2 = \sum_{k=1}^{K} n_{k\bullet} (\bar{y}_{k\bullet} - \bar{y})^2 + \sum_{k=1}^{K} \sum_{j=1}^{n_{k\bullet}} (y_{j,k} - \bar{y}_{k\bullet})^2, \qquad (93)$$

where $\bar{y}_{k\bullet}$ is the mean observation given treatment k (Wilks, 2019). The first term is the sum of squared differences between the observations and their global mean, termed the total sum of squares (TSS). This term is written as the sum of the treatment sum of squares (SST), denoting the weighted deviation between the conditional and global means, and the sum of squared error (SSE), the variation of the observations from the conditional means. Dividing these terms by the total number of observations recovers the law of total variance: the variance in the observations (scaled TSS) is decomposed into the between-treatment group variation (scaled SST) and the within-group variation (scaled SSE). That is, the (scaled) treatment sum of squares, represents the amount of variation in the observations that is captured by the treatment groups, whereas the (scaled) sum of squared error is the remaining, unexplained variance. A common analysis might then test whether SST is significantly different from zero, which would suggest the treatment groups explain a non-negligible amount of variation in the data.

This decomposition does not depend on a specific prediction as such, but only the extent to which the treatment groups can distinguish between different observations. However, the sum of squared error can be decomposed further into two terms:

$$\sum_{k=1}^{K} \sum_{j=1}^{n_{k\bullet}} (y_{j,k} - \bar{y}_{k\bullet})^2 = \sum_{k=1}^{K} \sum_{j=1}^{n_{k\bullet}} (y_{j,k} - P_k)^2 - \sum_{k=1}^{K} n_{k\bullet} (P_k - \bar{y}_{k\bullet})^2, \qquad (94)$$

where P_k represents a prediction pertaining to treatment group k (Brook and Arnold, 2018). The first term on the right-hand side is the squared forecast error, while the second term is often termed the sum of squares due to lack of fit (SSLF), which represents the deviation between the forecasts and the conditional mean given the treatment group (Brook and Arnold, 2018). Note, however, that the ANOVA decomposition is typically applied in-sample, with the forecasts constructed to equal the conditional means observed in the data. Hence, the SSLF component is typically zero, and this decomposition of the SSE is rarely considered in practice. Combining and rearranging Equations 93 and 94, and scaling by the number of observations gives

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_{k\bullet}}(y_{j,k}-P_k)^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_{k\bullet}}(y_{j,k}-\bar{y})^2-\sum_{k=1}^{K}\frac{n_{k\bullet}}{n}(\bar{y}_{k\bullet}-\bar{y})^2+\sum_{k=1}^{K}\frac{n_{k\bullet}}{n}(P_k-\bar{y}_{k\bullet})^2. \quad (95)$$

Here, the mean squared error of the predictions has been decomposed into the variance of the observations, minus the expected squared deviation between the conditional and global means, plus the expected squared difference between the predicted value in treatment group k and the associated conditional mean. These terms clearly resemble the uncertainty, resolution and reliability of the forecast, respectively, as defined in Section 7.2. The (scaled) ANOVA decomposition is thus equivalent to a factorisation of the mean squared error into uncertainty, resolution and reliability terms, where the treatment groups correspond to the possible forecast values. In the case of binary outcomes, Equation 95 is equivalent to the well-known decomposition of the Brier score (Murphy, 1973b).

As noted by Blattenberger and Lad (1985), however, one key difference between the ANOVA and Brier score decompositions is the way in which they are implemented: the ANOVA framework is typically applied in-sample, whereas the Brier score decomposition is almost always used to assess out-of-sample forecast performance. Moreover, the decomposition in Blattenberger and Lad (1985) differs slightly from that discussed here. In particular, $\bar{y}$ in Equation 93 is replaced by a hypothesised conditional mean, which acts as the forecast. As a result, TSS as defined in Blattenberger and Lad (1985) is akin to the squared forecast error here, and hence their SST denotes the forecast reliability, rather than resolution. The final term of Equation 93 is also not decomposed further, and is interpreted as the refinement, or sharpness of the forecast.

In general, for other scores, the uncertainty, resolution and reliability terms are not equivalent to those in Equation 95, and hence the exact mathematical equality of the analysis of variance and score decomposition frameworks does not hold. Nevertheless, the broad interpretation of the terms remains similar regardless of the score used to assess the forecasts: the resolution represents the amount of uncertainty that can be explained by the forecast, the reliability quantifies the information that is lost due to miscalibration, and, in both cases, the aim of the forecaster is to maximise the proportion of uncertainty that is explained by the forecast. The decomposition of scoring rules therefore provides a generalised analysis of variance framework.

Furthermore, there are parallels in some common applications of the ANOVA decomposition. For example, the terms of Equation 93 are often used to calculate the coefficient of determination, or R^2 value of a prediction system. This is a well-known goodness-of-fit statistic that quantifies the proportion of variation in the observations that is explained by the forecasts, defined mathematically as the ratio of the explained variance ($SST - SSLF$[1]) to the total variance (TSS). Analogously, the proportion of uncertainty that is captured by the forecast is equivalent to the ratio of $RES_F - REL_F$ to UNC_Y, which is equivalent to the skill score obtained when using the unconditional (i.e. climatological, in the context of meteorology) forecast distribution as a reference (Mason, 2004).

The terms of the new decomposition (Equation 90) also align with the analysis of variance framework. Another common application of the ANOVA decomposition is to aid the comparison of statistical models by testing whether or not additional variables should be included to capture some of the remaining, unexplained variation in the data. In particular, a sequential ANOVA approach can be used to test whether potential covariates have a significant effect on the model fit, given those already included. The proportion of previously unexplained variation that is captured by the extra predictor is quantified by the coefficient of partial determination.

Consider now a reliable forecast. In the case of reliable predictions, the squared forecast error is equal to the sum of squared error (Equation 94). In this sense, the average score can be thought of as the amount of variation in the observations that is not explained by the forecast. As remarked in the previous section, the $RES_{A|F}$ component represents the extent to which the forecast does not accurately model the dependence between the observations and the chosen states. Therefore, the ratio of $RES_{A|F}$ to the forecast's score is the proportion of unexplained variation that would be explained if the forecast captured all uncertainty in the observations due to changes in the state. This is equivalent to the skill score for a prediction obtained by (perfectly) recalibrating the original forecast with respect to the states, relative to the original forecast itself. The terms of Equation 90 thus help to quantify the improvement in score gained from utilising a larger information set (Holzmann and Eulert, 2014; Strähl et al., 2017). More generally, the skill score of one forecast relative to another represents the proportion of previously unexplained uncertainty that has now been captured, and

[1] The $SSLF$ component is rarely included since ANOVA models are generally constructed to be unbiased.

this thus constitutes a natural analogue of the partial coefficient of determination.

In the field of weather forecasting, the decomposition presented here is therefore relevant for use alongside statistical post-processing methods, where recent advances have focused on including additional variables into recalibration models. The decomposition can be applied several times with different choices of the states, and those that result in a large $RES_{A|F}$ component may be a source of significant conditional biases in the forecast. Incorporating information regarding these states into post-processing models should in turn yield significant improvements. However, this is true only in theory, and errors may occur in practice due to a finite sample size. Commonly, the aim of post-processing methods is to calibrate the weather model output without sacrificing the information content in the numerical forecast. Conversely, adding more information seeks to increase the forecast resolution without using a model that is so complex that it overfits the data, and thus hinders calibration. This aligns with the notion of yielding predictive distributions that are sharp subject to being calibrated (Murphy and Winkler, 1987; Gneiting et al., 2007).

7.4 The Brier score

7.4.1 Decomposition of the Brier score

In this section the (half-)Brier score is decomposed according to both Equation 87 and Equation 90. The Brier, or probability, score is used to assess forecasts for the occurrence of a binary event, and is defined as the average squared difference between a probability forecast and the corresponding binary outcome (Brier, 1950):

$$BS(P, y) = (P - y)^2. \tag{96}$$

Here, P is the forecast probability of the event occurring, and y is the associated observation, which takes the value one if the event under consideration occurs, and zero otherwise. A smaller score clearly indicates closer alignment between the forecasts and observations, and a perfect forecast is one that can always recognise when the event will and will not occur, so that P is always equal to y.

Consider the case where the forecast P can only take one of a finite number of values, $P \in \{P_1, ..., P_K\}$. Murphy (1973b) shows that in such circumstances, the average Brier score over n forecast instances can be divided into three components. Let $n_{k\bullet}$ represent the number of times forecast P_k was issued, with $\sum_{k=1}^{K} n_{k\bullet} = n$, and let $o_{k\bullet}$ denote the frequency with which the event occurred given that the k-th forecast was issued. The average event frequency given the k-th forecast is then defined as the ratio of $o_{k\bullet}$ to

$n_{k\bullet}$, denoted by $\bar{y}_{k\bullet}$, while the climatological event frequency is

$$\bar{y} = \frac{\sum_{k=1}^{K} o_{k\bullet}}{\sum_{k=1}^{K} n_{k\bullet}} = \frac{\sum_{i=1}^{n} y^{(i)}}{n}.$$

The classical decomposition (Murphy, 1973b) into uncertainty, resolution and reliability components is then

$$\frac{1}{n}\sum_{i=1}^{n} BS(P^{(i)}, y^{(i)}) = \bar{y}(1-\bar{y}) - \sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(\bar{y}_{k\bullet} - \bar{y})^2 + \sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(P_k - \bar{y}_{k\bullet})^2, \tag{97}$$

where $P^{(i)}$ and $y^{(i)}$ represent, respectively, the predicted probability and observation on the i-th forecast case. The first term is an estimator for the uncertainty component $\widehat{UNC}_Y$, the second estimates the resolution $\widehat{RES}_F$, and the final term the reliability $\widehat{REL}_F$. This is equivalent to the ANOVA decomposition in Equation 95.

Suppose now that, as before, we wish to assess the forecast conditional on the J possible states that could arise. Let $n_{\bullet j}$ denote the number of times state j occurred, and n_{kj} the number of times forecast k was issued given that state j occurred. Define also $\bar{y}_{\bullet j}$ as the climatological event frequency given state j, and $\bar{y}_{kj}$ as that given the j-th state and the k-th forecast value. The terms of Equation 90 for the Brier score can then be expressed as

$$\begin{aligned}
\widehat{UNC}_{Y|A} &= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n} \bar{y}_{\bullet j}(1 - \bar{y}_{\bullet j}),\\
\widehat{RES}_{A} &= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n} (\bar{y}_{\bullet j} - \bar{y})^2,\\
\widehat{RES}_{F|A} &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{\bullet j} - \bar{y}_{kj})^2,\\
\widehat{RES}_{A|F} &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y}_{kj})^2,\\
\widehat{REL}_{F|A} &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (P_k - \bar{y}_{kj})^2,
\end{aligned} \tag{98}$$

with $\sum_{j,k}^{J,K}$ denoting the double summation over all possible forecasts and states. A hat is again used to emphasise that these terms are estimators in the presence of a finite amount of data. We show in Appendix 7.3 that when these terms are combined as in Equation 90, then we recover the terms in Equation 97.

7.4.2 Bias corrected components

In practice, the score and its constituent terms can only be estimated from a finite amount of data. The sample mean score is an unbiased estimator for the expected score, but, in general, the empirical uncertainty, resolution and reliability terms are biased (Bröcker, 2012a). Although no unbiased estimators exist for the resolution and reliability components of the Brier score (Ferro and Fricker, 2012), both Bröcker (2012a) and Ferro and Fricker (2012) have proposed bias corrections whose biases decay at a much faster rate than those of Equation 97. In the following section, the bias corrected decomposition of Ferro and Fricker (2012) is utilised:

$$\begin{aligned}
\widetilde{UNC}_Y &= \widehat{UNC}_Y + \frac{\bar{y}(1-\bar{y})}{n-1}, \\
\widetilde{RES}_F &= \widehat{RES}_F + \frac{\bar{y}(1-\bar{y})}{n-1} - \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}), \\
\widetilde{REL}_F &= \widehat{REL}_F - \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}).
\end{aligned} \tag{99}$$

We use similar ideas to obtain bias corrections for the terms in Equation 98. In doing so, the resulting estimates are less sensitive to the amount of data available:

$$\begin{aligned}
\widetilde{UNC}_{Y|A} &= \widehat{UNC}_{Y|A} + \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}), \\
\widetilde{RES}_A &= \widehat{RES}_A + \frac{\bar{y}(1-\bar{y})}{n-1} - \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}), \\
\widetilde{RES}_{F|A} &= \widehat{RES}_{F|A} + \frac{1}{n}\sum_{j=1}^{J} \frac{n_{\bullet j}}{n_{\bullet j}-1}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}) - \frac{1}{n}\sum_{j,k}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}), \\
\widetilde{RES}_{A|F} &= \widehat{RES}_{A|F} - \frac{1}{n}\sum_{j,k}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}) + \frac{1}{n}\sum_{k=1}^{K} \frac{n_{k\bullet}}{n_{k\bullet}-1}\bar{y}_{k\bullet}(1-\bar{y}_{k\bullet}), \\
\widetilde{REL}_{F|A} &= \widehat{REL}_{F|A} - \frac{1}{n}\sum_{j,k}^{J,K} \frac{n_{kj}}{n_{kj}-1}\bar{y}_{kj}(1-\bar{y}_{kj}).
\end{aligned} \tag{100}$$

It is assumed here that the forecast-observation pairs are independent and identically distributed, and also that the forecasts and states are chosen such that none of $n_{k\bullet}$, $n_{\bullet j}$, and n_{kj} are equal to one. It is straightforward to verify that the bias corrections above cancel each other out so that the estimator for the expected score is unbiased, and also that the bias corrections for the corresponding UNC_Y, RES_F and REL_F estimators agree with those in Equation 99. A more thorough investigation of the properties of these bias corrections is conducted in Appendix 7.4. In the next section, we use these bias corrected components of the Brier score to assess probability of precipitation forecasts.

7.5 Case study

7.5.1 Data

This section demonstrates how the Brier score can be applied to assess forecasts under a variety of different circumstances, highlighting the merits of the novel decomposition discussed in Section 7.2. Several competing probability of precipitation forecasts are assessed using both the classical and novel decompositions. Forecast performance is evaluated conditional on three different choices of states: the season at the forecast initialisation time, the predicted weather regime at the forecast validation time, and the spatial region to which an observing station belongs.

Ensemble forecasts for 24 hour precipitation accumulation are obtained from the Met Office's MOGREPS-UK ensemble prediction system. The model operates at a 2.2km resolution over the UK and surrounding area, and the forecasts are bilinearly interpolated to 140 synoptic stations across the UK and Ireland. The locations considered are shown in Figure 50. Precipitation measurements at these sites are used to verify the forecasts. In particular, interest lies on the precipitation accumulation at each station between 24 and 48 hours after the forecast is initialised.

Since the accumulations must be non-negative, interpolation schemes may inflate forecasts of zero precipitation (Accadia et al., 2003). To counteract this, a nonzero precipitation event is said to occur only if the 24 hour accumulation exceeds 0.3mm. Forecasts are considered daily, initialised at 0300 UTC between the 1st January 2017 and the 31st December 2018. The two year period is chosen since it lies between two major model upgrades, ensuring fairly homogeneous forecast errors. After accounting for missing data, the total number of forecast-observation pairs available is roughly 100,000.

7.5.2 States

The first set of states considered are the seasons. Forecasts initialised between December and February are classified as Winter, March to May as Spring, June to August as Summer, and Autumn is those in September, October or November. There are hence four states that occur with very similar frequencies.

Alternatively, time can be stratified by the occurrence of synoptic-scale weather regimes. Neal et al. (2016) identify 30 weather patterns over the UK from mean sea

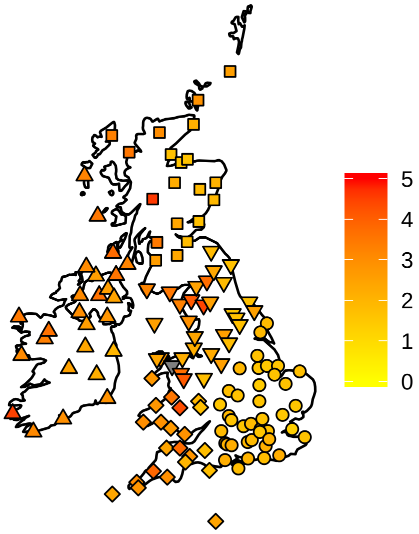

Figure 50: The 140 station locations where precipitation forecasts are considered. The five different shapes are used to distinguish between stations assigned to different regions, while the colours represent the average 24 hour precipitation accumulations at each site (in mm). One location in northwest Wales has a mean precipitation accumulation of 7.3, which is shown in grey to allow comparison between the remaining stations.

level pressure anomaly fields, which are later condensed to eight more general regimes. These regimes are used operationally in the Met Office's Decider tool, and they are used here as states to investigate how forecast quality depends on the prevailing atmospheric behaviour. A plot of the regime centres is shown in Figure 27 in Chapter 5, and a thorough interpretation of the regimes is available in Neal et al. (2016). Regimes 7 and 8 occur only a handful of times over our verification set and hence are merged with other regimes, those to which they are most similar. Occurrences of Regime 7 are instead classified as Regime 5, and those of Regime 8 are mapped to Regime 6. This results in six weather regimes that all occur with similar frequencies.

Studies of model performance in different weather regimes typically define the regimes as those that materialise at the forecast initialisation time. Results in Chapter 4, however, suggest that model biases vary more depending on the regime predicted by a numerical weather model at the forecast validation time. Therefore, we assign forecast-observation pairs to a weather regime depending on the regime that is predicted to occur by the Met Office's global deterministic model at 1200 UTC on the day of validation. This is equivalent to how the regime of a forecast is defined in Chapter 5.

Figure 51 displays the distribution of 24 hour precipitation accumulations in each season and weather regime at all locations under consideration. There is a clear de-

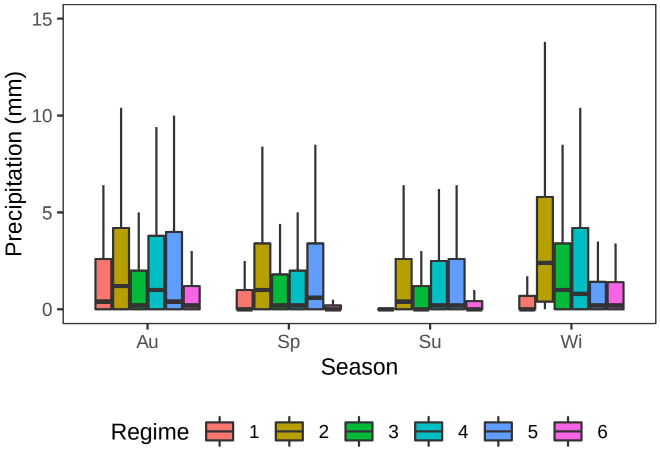

Figure 51: Distribution of 24 hour precipitation accumulations in each season and weather regime, across all locations.
Outliers, defined as precipitation values more than 1.5 times the interquartile range above the upper quartile, have been omitted for ease of comparison.

pendence on season, with large accumulations occurring in winter and autumn, and comparatively less precipitation in spring and summer. There is also a reliance on the prevailing weather regime, which may itself depend on the season. Regime 2, corresponding to the positive phase of the North Atlantic Oscillation (NAO; see Chapter 5) is associated with relatively high precipitation accumulations throughout the year, whereas the negative phase of the NAO (Regime 1) is linked to drier periods in the UK.

Lastly, we examine how forecasts perform over distinct spatial regions across the UK and Ireland. The locations are stratified into five clusters using a k-means clustering approach applied to the latitude-longitude coordinates of the stations, minimising the average distance between points in the same cluster (Wilks, 2019). These clusters then act as the states. The grouping is shown in Figure 50, with the five clusters roughly corresponding to the south-east of England, the south-west of the UK, northern England and Wales, Scotland, and Ireland and Northern Ireland.

Of course, there are several alternative ways to partition the stations, perhaps based on geographical properties (Hamill et al., 2017), local orography (Friedli et al., 2020), or historical model biases (Lerch and Baran, 2017). For this data set, however, results suggest that the conclusions drawn from these more advanced methods would be similar to those obtained from the simple approach implemented here. The number of clusters is also chosen arbitrarily, though this is again not found to have a large effect on the results.

7.5.3 Forecasts

For the time period considered here, the MOGREPS-UK model issues ensemble forecasts comprised of $M = 12$ members. A probability of precipitation forecast is then extracted from the ensemble using the following formula:

$$P = \frac{(M^{+} + 1) - 1/3}{(M + 1) + 1/3}, \tag{101}$$

where M^{+} denotes the number of ensemble members that predict non-zero precipitation, and the adjustments of $\pm 1/3$ ensure the forecast probability is never 0 or 1 (Wilks, 2006; Williams et al., 2014). The resulting probabilities range from 2/40 to 38/40, and intermediate forecast values differ by a probability of 3/40, resulting in 13 $(M + 1)$ evenly spaced possible forecasts.

If the range of possible forecast values were continuous, it may be necessary to group together similar forecasts in order to estimate the conditional event frequencies. Stephenson et al. (2008) show that such a binning can introduce biases, and two additional terms should be included in the Brier score decomposition (Equation 97) to address this. Since the number of possible forecast values here is discrete, the Brier score components can easily be calculated without any grouping of the forecasts.

7.5.4 Results

One of the most simple forecasts to issue is the unconditional distribution, or the historical, long-run average event frequency. Such a forecast is known to be reliable, but also uninformative, and it is hence often considered as a baseline to which other forecast schemes can be compared (Mason, 2004). In this case, the unconditional, or climatological forecast probability of precipitation is obtained on each day via leave-one-out cross-validation, whereby it is estimated by the proportion of precipitation event occurrences over all other days in the two-year verification period. Data from all locations is amalgamated, and the climatological forecast therefore does not react to changes in the location, season or regime. The forecast is then mapped to the permissible forecast to which it is closest. In all cases, this forecast is equal to 0.5, suggesting precipitation occurs on roughly half of the days under consideration. The Brier score for this climatological forecast is decomposed in Table 16. Firstly, note that the terms of the classical decomposition, as well as the overall score, are independent of the choice of states, with both the forecast resolution and reliability almost equal to zero, as would be expected from a climatological forecast. The ratio of RES_A to UNC_Y then provides the proportion of variation in the observations that is attributable to changes in the state. This term is larger for the regime states than for the seasons

State	UNC$_Y$	RES$_F$	REL$_F$	UNC$_{Y\|A}$	RES$_A$	RES$_{F\|A}$	RES$_{A\|F}$	REL$_{F\|A}$
Seasons	2499.8	0	0.2	2454.9	44.9	0	44.9	45.1
Regimes	2499.8	0	0.2	2368.7	131.1	0	131.1	131.3
Locations	2499.8	0	0.2	2439.0	60.8	0	60.8	61.0

Table 16: Brier score for a climatological forecast decomposed using three choices of states.
Terms have been scaled by 10^4 to aid interpretation. The total Brier score is equal to 2500 (±0.0).

State	UNC$_Y$	RES$_F$	REL$_F$	UNC$_{Y\|A}$	RES$_A$	RES$_{F\|A}$	RES$_{A\|F}$	REL$_{F\|A}$
Regimes	2499.8	130.7	1.3	2368.7	131.1	0	0.3	1.7

Table 17: Brier score for a climatological forecast that has been conditioned on the predicted weather regime, decomposed using the regimes as states.
Terms have been scaled by 10^4 to aid interpretation. The total Brier score is equal to 2370.4 (±3.3).

and locations, with changes in the regime accounting for just over 5% of the variation in the observed precipitation distribution.

Secondly, we notice that the RES_A and $RES_{A|F}$ components are equal. The difference between these terms can be interpreted as the amount of state-dependent variation that is captured by the forecast. As expected from a climatological forecast, this difference is zero, since the forecast is constructed to not depend on the prevailing state. The $RES_{A|F}$ component is also very close to $REL_{F|A}$, resulting in a reliability term that has a negligible impact on the score. However, both of these terms are positive, which suggests that, although the forecasts are reliable, they are not reliable with respect to the chosen states. For a climatological forecast, this miscalibration is, in theory, equivalent to the total variation in the observations due to changes in the state, which is largest for the regimes in this data set.

Just as the climatological forecast is reliable overall, a forecast that is calibrated with respect to the states could be constructed by deriving a separate climatology for each state, and issuing that which corresponds to the prevailing state. Results in Table 17 show the decomposition that would be obtained from a regime-dependent climatological forecast. In this case, the $RES_{A|F}$ and $REL_{F|A}$ components are again almost the same, suggesting a reliable forecast, but they are now very close to zero as well. This highlights how the forecasts are calibrated with respect to the regimes, and the resulting forecasts exhibit an increased resolution, and hence better score, as a result.

However, using information from the states alone means the forecasts now explain at most 5% of the variation in the data, and hence are still not particularly useful.

State	UNC_Y	RES_F	REL_F	$\mathrm{UNC}_{Y\|A}$	RES_A	$\mathrm{RES}_{F\|A}$	$\mathrm{RES}_{A\|F}$	$\mathrm{REL}_{F\|A}$
Seasons	2499.8	1393.7	1.8	2454.9	44.9	1352.0	3.2	4.9
Regimes	2499.8	1393.7	1.8	2368.7	131.1	1264.1	1.4	3.2
Locations	2499.8	1393.7	1.8	2439.0	60.8	1336.0	3.1	4.8

Table 18: Brier score for a Met Office MOGREPS-UK ensemble forecast decomposed using three choices of states.
Terms have been scaled by 10^4 to aid interpretation. The total Brier score is equal to 1107.9 (±6.6).

Numerical weather prediction models, on the other hand, aim to reproduce the physical processes governing the atmosphere's evolution, and thus should contain considerably more information than a purely statistical forecast, particularly in the short- and medium-range. Indeed, this is highlighted in Table 18, which shows the decomposed score for this ensemble forecast: as expected, the resolution term is considerably larger than for the climatological forecasts.

Table 19 shows the Brier scores for the forecasts discussed above, as well as the Brier skill score relative to climatology. Incorporating regime information into the climatological probability of precipitation improves forecast performance (as assessed using the Brier score) by roughly 5%, whereas the numerically driven forecast explains 55% of the variation in the observations, resulting in a considerably lower score. The increase in information from the numerical forecast often comes at the expense of reliability, though in this example the forecasts are also well-calibrated. This is reinforced by the reliability diagram in Figure 52, displaying the forecast probability plotted against the conditional event frequency (Bröcker and Smith, 2007a). These quantities should coincide if the forecast is reliable, and hence points that lie along the diagonal, as is seen here, are synonymous with a well-calibrated prediction.

As has been discussed extensively throughout the previous chapters of this thesis, a recalibrated forecast could be obtained by issuing the conditional distribution of the outcome given the forecast. Standard recalibration methods do not utilise any additional information, other than that already present in the forecast, and hence they are only beneficial when the initial prediction is highly unreliable. This suggests a recalibration of the forecast here would be redundant. However, although the forecast is reliable, it may still exhibit conditional biases, which may partially account for some of the unexplained variation in the observations. Table 18 again reveals that $REL_{F|A}$ and $RES_{A|F}$ are both very close to zero for all choices of states, and hence the MOGREPS-UK forecasts already explain most of the uncertainty owing to the seasons, regimes and spatial regions. Recalibrating the forecasts conditionally on these states would therefore also provide little benefit. A yet more detailed description of forecast performance could be obtained by looking at the decomposition components specific to each state (i.e.

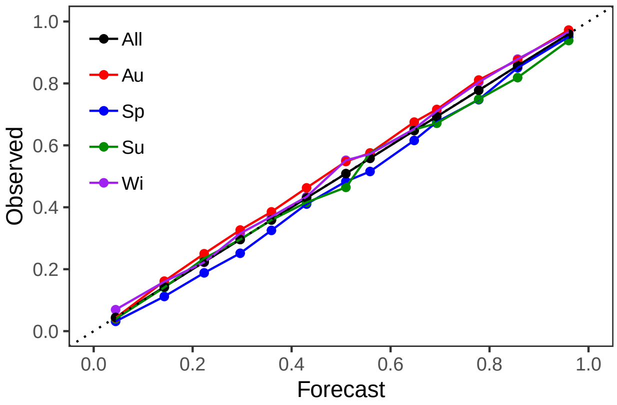

Figure 52: Reliability diagram for the MOGREPS-UK ensemble forecast for the probability of precipitation.
A dotted line is drawn along the line of equality to indicate a perfectly reliable forecast. The forecast and observed event frequencies are displayed separately for each season: autumn (Au), spring (Sp), summer (Su) and winter (Wi), and also for all days (All).

$UNC_{Y|j}$, $RES_{F|j}$, $REL_{F|j}$ in Equation 88). However, since the forecasts are almost perfectly reliable conditional on all choices of the states, the ensemble forecast in all cases explains between 50 and 60% of the uncertainty in the observations (not shown).

7.6 Discussion

This chapter has studied decompositions of proper scoring rules into uncertainty, resolution and reliability components. We distinguish between two alternative partitions, and propose an extension that utilises information from both. The decomposition introduced here maintains the established interpretation of the components, while also allowing the forecast quality to be assessed in different situations. The motivation behind such an extension is that it provides additional information to forecasters, helping them to identify when and why forecast performance changes. In particular, the decomposition divides both the uncertainty and the resolution into within-state and between-state contributions, while also allowing for the simultaneous evaluation of the

Forecast Scheme	BS	BSS
Climatology	2500.0	0.0
Regime-dependent Climatology	2370.4	5.2
MOGREPS-UK Ensemble	1107.9	55.7

Table 19: Brier score (BS) and Brier skill score (BSS) relative to climatology, for the climatological forecast, the regime-dependent climatological forecast, and the MOGREPS-UK Ensemble.
BS has been scaled by 10^4 to aid interpretation, and BSS by 10^2 so it can be interpreted as a percentage improvement.

overall and the conditional calibration. It thus becomes easy to identify states on which the outcome depends strongly, yet the forecasts do not. Such information can then easily be incorporated into the forecast, potentially via statistical post-processing methods.

The extended decomposition therefore also allows the user to quantify the contributions from several sources of uncertainty in the observations. This aligns with the idea that score decompositions can be thought of as a generalised analysis of variance framework, with further analogies related to the coefficients of determination and partial determination. More precisely, Equation 90 allows for a sequential breakdown of the forecast information, which means several sets of states can be considered in the decomposition simultaneously, as is discussed in Section 7.2.2.

An example of the decomposition is provided for the Brier score, including formulae for appropriate bias corrections when estimated over a finite sample size. This is then applied to 24-48 hour probability of precipitation forecasts from the Met Office's MOGREPS-UK ensemble prediction system, at several synoptic stations over the UK and Ireland. The decomposition is applied using three choices of the state, including both temporal and spatial information. At this short lead time, the forecast is found to be almost perfectly reliable, so that a recalibration scheme would not provide substantial benefit to the prediction. The forecast explains 55% of the uncertainty in the observations, and also accurately captures the seasonality and spatial structure of the data, and hence only minimal additional information would be provided by including these factors into post-processing models. Similarly, the ensemble prediction system represents well the dependence of the precipitation on a set of local weather regimes, meaning the regime-dependent post-processing methods presented in Chapters 3 - 5 would not be beneficial in this instance.

Future research in this area could consider the decomposition of multivariate scores. The partition presented here is able to separate contributions to the score components from different sources, and it would be interesting to study whether this could be used to extract the marginal contributions to a score used to assess several variables simultaneously. This would also help to alleviate well-documented shortcomings in decompositions of the discrete and continuous ranked probability scores (Ferro and Fricker, 2012). Finally, to estimate the terms of this decomposition in practice, it is generally necessary to assume the forecasts can manifest as only a finite number of possible options. This can often easily be achieved by grouping together similar forecasts, but a way to include the state-dependent information without some binning of the forecasts may be more convenient for forecasts in the form of predictive distributions.

Appendix 7.1: Ehm and Ovcharov (2017) decomposition

Ehm and Ovcharov (2017) introduce a similar factorisation of scoring functions in the presence of auxiliary information to that presented in Section 7.2. We show in this section how Equation 90 relates to the decomposition presented therein. The framework differs slightly in that Ehm and Ovcharov (2017) consider consistent scoring functions that assess point forecasts obtained by applying a functional to a predictive distribution. Nonetheless, there is a clear duality between consistent scoring functions and proper scoring rules (Gneiting and Ranjan, 2011; Thorarinsdottir and Schuhen, 2018), and the decompositions can thus easily be compared.

As alluded to in Section 7.2, Ehm and Ovcharov (2017) consider pairs of the form $W = (F, A)$, where A again denotes the possible states, while F represents the forecast. The decomposition thus allows predictions to be assessed separately depending on the value of the forecast as well as the prevailing state. This leads to the following local decomposition:

$$E_Y[S(F,Y)|W] = e(p(y|W)) + d(F, p(y|W)). \tag{102}$$

The entropy and divergence are defined similarly to before, with $e(G) = s(T(G), G)$ and $d(H, G) = s(T(H), G) - s(T(G), G)$ for the functional of interest T. This partition conditions the expected score on both the forecasts and some auxiliary information, whereas the local decomposition in Equation 88 depends only on the states. Therefore, Equation 88 can be obtained by taking the expectation of the components in Equation 102 with respect to the possible forecasts, and rearranging the expected entropy into the local uncertainty and resolution.

Ehm and Ovcharov (2017) then note that taking the expectation of Equation 102 with respect to W yields the overall score:

$$\begin{aligned}
E_W[E_Y[S(F,Y)|W]] &= E_W[e(p(y|W)] + E_W[d(F, p(y|W))], \\
&= E_W[s(T(p(y)), p(y|W)) - d(p(y), p(y|W))] + E_W[d(F, p(y|W))], \\
&= e(p(y)) - E_W[d(p(y), p(y|W))] + E_W[d(F, p(y|W))].
\end{aligned} \tag{103}$$

Writing $W = (F, A)$, this can be expressed in the terms of Equations 87 and 90. The first term is clearly equivalent to the uncertainty, UNC_Y, while the third term is equal to $REL_{F|A}$. The second term, on the other hand, is the resolution of W, which, using results in Section 7.2, can itself be decomposed into $RES_A + RES_{F|A}$. We therefore note that the latter two terms of Equation 103 are not equivalent to the resolution and reliability as defined in Equation 87 (Bröcker, 2009). Instead, they are both overestimated, by an amount equal to $RES_{A|F}$.

Appendix 7.2: Conditional decomposition of proper scoring rules

We demonstrate here how the components of the classical decomposition of proper scores can be expressed as the terms in Equation 90.

Firstly, the uncertainty term can be rewritten as

$$\begin{aligned} UNC_Y &= e(p(y)),\\ &= E_Y[S(p(y),Y)],\\ &= E_A[E_Y[S(p(y),Y)|A]],\\ &= E_A[s(p(y),p(y|A))],\\ &= E_A[s(p(y|A),p(y|A)) + d(p(y),p(y|A))],\\ &= E_A[e(p(y|A))] + E_A[d(p(y),p(y|A))],\\ &= UNC_{Y|A} + RES_A. \end{aligned} \tag{104}$$

The reliability term can similarly be decomposed:

$$\begin{aligned} REL_F &= E_F[d(F,p(y|F))],\\ &= E_F[s(F,p(y|F)) - s(p(y|F),p(y|F))],\\ &= E_F[E_Y[S(F,Y) - S(p(y|F),Y)|F]],\\ &= E_A[E_F[E_Y[S(F,y) - S(p(y|F),Y)|F,A]]],\\ &= E_A[E_F[s(F,p(y|F,A)) - s(p(y|F),p(y|F,A))|A]],\\ &= E_A[E_F[d(F,p(y|F,A)) - d(p(y|F),p(y|F,A))|A]],\\ &= E_A[E_F[d(F,p(y|F,A))|A]] - E_A[E_F[d(p(y|F),p(y|F,A))|A]],\\ &= REL_{F|A} - RES_{A|F}. \end{aligned} \tag{105}$$

Finally, combining Equations 89 and 90 we get

$$E_Y[S(F,Y)] = UNC_Y - RES_F + REL_F = UNC_{Y|A} - RES_{F|A} + REL_{F|A},$$

into which we can substitute the expressions for UNC_Y and REL_F derived above. Rearranging this obtains the breakdown of the resolution as given in Equation 90:

$$\begin{aligned} RES_F &= (UNC_Y - UNC_{Y|A}) + RES_{F|A} + (REL_F - REL_{F|A}),\\ &= RES_A + RES_{F|A} - RES_{A|F}. \end{aligned} \tag{106}$$

Appendix 7.3: Conditional decomposition of the Brier score

The uncertainty component of the classical Brier score decomposition is

$$\begin{aligned}
\bar{y}(1-\bar{y}) &= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j} - \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}\bar{y} \\
&= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}) + \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}^2 - \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}\bar{y} \\
&= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}) + \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}(\bar{y}_{\bullet j}^2 - 2\bar{y}_{\bullet j}\bar{y} + \bar{y}^2) \\
&= \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}(1-\bar{y}_{\bullet j}) + \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}(\bar{y}_{\bullet j} - \bar{y})^2, \\
&= UNC_{Y|A} + RES_A.
\end{aligned} \tag{107}$$

where we use the fact that

$$\bar{y} = \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}, \quad \text{and hence} \quad \bar{y}^2 = \sum_{j=1}^{J} \frac{n_{\bullet j}}{n}\bar{y}_{\bullet j}\bar{y}.$$

The reliability component can be expressed as

$$\begin{aligned}
\sum_{k=1}^{K} \frac{n_{k\bullet}}{n}(P_k - \bar{y}_{k\bullet})^2 &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}(P_k - \bar{y}_{k\bullet})^2 \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}(P_k - \bar{y}_{kj} + \bar{y}_{kj} - \bar{y}_{k\bullet})^2 \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}\Big[(P_k - \bar{y}_{kj})^2 + (\bar{y}_{kj} - \bar{y}_{k\bullet})^2 + 2(P_k - \bar{y}_{kj})(\bar{y}_{kj} - \bar{y}_{k\bullet})\Big] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}\Big[(P_k - \bar{y}_{kj})^2 + (\bar{y}_{kj} - \bar{y}_{k\bullet})^2 - 2(\bar{y}_{kj} - \bar{y}_{k\bullet})^2\Big] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}\Big[(P_k - \bar{y}_{kj})^2 - (\bar{y}_{kj} - \bar{y}_{k\bullet})^2\Big] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n}(P_k - \bar{y}_{kj})^2 - \sum_{j,k}^{J,K} \frac{n_{kj}}{n}(\bar{y}_{k\bullet} - \bar{y}_{kj})^2, \\
&= REL_{F|A} - RES_{A|F}.
\end{aligned} \tag{108}$$

Lastly, for the resolution,

$$
\begin{aligned}
\sum_{k=1}^{K} \frac{n_{k\bullet}}{n} (\bar{y}_{k\bullet} - \bar{y})^2 &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y})^2 \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y}_{kj} + \bar{y}_{kj} - \bar{y})^2 \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} \left[(\bar{y}_{k\bullet} - \bar{y}_{kj})^2 + (\bar{y}_{kj} - \bar{y})^2 + 2(\bar{y}_{k\bullet} - \bar{y}_{kj})(\bar{y}_{kj} - \bar{y}) \right] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} \left[(\bar{y}_{k\bullet} - \bar{y}_{kj})^2 + (\bar{y}_{kj} - \bar{y})^2 - 2(\bar{y}_{k\bullet} - \bar{y}_{kj})^2 \right] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} \left[(\bar{y}_{kj} - \bar{y})^2 - (\bar{y}_{k\bullet} - \bar{y}_{kj})^2 \right] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{kj} - \bar{y})^2 - \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y}_{kj})^2.
\end{aligned} \tag{109}
$$

The first term is the joint resolution of the forecasts and the states, $RES_{F,A}$, while the second term is $RES_{A|F}$. The former component can be partitioned into

$$
\begin{aligned}
\sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{kj} - \bar{y})^2 &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{kj} - \bar{y}_{\bullet j} + \bar{y}_{\bullet j} - \bar{y})^2 \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} \left[(\bar{y}_{kj} - \bar{y}_{\bullet j})^2 + (\bar{y}_{\bullet j} - \bar{y})^2 + 2(\bar{y}_{kj} - \bar{y}_{\bullet j})(\bar{y}_{\bullet j} - \bar{y}) \right] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} \left[(\bar{y}_{kj} - \bar{y}_{\bullet j})^2 + (\bar{y}_{\bullet j} - \bar{y})^2 \right] \\
&= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{kj} - \bar{y}_{\bullet j})^2 + \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{\bullet j} - \bar{y})^2, \\
&= RES_{F|A} + RES_{A}.
\end{aligned} \tag{110}
$$

Finally, it follows that

$$
\begin{aligned}
\sum_{k}^{K} \frac{n_{k\bullet}}{n} (\bar{y}_{k\bullet} - \bar{y})^2 &= \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{\bullet j} - \bar{y})^2 + \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{kj} - \bar{y}_{\bullet j})^2 - \sum_{j,k}^{J,K} \frac{n_{kj}}{n} (\bar{y}_{k\bullet} - \bar{y}_{kj})^2, \\
&= RES_{A} + RES_{F|A} - RES_{A|F},
\end{aligned} \tag{111}
$$

as desired.

Equations 108 - 110 all make use of the fact that

$$\sum_{j=1}^{J} \frac{n_{kj}}{n_{k\bullet}} (\bar{y}_{kj} - \bar{y}_{k\bullet}) = \sum_{k=1}^{K} \frac{n_{kj}}{n_{\bullet j}} (\bar{y}_{kj} - \bar{y}_{\bullet j}) = 0.$$

Appendix 7.4: Bias of Brier score estimators

In this section we discuss the nature of the biases of the Brier score components, making extensive use of results in Bröcker (2012a) and Ferro and Fricker (2012). To study the biases of the decomposition terms, it is first necessary to consider the components in the presence of unlimited data. To do so, let $\mu = E[Y^{(i)}]$, $\mu_{k\bullet} = E[Y^{(i)}|P^{(i)} = P_k]$, $\mu_{\bullet j} = E[Y^{(i)}|A^{(i)} = A_j]$, $\mu_{kj} = E[Y^{(i)}|P^{(i)} = P_k, A^{(i)} = A_j]$, be the expected value of the outcome conditional on the possible scenarios, for every possible i. $A^{(i)}$ here represents the state corresponding to the $i-$th forecast and observation. Define also the probability of forecast k occurring as $\phi_{k\bullet}$, the probability of state j occurring as $\phi_{\bullet j}$, and the probability of forecast k and state j both occurring as ϕ_{kj}. Then the components of Equation 97 become:

$$\begin{aligned} UNC_Y &= \mu(1-\mu), \\ RES_F &= \sum_{k=1}^{K} \phi_{k\bullet}(\mu - \mu_{k\bullet})^2, \\ REL_F &= \sum_{k=1}^{K} \phi_{k\bullet}(P_k - \mu_{k\bullet})^2, \end{aligned} \tag{112}$$

and the components of the new decomposition become:

$$\begin{aligned} UNC_{Y|A} &= \sum_{j=1}^{J} \phi_{\bullet j}\mu_{\bullet j}(1-\mu_{\bullet j}), \\ RES_A &= \sum_{j=1}^{J} \phi_{\bullet j}(\mu - \mu_{\bullet j})^2, \\ RES_{F|A} &= \sum_{j,k}^{J,K} \phi_{kj}(\mu_{k\bullet} - \mu_{kj})^2, \\ RES_{A|F} &= \sum_{j,k}^{J,K} \phi_{kj}(\mu_{\bullet j} - \mu_{kj})^2, \\ REL_{F|A} &= \sum_{j,k}^{J,K} \phi_{kj}(P_k - \mu_{kj})^2. \end{aligned} \tag{113}$$

Bröcker (2012a) proves that in the presence of finite data, the estimators of Equation 97, although commonly used in practice, are biased:

$$
\begin{aligned}
E[\widehat{UNC}_Y] - UNC_Y &= -\frac{\mu(1-\mu)}{n}, \\
E[\widehat{RES}_F] - RES_F &= \frac{1}{n}\sum_{k=1}^{K} v_{k,\bullet,n}\mu_{k\bullet}(1-\mu_{k\bullet}) - \frac{\mu(1-\mu)}{n}, \\
E[\widehat{REL}_F] - REL_F &= \frac{1}{n}\sum_{k=1}^{K} v_{k,\bullet,n}\mu_{k\bullet}(1-\mu_{k\bullet}),
\end{aligned}
\tag{114}
$$

where $v_{k,\bullet,n}$ is the probability that $n_{k\bullet}$ is zero. Assuming independent and identically distributed forecast-observation pairs, both Bröcker (2012a) and Ferro and Fricker (2012) introduce bias corrections that yield estimators for these terms with biases that decay considerably faster than those of the standard decomposition. The bias corrected estimators of Ferro and Fricker (2012) (which are implemented here in Section 7.5) are given in Equation 99. Ferro and Fricker (2012) prove that the resulting uncertainty component is unbiased, but no unbiased estimators exist for either the resolution or reliability of the decomposition. Since the uncertainty is unbiased, the biases in the resolution and reliability must be the same, and are equal to:

$$
E[\widetilde{RES}_F] - RES_F = E[\widetilde{REL}_F] - REL_F = \sum_{k=1}^{K} \phi_{k\bullet}(1-\phi_{k\bullet})^{n-1}\mu_{k\bullet}(1-\mu_{k\bullet}). \tag{115}
$$

Although the biases have not been eradicated, they decay exponentially, at a much faster rate than those in Equation 114.

Using a very similar approach to that of Bröcker (2012a), it is possible to quantify the biases of the new decomposition terms:

$$
\begin{aligned}
E[\widehat{UNC}_{Y|A}] - UNC_{Y|A} &= -\frac{1}{n}\sum_{j=1}^{J} v_{\bullet,j,n}\mu_{\bullet j}(1-\mu_{\bullet j}), \\
E[\widehat{RES}_A] - RES_A &= \frac{1}{n}\sum_{j=1}^{J} v_{\bullet,j,n}\mu_{\bullet j}(1-\mu_{\bullet j}) - \mu(1-\mu), \\
E[\widehat{RES}_{F|A}] - RES_{F|A} &= -\frac{1}{n}\sum_{j=1}^{J} v_{\bullet,j,n}\mu_{\bullet j}(1-\mu_{\bullet j}) + \frac{1}{n}\sum_{j,k}^{J,K} v_{k,j,n}\mu_{kj}(1-\mu_{kj}), \\
E[\widehat{RES}_{A|F}] - RES_{A|F} &= \frac{1}{n}\sum_{j,k}^{J,K} v_{k,j,n}\mu_{kj}(1-\mu_{kj}) - \frac{1}{n}\sum_{k=1}^{K} v_{k,\bullet,n}\mu_{k\bullet}(1-\mu_{k\bullet}), \\
E[\widehat{REL}_{F|A}] - REL_{F|A} &= \frac{1}{n}\sum_{j,k}^{J,K} v_{k,j,n}\mu_{kj}(1-\mu_{kj}).
\end{aligned}
\tag{116}
$$

Here, $v_{\bullet,j,n}$ and $v_{k,j,n}$ similarly represent the probabilities that $n_{\bullet j}$ and n_{kj}, respectively, are zero. It is straightforward to verify that these biases cancel out to reproduce the biases of the standard decomposition terms above. Noting the similar form of the bias of $UNC_{Y|A}$ and $REL_{F|A}$ to that of REL_F, results in Ferro and Fricker (2012) can easily be extended to show that no unbiased estimator exists for these terms. Since an unbiased estimator is attainable for UNC_Y but not $UNC_{Y|A}$, there is also no unbiased estimator for RES_A; if there were an unbiased estimator for both UNC_Y and RES_A then an unbiased estimator for $UNC_{Y|A}$ could be obtained, which we have argued is impossible. It is not immediately obvious whether an unbiased estimator exists for $RES_{F|A}$ or $RES_{A|F}$.

Nonetheless, the terms of Equation 100 seek to reduce the bias of the estimators. Again, using the similarity between these bias corrections and those of Equation 99, along with results in Ferro and Fricker (2012), the biases of these terms can easily be calculated:

$$\begin{aligned}
E[\widetilde{UNC}_{Y|A}] - UNC_{Y|A} &= -\sum_{j=1}^{J} \phi_{\bullet j}(1-\phi_{\bullet j})^{n-1}\mu_{\bullet j}(1-\mu_{\bullet j}),\\
E[\widetilde{RES}_{A}] - RES_{A} &= \sum_{j=1}^{J} \phi_{\bullet j}(1-\phi_{\bullet j})^{n-1}\mu_{\bullet j}(1-\mu_{\bullet j}),\\
E[\widetilde{RES}_{F|A}] - RES_{F|A} &= -\sum_{j=1}^{J} \phi_{\bullet j}(1-\phi_{\bullet j})^{n-1}\mu_{\bullet j}(1-\mu_{\bullet j})\\
&\qquad + \sum_{j,k}^{J,K} \phi_{kj}(1-\phi_{kj})^{n-1}\mu_{kj}(1-\mu_{kj}), \qquad (117)\\
E[\widetilde{RES}_{A|F}] - RES_{A|F} &= \sum_{j,k}^{J,K} \phi_{kj}(1-\phi_{kj})^{n-1}\mu_{kj}(1-\mu_{kj})\\
&\qquad - \sum_{k=1}^{K} \phi_{k\bullet}(1-\phi_{k\bullet})^{n-1}\mu_{k\bullet}(1-\mu_{k\bullet}),\\
E[\widetilde{REL}_{F|A}] - REL_{F|A} &= \sum_{j,k}^{J,K} \phi_{kj}(1-\phi_{kj})^{n-1}\mu_{kj}(1-\mu_{kj}).
\end{aligned}$$

It is again straightforward to confirm that when combined as in Equation 90, these formulae give the same biases as previously derived for $\widetilde{UNC}_Y, \widetilde{RES}_F$ and $\widetilde{REL}_F$. Furthermore, although biases are still present for all terms, they again decay at a much faster rate than those in Equation 116.

8 Concluding remarks

Numerical weather prediction (NWP) models aim to replicate the physical laws governing the atmosphere's trajectory. Due to computational advancements, these models are becoming progressively more complex, operating at higher resolutions and incorporating more intricate model physics. Nonetheless, systematic biases remain present in the model output due to incomplete knowledge of atmospheric dynamics, and errors when specifying the initial forecast state. To address this, NWP models are run from a variety of initial conditions to obtain a sample of distinct forecasts (Leith, 1974), possibly also with differing model physics. Although this ensemble of model runs can account for the flow-dependent uncertainty in the forecast (Leutbecher and Palmer, 2008), the comprising members are products of an imperfect prediction system. The result is a biased, typically overconfident ensemble forecast. Hence, to avoid misguided conclusions drawn directly from the model output, it has become imperative that the forecast undergoes some form of post-processing, or recalibration (Vannitsem et al., 2018).

This thesis has considered a range of approaches suitable for extending established statistical post-processing methods, which have become an integral component of weather forecasting over the last two decades. Chapter 6, for example, extends the existing functionality within IMPROVER, a post-processing suite currently in development at the UK Met Office, to permit asymmetric predictive distributions when recalibrating temperature forecast fields over the UK. The primary focus of the work presented herein, however, concerns a framework for incorporating atmospheric regimes into statistical post-processing methods, thus continuing a recent trend in the field of post-processing whereby additional sources of information are utilised in the statistical models. Atmospheric regimes are known to have a profound influence on the behaviour of local weather systems, and incorporating this information into post-processing methods can account for biases that occur due to an incorrect representation of the regimes by numerical weather models.

To incorporate weather regimes into established post-processing methods, we employ a mixture distribution that comprises a separate predictive distribution corresponding to each regime under consideration. To understand the behaviour of these regime-dependent mixture distributions, they have been implemented in a hierarchy of progressively more realistic scenarios in Chapters 3 - 5. To summarise the results presented in these chapters, it is convenient to distinguish between two categories of regime-dependent forecast errors. Errors of the first type materialise when the output from the numerical weather model accurately predicts the future weather regime. Although high resolution weather models are typically able to simulate the regimes observed in the atmosphere, they may not exactly reproduce the spatial structure of the regime centres, or the intensity of the corresponding pressure systems (Dawson et al., 2012), leading to forecast biases that are specific to particular weather regimes. This type of regime-dependent biases tends to occur more regularly at short lead times, when the information content in the dynamical weather model is largest.

Conversely, the second category of regime-dependent forecast errors corresponds to when the numerical weather model does not correctly predict the future weather regime. In this case, there is additional error in the forecast owing to differences between the climatological distribution of the outcome variable associated with each regime. Hence, if the outcome depends strongly on the prevailing weather regime, then the errors belonging to this second category are typically much larger than those in the first. Moreover, as lead time increases, the quality of the dynamical becomes progressively worse, meaning it is less capable of predicting the future atmospheric state, resulting in regime-dependent errors that are dominated by those in this second category.

At these larger lead times, the regime at the forecast initialisation time is not representative of the regime that will occur in the future, and, as such, the forecast biases become independent of the initial flow regime. Therefore, incorporating this initial regime into statistical post-processing methods is not found to be particularly useful for forecasts beyond a few days in advance. The regime predicted by a numerical model, on the other hand, provides a better indication of the future atmospheric state at longer lead times. As such, using the forecast regime within regime-dependent post-processing methods results in improvements upon conventional approaches at a larger range of lead times compared with the initial regime. However, as the quality of the raw forecast deteriorates, most state-of-the-art post-processing methods reduce the influence that the numerical model output has on the resulting forecast, using instead more information from the climatological distribution of the observations in the training data. As a result, although errors in the numerical weather model depend on both the predicted and realised, or ‘true’, weather regimes, conventional post-processing methods exhibit biases at longer lead times that are dominated by the true regime at the validation time.

The lead time at which improvements diminish when using the forecast regime depends on the regimes under consideration, and the ability of the weather model to predict them. In Chapter 3, for example, the regimes are persistent and hence more predictable, and improvements thus extend to forecast horizons beyond one week, whereas improvements curtail after only a couple of days for the more transient weather patterns considered in Chapter 5. Moreover, if a post-processing method is applied that does not reduce the influence of the weather model output on the post-processed forecast as its quality deteriorates, then regime-dependent extensions of such an approach would benefit from utilising both the observed and forecast regimes simultaneously, as is discussed in Chapter 4.

Furthermore, Chapter 4 also demonstrates that if the regime that manifests at the forecast validation time were known, then regime-dependent post-processing methods can account for the second category of regime-dependent forecast errors, in turn offering considerable improvements upon conventional post-processing methods when the prevailing regime markedly affects the behaviour of the response variable. In this sense, utilising the true regime when post-processing can account for biases that arise even when the forecast does not correctly predict the regime that will occur. The improvements gained by regime-dependent post-processing in this case tend to increase with forecast lead time, rather than deteriorating: progressively less information regarding the future atmosphere is contained by the raw model output, and hence incorporating additional, external information into post-processing methods is more beneficial at these longer lead times. In the limiting case, when the response variable is independent of the model output, statistical post-processing should issue the climatological distribution of the outcome. The regime-dependent post-processing methods considered here, on the other hand, tend towards the climatological distribution of the outcome given the weather regime on which the post-processing methods are conditioned. Hence, in general, to generate the largest improvements upon conventional methods, post-processing should be conditioned on the most accurate forecast of the weather regime at the forecast validation time.

The fact that the true weather regime can be used successfully within post-processing methods may seem trivial, since, fundamentally, the future state of the atmosphere is unknown. However, the synoptic-scale atmosphere is considerably more predictable than the more turbulent surface weather, and hence it may be comparatively straightforward to obtain a more accurate forecast of the future weather regime. To do so, one might utilise the regime transition matrix alongside the sequence of preceding weather regimes to account for the regime persistence, while previous studies have also recognised that the occurrence of certain weather regimes depends on larger-scale circulation patterns, such as the Madden-Julian Oscillation (Cassou, 2008) and the El-Nino Southern Oscillation (Robertson and Ghil, 1999). If this information could be used effectively to obtain an accurate forecast of the regime that will occur at the forecast validation time, then the large improvements gained from regime-dependent post-processing methods in Chapter 4 may be attainable, particularly if the numerical weather model alone is unable to predict the regime. Future applications of regime-dependent post-processing may wish to investigate this.

Nonetheless, results presented throughout this thesis have indicated that forecasts of surface weather variables exhibit biases that depend on the atmospheric regime, regardless of whether the dynamical model correctly simulates the regime that will occur at the forecast validation time. Regime-dependent post-processing is found to be most beneficial at locations and times of the year where the weather depends strongly on the prevailing weather regime, a result identified in all applications of regime-dependent post-processing presented here. Furthermore, since more extreme weather events are often associated with the occurrence of particular weather patterns, incorporating these regimes into statistical post-processing methods can significantly enhance forecasts made for these high-impact events, which are of a high societal importance. This is demonstrated in Chapter 5, where extreme wind speed events over the UK and Ireland are associated with the occurrence of the positive phase of the North Atlantic Oscillation (NAO), and hence incorporating information regarding the prevailing phase of the NAO into post-processing methods significantly enhances predictions of these extreme events.

This summarises the situations in which regime-dependent post-processing is expected to be beneficial, though the exact instances in which operational weather forecasts will improve as a result of these methods will depend on a variety of factors, including, for example, the prediction system, spatial domain, lead time, and weather variable under consideration. The results presented herein can therefore be used as a general indication as to when incorporating regime information will enhance the resulting forecasts, but to fully appreciate whether regime-dependent approaches should be applied in practice requires a detailed analysis of the performance of particular prediction systems.

Future work into regime-dependent post-processing methods could also consider the dependence of the forecast biases on the duration of the weather regimes. As noted by Woollings et al. (2018), “the strongest impacts of (atmospheric) blocking occur due to its persistence, which can allow temperature and moisture anomalies to build up over one or more weeks.” The biases in numerical weather models may thus be dependent not only on the weather regime that occurs, but also for how long it has resided. As well as considering the temporal behaviour of the regimes, weather regimes could also affect the multivariate dependencies between different weather variables and spatial locations, and this may translate to the forecast biases. If so, then applying the regime-dependent framework within multivariate post-processing methods should be beneficial, helping to achieve a more coherent forecast field. This could easily be achieved, for example, by implementing a separate dependence structure corresponding to each regime within copula-based post-processing approaches.

Furthermore, we have focused in particular on the application of these regime-dependent post-processing methods to wind speed forecasts. Other weather variables, including temperature, precipitation, and cloud cover are also likely to depend on the prevailing weather regime, and hence it would be of interest to apply the regime-dependent frameworks presented herein to forecasts of these variables. In particular, extreme temperature and precipitation events are both closely linked to the occurrence of certain weather regimes - heatwaves and cold snaps tend to occur during anticyclonic weather regimes, whereas extreme rainfall is associated with cyclonic patterns - and hence regime-dependent statistical post-processing is expected to generate more accurate predictions of these high-impact events.

Throughout this thesis, weather regimes have been incorporated into statistical post-processing methods through a mixture distribution, which provides an appealingly flexible framework. Alternative approaches to utilise weather regimes, however, may also be of interest. It may be the case, for example, that the effect the regimes have on a particular weather variable can be described reasonably well by the behaviour of a few auxiliary variables. Using these auxiliary variables as predictors in more data-driven post-processing approaches, such as neural networks, may thus generate predictive distributions that capture the information provided by the regimes. Comparisons of approaches to include this regime information would therefore be revealing. Nonetheless, regardless of the framework used to post-process weather forecasts, the results presented in this thesis have illustrated that it is necessary to evaluate the performance of weather forecasts conditional on the occurrence of local weather regimes. A theoretically desirable method to achieve this based on the decomposition of proper scoring rules is introduced in Chapter 7. If the post-processed forecasts are not found to be calibrated conditional on the local weather regimes, then an improvement in predictive performance may be realisable through the application of the regime-dependent statistical post-processing methods presented herein.

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., and Speranza, A. (2003). Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather and Forecasting*, 18:918–932.

Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F. (2021a). Accounting for skew when post-processing MOGREPS-UK temperature forecasts. *Monthly Weather Review (Under Review).*

Allen, S., Evans, G. R., Buchanan, P., and Kwasniok, F. (2021b). Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, 147:1403–1418.

Allen, S., Ferro, C. A. T., and Kwasniok, F. (2019). Regime-dependent statistical post-processing of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145:3535–3552.

Allen, S., Ferro, C. A. T., and Kwasniok, F. (2020). Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 146:2576–2596.

Alley, R. B., Emanuel, K. A., and Zhang, F. (2019). Advances in weather prediction. *Science*, 363:342–344.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9:1518–1530.

Atger, F. (2004). Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 130:627–646.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178.

Baldwin, M. P. (2001). Annular modes in global daily surface pressure. *Geophysical Research Letters*, 28(21):4115–4118.

Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, 75:227–238.

Baran, S. and Lerch, S. (2015). Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141:2289–2299.

Baran, S. and Lerch, S. (2016). Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27:116–130.

Baran, S. and Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34(3):477–496.

Baran, S. and Möller, A. (2017). Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorology and Atmospheric Physics*, 129:99–112.

Barnes, C., Brierley, C. M., and Chandler, R. E. (2019). New approaches to postprocessing of multi-model ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 145(725):3479–3498.

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525:47–55.

Baur, F., Hess, P., and Nagel, H. (1944). Kalender der grosswetterlagen Europas 1881–1939. *Bad Homburg*, 35.

Benedetti, R. (2010). Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211.

Bentzien, S. and Friederichs, P. (2014). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140:1924–1934.

Bernardo, J. M. (1979). Expected information as expected utility. *the Annals of Statistics*, pages 686–690.

Blattenberger, G. and Lad, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39:26–32.

Bowler, N. E. (2006). Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus A*, 58:538–548.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26:211–243.

Bremnes, J. B. (2004). Probabilistic wind power forecasts using local quantile regression. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 7:47–54.

Bremnes, J. B. (2019). Constrained quantile regression splines for ensemble postprocessing. *Monthly Weather Review*, 147:1769–1780.

Bremnes, J. B. (2020). Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, 148:403–414.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135:1512–1519.

Bröcker, J. (2012a). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, 39:655–667.

Bröcker, J. (2012b). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138:1611–1617.

Bröcker, J. and Smith, L. A. (2007a). Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22:651–661.

Bröcker, J. and Smith, L. A. (2007b). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22:382–388.

Brook, R. J. and Arnold, G. C. (2018). *Applied regression analysis and experimental design*. Routledge.

Buizza, R. (2018). Ensemble forecasting and the need for calibration. In *Statistical postprocessing of ensemble forecasts*, pages 15–48. Elsevier.

Buizza, R., Houtekamer, P., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133:1076–1097.

Buizza, R., Leutbecher, M., and Isaksen, L. (2008). Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134:2051–2066.

Buizza, R., Miller, M., and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125:2887–2908.

Buizza, R. and Palmer, T. N. (1995). The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Sciences*, 52:1434–1456.

Buizza, R. and Palmer, T. N. (1998). Impact of ensemble size on ensemble prediction. *Monthly Weather Review*, 126:2503–2518.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131:2131–2150.

Carrera, M., Higgins, R., and Kousky, V. (2004). Downstream weather impacts associated with atmospheric blocking over the northeast pacific. *Journal of Climate*, 17:4823–4839.

Cassou, C. (2008). Intraseasonal interaction between the madden–julian oscillation and the north atlantic oscillation. *Nature*, 455:523–527.

Cassou, C., Terray, L., Hurrell, J. W., and Deser, C. (2004). North Atlantic winter climate regimes: Spatial asymmetry, stationarity with time, and oceanic forcing. *Journal of Climate*, 17:1055–1068.

Charney, J. G. and DeVore, J. G. (1979). Multiple flow equilibria in the atmosphere and blocking. *Journal of the Atmospheric Sciences*, 36:1205–1216.

Charney, J. G., Fjörtoft, R., and Neumann, J. v. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, 2:237–254.

Cheng, X. and Wallace, J. M. (1993). Cluster analysis of the Northern Hemisphere wintertime 500-hpa height field: Spatial patterns. *Journal of the Atmospheric Sciences.*, 50:2674–2696.

Christensen, H. M., Moroz, I. M., and Palmer, T. N. (2015). Simulating weather regimes: Impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, 44:2195–2214.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5:243–262.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., et al. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137:1–28.

Corti, S., Molteni, F., and Palmer, T. N. (1999). Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, 398:799–802.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A. (2017). Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143:909–916.

Dabernig, M., Schicker, I., Kann, A., Wang, Y., and Lang, M. N. (2020). Statistical post-processing with standardized anomalies based on a 1 km gridded analysis. *Meteorologische Zeitschrift*, pages 265–275.

Daley, R. (1993). *Atmospheric data analysis*. Cambridge university press.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147:278–290.

Dawson, A. and Palmer, T. N. (2015). Simulating weather regimes: Impact of model resolution and stochastic parameterization. *Climate Dynamics*, 44:2177–2193.

Dawson, A., Palmer, T. N., and Corti, S. (2012). Simulating regime structures in weather and climate prediction models. *Geophysical Research Letters*, 39.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al. (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

DeGroot, M. H. and Fienberg, S. E. (1982). Assessing probability assessors: Calibration and refinement.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32:12–22.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141:3498–3516.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, 111.

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., and Stull, R. (2011). Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139:3554–3570.

Deloncle, A., Berk, R., d'Andrea, F., and Ghil, M. (2007). Weather regime prediction using statistical learning. *Journal of the Atmospheric Sciences*, 64:1619–1635.

Demaeyer, J. and Vannitsem, S. (2020). Correcting for model changes in statistical postprocessing–an approach based on response theory. *Nonlinear Processes in Geophysics*, 27:307–327.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–22.

Di Narzo, A. and Cocchi, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(3):405–422.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20:134–144.

Dole, R. M. and Gordon, N. D. (1983). Persistent anomalies of the extratropical northern hemisphere wintertime circulation: Geographical distribution and regional persistence characteristics. *Monthly Weather Review*, 111:1567–1586.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

Ehm, W. and Ovcharov, E. (2017). Bias-corrected score decomposition for generalized quantiles. *Biometrika*, 104:473–480.

Eide, S. S., Bremnes, J. B., and Steinsland, I. (2017). Bayesian model averaging for wind speed ensemble forecasts using wind speed and direction. *Weather and Forecasting*, 32:2217–2227.

Epstein, E. S. (1969a). The role of initial uncertainties in predicion. *Journal of Applied Meteorology*, 8:190–198.

Epstein, E. S. (1969b). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987.

Epstein, E. S. (1969c). Stochastic dynamic prediction. *Tellus*, 21:739–759.

Evans, G. R., Ayliffe, B. A., Sandford, C., Rust, F. M., Jones, C., Fitzpatrick, B. J. R., Moseley, S. R., Hopkinson, A. R., Pillinger, T., Baker, M., Crosswaite, N., Howard, K., Beard, L., Worsfold, M., Abernethy, P., Trzeciak, T. M., Gale, T., Jackson, S. D., Booton, A., Smith, E., Kinoshita, B. P., Sampson, C., Hume, T., Canvin, J., Friedrich, M., Roberts, N. M., Wright, B. J., and Nic Guidhir, M. (2020). metoppv/IMPROVER: IMPROVER: A library of algorithms for meteorological post-processing. (version 0.10.0).

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99:10143–10162.

Falkena, S. K., de Wiljes, J., Weisheimer, A., and Shepherd, T. G. (2020). Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 146:2801–2814.

Feldmann, K., Richardson, D. S., and Gneiting, T. (2019). Grid-versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, 46:7744–7751.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review*, 143:955–971.

Ferranti, L. and Corti, S. (2011). *New clustering products*. ECMWF.

Ferranti, L., Corti, S., and Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141:916–924.

Ferro, C. A. T. (2017). Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society*, 143:2665–2676.

Ferro, C. A. T. and Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138:1954–1960.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15:19–24.

Fleming, R. J. (1971). On stochastic dynamic prediction. *Monthly Weather Review*, 99:1236.

Flowerdew, J. (2014). Calibrating ensemble reliability whilst preserving spatial structure. *Tellus A*, 66:22662.

Fraley, C., Raftery, A. E., and Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138:190–202.

Franzke, C., Crommelin, D., Fischer, A., and Majda, A. J. (2008). A hidden Markov model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, 21:1740–1757.

Franzke, C., Woollings, T., and Martius, O. (2011). Persistent circulation regimes and preferred regime transitions in the North Atlantic. *Journal of the Atmospheric Sciences.*, 68:2809–2825.

Friederichs, P. (2010). Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13:109–132.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23:579–594.

Friederichs, P., Wahl, S., and Buschow, S. (2018). Postprocessing for extreme events. In *Statistical Postprocessing of Ensemble Forecasts*, pages 127–154. Elsevier.

Friedli, L., Ginsbourger, D., and Bhend, J. (2020). Area-covering postprocessing of ensemble precipitation forecasts using topographical and seasonal conditions. *Stochastic Environmental Research and Risk Assessment*, pages 1–16.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). Fine-tuning nonhomogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Monthly Weather Review*, 145:4693–4708.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2018). Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146:4323–4338.

Gebetsberger, M., Stauffer, R., Mayr, G. J., and Zeileis, A. (2019). Skewed logistic distribution for statistical temperature post-processing in mountainous areas. *Advances in Statistical Climatology, Meteorology and Oceanography*, 5:87–100.

Glahn, B., Peroutka, M., Wiedenfeld, J., Wagner, J., Zylstra, G., Schuknecht, B., and Jackson, B. (2009). MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137:246–268.

Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11:1203–1211.

Gleeson, T. A. (1967). On theoretical limits of predictability. *Journal of Applied Meteorology*, 6:213–215.

Gleeson, T. A. (1970). Statistical-dynamical predictions. *Journal of Applied Meteorology*, 9:333–344.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101:968–979.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133:1098–1118.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29:411–422.

Gneiting, T., Ranjan, R., et al. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:211.

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society*, 14A:107–114.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H. (2017). Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, 7:557–562.

Greybush, S. J., Haupt, S. E., and Young, G. S. (2008). The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Weather and Forecasting*, 23:1146–1161.

Grimit, E. P., Gneiting, T., Berrocal, V., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132:2925–2942.

Gupta, R. D. and Kundu, D. (2010). Generalized logistic distributions. *Journal of Applied Statistical Science*, 18:51.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A*, 57:219–233.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W. (2017). The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 143:2846–2861.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560.

Hamill, T. M. (2018). Practical aspects of statistical postprocessing. In *Statistical postprocessing of ensemble forecasts*, pages 187–217. Elsevier.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., Zhu, Y., and Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc*, 94:1553–1565.

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125:1312–1327.

Hamill, T. M., Engle, E., Myrick, D., Peroutka, M., Finan, C., and Scheuerer, M. (2017). The US national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Monthly Weather Review*, 145:3441–3463.

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134:3209–3229.

Hamill, T. M., Whitaker, J. S., and Wei, X. (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132:1434–1447.

Hannachi, A., Straus, D. M., Franzke, C. L., Corti, S., and Woollings, T. (2017). Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55:199–234.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hemri, S., Haiden, T., and Pappenberger, F. (2016). Discrete postprocessing of total cloud cover ensemble forecasts. *Monthly Weather Review*, 144(7):2565–2577.

Hemri, S., Lisniak, D., and Klein, B. (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resources Research*, 51:7436–7451.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41:9197–9205.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2019). Isotonic distributional regression. *arXiv preprint arXiv:1909.03725*.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting—with applications to risk management. *The Annals of Applied Statistics*, 8:595–621.

Holzmann, H., Klar, B., et al. (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics*, 11:2404–2431.

Horel, J. D. (1981). A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. *Monthly Weather Review*, 109:2080–2092.

Horel, J. D. (1985). Persistence of the 500 mb height field during Northern Hemisphere winter. *Monthly Weather Review*, 113:2030–2042.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124:1225–1242.

Hurrell, J. W. (1995). Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, 269:676–679.

Hurrell, J. W. and Deser, C. (2009). North Atlantic climate variability: the role of the North Atlantic Oscillation. *Journal of Marine Systems*, 79:231–244.

Johnson, C. and Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, 135:777–794.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*. John Wiley & Sons, Ltd.

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Jordan, A., Krüger, F., and Lerch, S. (2017). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90:1–37.

Junk, C., Delle Monache, L., and Alessandrini, S. (2015). Analog-based ensemble model output statistics. *Monthly Weather Review*, 143:2909–2917.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77:437–472.

Kharin, V. V. and Zwiers, F. W. (2002). Climate predictions with multimodel ensembles. *Journal of Climate*, 15:793–799.

Kimoto, M. and Ghil, M. (1993). Multiple flow regimes in the Northern Hemisphere winter. Part I: Methodology and hemispheric regimes. *Journal of the Atmospheric Sciences.*, 50:2625–2644.

Klein, N., Kneib, T., Lang, S., Sohn, A., et al. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in germany. *The Annals of Applied Statistics*, 9:1024–1052.

Klein, W. H., Lewis, B. M., and Enger, I. (1959). Objective prediction of five-day mean temperatures during winter. *Journal of Meteorology*, 16:672–682.

Kober, K., Craig, G., and Keil, C. (2014). Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 140:1179–1188.

Koch, S. E., Skillman, W. C., Kocin, P. J., Wetzel, P. J., Brill, K. F., Keyser, D. A., and McCumber, M. C. (1985). Synoptic scale forecast skill and systematic errors in the MASS 2.0 model. *Monthly Weather Review*, 113:1714–1737.

Kondrashov, D., Ide, K., and Ghil, M. (2004). Weather regimes and preferred transition paths in a three-level quasigeostrophic model. *Journal of the Atmospheric Sciences.*, 61:568–587.

Krüger, F., Lerch, S., Thorarinsdottir, T. L., and Gneiting, T. (2016). Predictive inference based on Markov chain Monte Carlo output. *arXiv preprint arXiv:1608.06802*.

Kwasniok, F. (2007). Reduced atmospheric models using dynamically motivated basis functions. *Journal of the Atmospheric Sciences.*, 64:3452–3474.

Kwasniok, F. (2012). Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society*, 370:1061–1086.

Kwasniok, F. (2019). Fluctuations of finite-time Lyapunov exponents in an intermediate-complexity atmospheric model: a multivariate and large-deviation perspective. *Nonlinear Processes in Geophysics*, 26:195–209.

Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R., and Zeileis, A. (2019a). Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes in Geophysics*, 27.

Lang, M. N., Mayr, G. J., Stauffer, R., and Zeileis, A. (2019b). Bivariate Gaussian models for wind vectors in a distributional regression framework. *Advances in Statistical Climatology, Meteorology and Oceanography*, 5:115–132.

Leathers, D. J., Yarnal, B., and Palecki, M. A. (1991). The Pacific/North American teleconnection pattern and united states climate. part I: Regional temperature and precipitation associations. *Journal of Climate*, 4:517–528.

Leith, C. (1971). Atmospheric predictability and two-dimensional turbulence. *Journal of the Atmospheric Sciences*, 28(2):145–161.

Leith, C. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102:409–418.

Lemcke, C. and Kruizinga, S. (1988). Model output statistics forecasts: Three years of operational experience in the Netherlands. *Monthly Weather Review*, 116:1077–1090.

Lerch, S. and Baran, S. (2017). Similarity-based semilocal estimation of post-processing models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66:29–51.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S., and Graeter, M. (2020). Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, 27:349–371.

Lerch, S. and Thorarinsdottir, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65:21206.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, pages 106–127.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H. M., Diamantakis, M., Dutra, E., et al. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143:2315–2339.

Leutbecher, M. and Palmer, T. N. (2008). Ensemble forecasting. *Journal of computational physics*, 227:3515–3539.

Lewis, J. M. (2005). Roots of ensemble forecasting. *Monthly Weather Review*, 133:1865–1885.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21:289–307.

Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1.

Lorenz, E. N. (2014). *The essence of chaos*. CRC Press.

Lynch, P. (2006). *The emergence of numerical weather prediction: Richardson's dream*. Cambridge University Press.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of computational physics*, 227:3431–3444.

Magnusson, L., Leutbecher, M., and Källén, E. (2008). Comparison between singular vectors and breeding vectors as initial perturbations for the ECMWF ensemble prediction system. *Monthly Weather Review*, 136:4092–4104.

Majda, A. J., Franzke, C. L., Fischer, A., and Crommelin, D. T. (2006). Distinct metastable atmospheric regimes despite nearly Gaussian statistics: A paradigm model. *Proceedings of the National Academy of Sciences*, 103:8309–8314.

Marshall, J. and Molteni, F. (1993). Toward a dynamical understanding of planetary-scale flow regimes. *Journal of the Atmospheric Sciences*, 50:1792–1818.

Mason, S. J. (2004). On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Monthly Weather Review*, 132:1891–1895.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

Matsueda, M. and Palmer, T. (2018). Estimates of flow-dependent predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 144:1012–1027.

McCullagh, P. (2018). *Generalized linear models*. Routledge.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98:2073–2090.

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142:3003–3014.

Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145:137–147.

Michelangeli, P.-A., Vautard, R., and Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences.*, 52:1237–1256.

Mitchell, K. (2020). *Score Decompositions in Forecast Verification*. PhD thesis, University of Exeter.

Möller, A. and Groß, J. (2020). Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model. *Quarterly Journal of the Royal Meteorological Society*, 146:211–224.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: Part I. two-state situation. *Journal of Applied Meteorology*, 11:273–282.

Murphy, A. H. (1973a). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12:215–223.

Murphy, A. H. (1973b). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8:281–293.

Murphy, A. H. (1995). A coherent method of stratification within a general framework for forecast verification. *Monthly Weather Review*, 123:1582–1588.

Murphy, A. H., Brown, B. G., and Chen, Y.-S. (1989). Diagnostic verification of temperature forecasts. *Weather and Forecasting*, 4:485–501.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Neal, R., Fereday, D., Crocker, R., and Comer, R. E. (2016). A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, 23:389–400.

O'Lenic, E. A. and Livezey, R. E. (1989). Relationships between systematic errors in medium range numerical forecasts and some of the principal modes of low-frequency variability of the Northern Hemisphere 700 mb circulation. *Monthly Weather Review*, 117:1262–1280.

Oliver, H., Shin, M., Matthews, D., Sanders, O., Bartholomew, S., Clark, A., Fitzpatrick, B., van Haren, R., Hut, R., and Drost, N. (2019). Workflow automation for cycling systems: The Cylc workflow engine. *Computing in Science & Engineering*.

Oliver, H., Shin, M., and Sanders, O. (2018). Cylc: A workflow engine for cycling systems. *J. Open Source Software*, 3:737.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128:747–774.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U. (2018). Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, 144:1864–1881.

Philipp, A., Della-Marta, P.-M., Jacobeit, J., Fereday, D. R., Jones, P. D., Moberg, A., and Wanner, H. (2007). Long-term variability of daily North Atlantic–European pressure patterns since 1850 classified by simulated annealing clustering. *Journal of Climate*, 20:4065–4095.

Pinson, P. (2012). Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138(666):1273–1284.

Pinson, P. and Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20.

Porson, A. N., Carr, J. M., Hagelin, S., Darvell, R., North, R., Walters, D., Mylne, K. R., Mittermaier, M. P., Willington, S., and Macpherson, B. (2020). Recent upgrades to the Met Office convective-scale ensemble: an hourly time-lagged 5-day ensemble. *Quarterly Journal of the Royal Meteorological Society*.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900.

Reinhold, B. B. and Pierrehumbert, R. T. (1982). Dynamics of weather regimes: Quasi-stationary waves and blocking. *Monthly Weather Review*, 110:1105–1145.

Richardson, L. F. (2007). *Weather prediction by numerical process*. Cambridge university press.

Robertson, A. W. and Ghil, M. (1999). Large-scale weather regimes and local climate over the western united states. *Journal of Climate*, 12:1796–1813.

Roebber, P. J. (1998). The regime dependence of degree day forecast technique, skill, and value. *Weather and Forecasting*, 13:783–794.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660.

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A*, 55:16–30.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, 2:191–201.

Sansom, P. G., Ferro, C. A. T., Stephenson, D. B., Goddard, L., and Mason, S. J. (2016). Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. *Journal of Climate*, 29:7247–7264.

Schefzik, R. (2016). A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, 144:1909–1921.

Schefzik, R. and Möller, A. (2018). Ensemble postprocessing methods incorporating dependence structures. In *Statistical Postprocessing of Ensemble Forecasts*, pages 91–125. Elsevier.

Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., et al. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28:616–640.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140:1086–1096.

Scheuerer, M. and Büermann, L. (2014). Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 405–422.

Scheuerer, M. and Hamill, T. M. (2015a). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143:4578–4596.

Scheuerer, M. and Hamill, T. M. (2015b). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143:1321–1334.

Scheuerer, M. and König, G. (2014). Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140:2582–2590.

Scheuerer, M. and Möller, D. (2015). Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, 9:1328–1349.

Scheuerer, M., Schaback, R., and Schlather, M. (2013). Interpolation of spatial data–A stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24:601–629.

Schuhen, N., Thorarinsdottir, T., and Lenkoski, A. (2020). Rapid adjustment and post-processing of temperature forecast trajectories. *Quarterly Journal of the Royal Meteorological Society*, 146:963–978.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, 140:3204–3219.

Serreze, M. C., Carse, F., Barry, R. G., and Rogers, J. C. (1997). Icelandic low cyclone activity: Climatological features, linkages with the NAO, and relationships with recent changes in the northern hemisphere circulation. *Journal of Climate*, 10:453–464.

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143:1178–1183.

Siegert, S., Sansom, P. G., and Williams, R. M. (2016a). Parameter uncertainty in forecast recalibration. *Quarterly Journal of the Royal Meteorological Society*, 142:1213–1221.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., and Arribas, A. (2016b). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *Journal of Climate*, 29:995–1012.

Slingo, J. and Palmer, T. (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369:4751–4767.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105:25–35.

Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135:3209–3220.

Smyth, P., Ide, K., and Ghil, M. (1999). Multiple regimes in northern hemisphere height fields via mixture model clustering. *Journal of the Atmospheric Sciences.*, 56:3704–3723.

Stephenson, D. B., Coelho, C. A. S., Doblas-Reyes, F. J., and Balmaseda, M. (2005). Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, 57:253–264.

Stephenson, D. B., Coelho, C. A. S., and Jolliffe, I. T. (2008). Two extra components in the Brier score decomposition. *Weather and Forecasting*, 23:752–757.

Stephenson, D. B. and Doblas-Reyes, F. J. (2000). Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A*, 52:300–322.

Stephenson, D. B., Hannachi, A., and O'Neill, A. (2004). On the existence of multiple climate regimes. *Quarterly Journal of the Royal Meteorological Society*, 130:583–605.

Stoss, L. A. and Mullen, S. L. (1995). The dependence of short-range 500-mb height forecasts on the initial flow regime. *Weather and Forecasting*, 10:353–368.

Strähl, C., Ziegel, J., et al. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, 11:608–639.

Taillardat, M., Fougères, A.-L., Naveau, P., and Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34:617–634.

Taillardat, M. and Mestre, O. (2020). From research to applications–examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, 27:329–347.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144:2375–2393.

Talagrand, O. (1997a). Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan.*, 75:191–209.

Talagrand, O. (1997b). Evaluation of probabilistic prediction systems. In *Proc. ECMWF Workshop on predictability, ECMWF, Reading, UK.*

Tang, Y., Lean, H. W., and Bornemann, J. (2013). The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteorological Applications*, 20:417–426.

Thompson, D. W. and Wallace, J. M. (1998). The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Research Letters*, 25:1297–1300.

Thompson, P. D. (1957). Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, 9:275–295.

Thorarinsdottir, T. L. and Gneiting, T. (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society*, 173:371–388.

Thorarinsdottir, T. L. and Schuhen, N. (2018). Verification: Assessment of calibration and accuracy. In *Statistical postprocessing of ensemble forecasts*, pages 155–186. Elsevier.

Tibaldi, S. and Molteni, F. (1990). On the operational predictability of blocking. *Tellus A*, 42:343–365.

Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140:2005–2017.

Toth, Z. and Kalnay, E. (1993). Ensemble forecasting at NMC: The generation of perturbations. *Bulletin of the American Meteorological Society*, 74:2317–2330.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. *Available at SSRN 2236605*.

Vallis, G. K. (2017). *Atmospheric and oceanic fluid dynamics*. Cambridge University Press.

Van den Dool, H. (1989). A new look at weather forecasting through analogues. *Monthly Weather Review*, 117:2230–2247.

Van Schaeybroeck, B. and Vannitsem, S. (2015). Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141:807–818.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., et al. (2020). Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world. *arXiv preprint arXiv:2004.06582*.

Vannitsem, S., Wilks, D. S., and Messner, J. (2018). *Statistical postprocessing of ensemble forecasts*. Elsevier.

Vautard, R. (1990). Multiple weather regimes over the north atlantic: Analysis of precursors and successors. *Monthly Weather Review*, 118:2056–2081.

Von Storch, H. and Zwiers, F. W. (2001). *Statistical analysis in climate research*. Cambridge university press.

Wallace, J. M. and Gutzler, D. S. (1981). Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 109:784–812.

Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., et al. (2017). The Met Office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations. *Geoscientific Model Development*, 10:1487–1520.

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131:965–986.

Wei, M., Toth, Z., Wobus, R., and Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, 60:62–79.

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2007). The discrete Brier and ranked probability skill scores. *Monthly Weather Review*, 135:118–124.

Weijs, S. V., Van Nooijen, R., and Van De Giesen, N. (2010). Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138:3387–3399.

Whitaker, J. S. and Loughe, A. F. (1998). The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, 126:3292–3302.

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128:2821–2836.

Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, 131:389–407.

Wilks, D. S. (2006). Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, 13:243–256.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution mos forecasts. *Meteorological Applications*, 16:361–368.

Wilks, D. S. (2016). "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society*, 97:2263–2273.

Wilks, D. S. (2018). Univariate ensemble postprocessing. In *Statistical postprocessing of ensemble forecasts*, pages 49–89. Elsevier.

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences*. Amsterdam: Elsevier.

Williams, R., Ferro, C., and Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140:1112–1120.

Williams, R. M. (2016). *Statistical methods for post-processing ensemble weather forecasts*. PhD thesis, University of Exeter.

Woollings, T., Barriopedro, D., Methven, J., Son, S.-W., Martius, O., Harvey, B., Sillmann, J., Lupo, A. R., and Seneviratne, S. (2018). Blocking and its response to climate change. *Current Climate Change Reports*, 4:287–300.

Woollings, T., Hannachi, A., and Hoskins, B. (2010). Variability of the North Atlantic eddy-driven jet stream. *Quarterly Journal of the Royal Meteorological Society*, 136:856–868.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30:132–156.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Young, R. (2010). Decomposition of the Brier score for weighted forecast-verification pairs. *Quarterly Journal of the Royal Meteorological Society*, 136:1364–1370.

Zorita, E. and Von Storch, H. (1999). The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of climate*, 12:2474–2489.