Stochastic Downscaling to Chaotic Weather Regimes Using Spatially Conditioned Gaussian Random Fields with Adaptive Covariance

RACHEL PRUDDEN,[a,b] NIALL ROBINSON,[a,b] PETER CHALLENOR,[b] AND RICHARD EVERSON[b]

[a] *Met Office Informatics Lab, Exeter, United Kingdom*
[b] *University of Exeter, Exeter, United Kingdom*

(Manuscript received 20 November 2020, in final form 3 September 2021)

ABSTRACT: Downscaling aims to link the behavior of the atmosphere at fine scales to properties measurable at coarser scales, and has the potential to provide high-resolution information at a lower computational and storage cost than numerical simulation alone. This is especially appealing for targeting convective scales, which are at the edge of what is possible to simulate operationally. Since convective-scale weather has a high degree of independence from larger scales, a generative approach is essential. We here propose a statistical method for downscaling moist variables to convective scales using conditional Gaussian random fields, with an application to wet bulb potential temperature (WBPT) data over the United Kingdom. Our model uses an adaptive covariance estimation to capture the variable spatial properties at convective scales. We further propose a method for the validation, which has historically been a challenge for generative models.

KEYWORDS: Inverse methods; Statistical techniques; Probability forecasts/models/distribution

1. Introduction

Accurate numerical simulation of convective-scale weather is intensive in its demand for computational resources. This is partly a consequence of increasing spatiotemporal resolution, which increases computational demand geometrically. Beyond this, the atmosphere has a high level of inherent chaotic variability at convective scales, leading to a demand for larger ensembles to capture the predictive distribution to the fullest extent possible. The challenge of meeting these competing demands for computing power, as well as the difficulty of utilizing the vast amounts of data produced, leads to the question: can more be done with lower-resolution data?

Downscaling aims to address this question by linking the behavior of the atmosphere at fine scales to properties at coarser scales. In particular, statistical downscaling defines these connections without the use of a physics-based dynamical model. Statistical downscaling thus has the potential to provide high-resolution information at a lower computational and storage cost than numerical simulation alone.

In some instances small scale structure is deterministically and directly related to the large scales, for instance, when the structure is driven by orographic forcing. In such cases, downscaling can be treated as a deterministic regression problem. The situation is similar when the target variables consist of temporal averages, as is usually the case in climate applications, because the temporal averaging has the effect of smoothing out small-scale variability.

In other cases, notably the generation of chaotic structure, this approach becomes ineffective. However, there is reason for optimism, as bulk statistical properties of the target field may still be strongly constrained by the larger scales: the spatial characteristics of the target field may be highly predictable even if its point values are not. In such cases, rather than making a deterministic prediction, downscaling is best viewed probabilistically. The aim is then to determine the distribution of possible target fields and draw samples from this distribution.

An example of where this approach can be valuable is in downscaling to convective scales. While convective-scale weather has a high degree of independence from larger scales, its overall spatial properties tend to be more predictable. Furthermore, the spatial properties of cloud and rain are highly relevant for applications such as energy forecasting and hydrology, where spatiotemporal variability can be as important as the expected value (Liu et al. 2019; Schaake et al. 2007). This work directly targets stochastic downscaling of moist variables to convective scales, using a statistical model called Gaussian random fields (GRFs). GRFs are statistical models of spatial functions; like Gaussian processes (Rasmussen and Williams 2006) they define a distribution over such functions with a single sample corresponding to a spatially coherent function. This spatial interpretation, combined with convenient methods for conditioning on observed data, makes them a useful tool for statistical downscaling.

In contrast with most of the GRF literature, we use spatial conditioning to model the connection between small and large scales: that is, we model our low-resolution data as an areal average over the high-resolution grid where the target data are defined. A more standard approach would be to model all observations as belonging to point locations, as in Wikle et al. (2001), but this would fail to account for the ensemble properties of moist variables associated with convection at the subgrid scale (Arakawa and Schubert 1974). Spatial conditioning enables a faithful representation of both scales and the relationship between them (Gotway and Young 2002; Kyriakidis 2004).

Historically, the lack of a standardized and comprehensive approach to verification has been a challenge for generative models, including their application to downscaling (Mathieu et al. 2015). Standard point-based scores such as mean-squared error (MSE) are not well suited to assessing the skill of

Corresponding author: Rachel Prudden, rachel.prudden@informaticslab.co.uk

DOI: 10.1175/WAF-D-20-0217.1

stochastic models due to issues like the double-penalty problem (Gilleland et al. 2009). This is a known challenge in generative modeling (Mathieu et al. 2015), with the consequence that subjective value judgements and other heuristic scores are common (Barratt and Sharma 2018). In this work we take a multifaceted approach to verification and propose a novel spatial verification metric that allows for simple comparisons between stochastic models (see also references in section 5b).

The main contributions of this paper are as follows:

- we explain the spatial conditioning of GRFs with reference to stochastic downscaling of atmospheric variables
- we introduce an adaptive GRF model for downscaling, which combines spatial conditioning with an instantaneous length scale estimation, allowing for differing weather regimes
- we introduce a novel verification score for stochastic models, the neighborhood Wasserstein score.

We begin by describing related work in section 2. The problem setup is outlined in section 3, the theory behind our GRF model in section 4, our approach to verification in section 5, and section 6 describes an application to wet bulb potential temperature over the United Kingdom.

2. Related work

a. Deterministic downscaling

Deterministic methods remain widely used in downscaling, especially for temporally aggregated and non-moist variables. Downscaling can be treated as a deterministic regression problem by expressing the values of a high-resolution weather field in terms of low-resolution covariates:

$$\mathbf{y} = f(\mathbf{X}),$$

where f is referred to as a regression or transfer function. As discussed in Wilby et al. (2004), the philosophy underlying these methods is that the local weather or climate scenario depends on a combination of the large-scale situation and local geographic features. From this perspective, the function of f is to encode the effects of these local features.

A diverse range of algorithms may be used for the transfer function. For example, linear and ridge regression were used for site-specific downscaling of temperature and precipitation in Hessami et al. (2008), while Ghosh (2010) uses support vector regression for a similar use-case. More recently, Vandal et al. (2017) has applied convolutional neural networks to gridded precipitation downscaling. Simple methods are often effective: BCSD (Wood 2002), a linear method based on single shifting and scaling factors, was found by Vandal et al. (2019) to outperform more complex, nonlinear approaches.

While deterministic methods can be effective for some variables, due to their underlying assumptions they are ill-suited for modeling moist variables over short time periods (Raut et al. 2019; Schoof and Pryor 2001). This has motivated research into stochastic methods, outlined in the next section.

b. Stochastic downscaling

Stochastic downscaling using statistical models is orders of magnitude less computationally demanding than the dynamical equivalent (Bordoy and Burlando 2014). In particular they do not require the use of a supercomputer, meaning that they can be employed on demand by users outside of traditional HPC centers for regions of interest. Their relatively low cost also makes it possible to create larger ensembles that would be infeasible using dynamical models.

The principle behind stochastic downscaling is to use information about the small-scale structural properties of the target field to inform the construction of the downscaled fields. To do stochastic downscaling, we must make some assumption about how the field looks at smaller scales. For instance, we might assume that the field looks the same as past data, that it follows a power scaling law, or that it has a known covariance structure.

Assuming similarity with past data leads us to the use of ensemble analogs. The idea is essentially the same as the k-nearest-neighbors algorithm: a database of previously observed, high-resolution fields is searched for the fields which most closely resemble the current situation, with similarity based on smooth synoptic variables such as pressure (Delle Monache et al. 2013). Analog methods have the advantage of being nonparametric, meaning they are easy to apply to variables with arbitrary distributions, and able to capture consistent spatial and multivariate properties (Zorita and von Storch 1999). However, the analog approach relies on having a large enough dataset of potential analogs to be confident of finding a close match, which Van Den Dool (1994) argues is unrealistic. In general, these methods are constrained by the completeness of their datasets; for example, they are unable to predict extremes which lie outside the previously observed range.

An alternative approach is to make use of approximate power-scaling laws which are empirically observed for variables such as precipitation (Gupta and Waymire 1993). Power-scaling law assumptions can be used as a basis for various models (including random field models, as in Wikle et al. (2001)). One example is multiplicative cascade methods (Perica and Foufoula-Georgiou 1996), which are based on repeated use of the Haar wavelet transform (Strang 1993) to create a field at twice the original resolution by stochastically sampling directional fluctuations. The power law assumption means that, unlike analogs, multiplicative cascade methods can be effective with small datasets. They are also generally successful in simulating realistic distributions of rainfall intensities. However, multiplicative cascade methods do not lend themselves to modeling spatial correlations and tend to introduce artificial discontinuities (Gagnon et al. 2012). This provides the motivation for methods based on random fields, which directly model the covariance between grid cells and can better represent spatial structure.

Gaussian random fields approach downscaling by using an assumed covariance form to infer small-scale structure. As the main topic of this work, they are reviewed in more detail in the following section.

c. Gaussian random fields

Gaussian random fields are a geostatistical method used in a range of spatial inference problems in fields such as geological sciences (Li et al. 2016) and disease mapping (usually in the form

of Gauss Markov random fields—see e.g., Martinez-Beneito (2013)). They are mathematically equivalent to the machine learning method Gaussian processes (Rasmussen and Williams 2006), with the former term reserved for models with a spatial interpretation. The term "kriging" is also found in the literature and is more or less equivalent. GRFs have been applied to various problems in atmospheric science, such as modeling wind (Hewer et al. 2017; Wikle et al. 2001), precipitation (Nychka et al. 2015), and energy production (Wytock and Kolter 2013; Zhang et al. 2016).

Besides the usual application of interpolation from point observations, Gaussian random fields can be applied in the more general case of observations based on areal averages, sometimes referred to as area-to-point (ATP) or area-to-area (ATA) kriging (Gotway and Young 2002; Kyriakidis 2004; Ge et al. 2019; Hu and Huang 2020). This broader view has found application to a number of interpolation tasks in disease risk modeling (Kelsall and Wakefield 2002), soil nutrient mapping (Schirrmann et al. 2012), and remote sensing (Pardo-Iguzquiza et al. 2006). Our approach to spatial conditioning belongs to this family of methods, although with a greater emphasis on stochastic field generation than interpolation.

In the stochastic downscaling paradigm, the most closely related work is that of Allcroft and Glasbey (2003) and Gagnon et al. (2012) on precipitation downscaling. Both use a Gibbs sampling approach to approximate a conditional GMRF, alternating between a step in which pixels are iteratively resampled based on their neighbors, and a rescaling step which matches the field to observed spatial averages. Our approach differs from these in the explicit use of spatial conditioning, and the use of an adaptive covariance structure.

A challenge of downscaling to convective scales is that the spatial properties of the target field are themselves highly variable (Flack et al. 2016). This complicates the task of finding a suitable assumption about the small-scale properties, which is needed for stochastic downscaling. GRFs provide a plausible line of approach on this problem, as a small number of parameters can be used to express a wide range of spatial properties. This approach has not yet been fully utilized in previous work. A time-constant covariance is used for the model in Allcroft and Glasbey (2003); Gagnon et al. (2012) also uses a constant covariance structure, but with a variable standard deviation estimated from the large-scale convective available potential energy (CAPE) value.

We here take a different approach, using an adaptive covariance which is estimated from the observed low-resolution field using the maximum likelihood. The intention is to account for the widely varying spatial structures of convective fields, themselves driven by changes in the prevailing synoptic situation.

3. Problem setup

In this work, we consider downscaling from synthetically coarse-grained weather fields, produced by block averaging high-resolution fields (Fig. 1). This approach is in line with the bulk mass flux view of convective parameterization, in which subgrid-scale clouds are treated as an ensemble (Arakawa and Schubert 1974; Gregory and Rowntree 1990). From the coarse-grained field, our aim is then to generate plausible reconstructions of the original atmospheric conditions. The coarse-graining can be performed with different block sizes to produce fields of different resolutions.

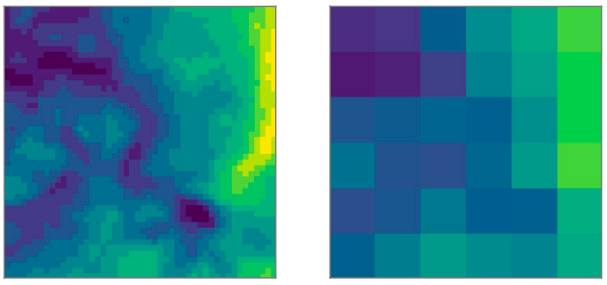


FIG. 1. (left) Original wet bulb potential temperature field together with (right) synthetic synoptic scale field formed by block averaging with block size 8×8.

Our high-resolution data is taken from the Met Office operational 2016 MOGREPS-U.K. model, which has a resolution of 2.2 km and so is convection permitting (Golding et al. 2016). We then coarse-grain this to two different extents: a synoptic scale of ~20 km (equivalent to an 8×8 coarsening) and a mesoscale of ~10 km (equivalent to a 4×4 coarsening).

Our downscaling task is, given a coarse-grained field, to produce samples which match the properties of the original field. In particular, we will compare the following metrics: the mean squared error (MSE), the power-spectral density (PSD), the continuous ranked probability score (CRPS), and a new metric called the neighborhood Wasserstein score (NWass) (see section 5). Each metric targets a different property of the sample fields, and whether or not it matches the target field. The point-by-point similarity is measured by MSE, the spectral properties such as the overall roughness or smoothness by PSD, the similarity of values over larger areas are compared by NWass, and the distributional properties are measured by CRPS.

4. Method

In this paper, we adopt a Gaussian random field approach. Our method has two main components: a linear operation which defines the distribution of the target field conditional on the coarse-grained field, given an estimated covariance of the target field; and an estimation of the covariance of the target field from the coarse-grained field by maximum likelihood. Taken together, these make it possible to construct sample outputs which match the properties of the original field, given a coarse-grained version of the field.

Before describing our model in detail in section 4b, we briefly introduce Gaussian random fields and their parameterization using kernels.

a. Gaussian random fields

A GRF is formally defined to be a collection of random variables indexed by one or more spatial variables, such that any finite subcollection has a multivariate Gaussian distribution

(Rasmussen and Williams 2006). In our context, these variables are our target grid point values, so our geospatial fields are represented by high-dimensional Gaussian distributions with each grid point represented by a different orthogonal dimension. This assumes a Gaussian distribution of values which is approximately true for the variables we are considering.[1]

GRFs have several advantages as a machine learning model, being intrinsically probabilistic and able to flexibly incorporate expert knowledge through the form of the mean and covariance (Rasmussen and Williams 2006). In addition, GRFs have convenient mathematical properties which make it possible to evaluate many quantities of interest analytically. Conditioning the distribution on data and drawing samples are important examples of this, both of which reduce to simple matrix algebra in the Gaussian case. We describe how our model makes use of these properties in section 4b(1).

A disadvantage of GRFs is that they can be computationally demanding for large datasets, scaling cubically with the number of sample points. Computational efficiency is an important consideration for statistical downscaling models, but this limitation does not excessively impact us in the present use-case.

x_{11}	x_{21}	x_{31}
x_{12}	x_{22}	x_{32}
x_{13}	x_{23}	x_{33}

FIG. 2. Grid on which we may define a Gaussian random field.

As an example of a GRF, consider the finite case of a 3×3 spatial grid X shown in Fig. 2. Then a Gaussian field G defined over X is a nine-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ given by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{33} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \operatorname{cov}(x_{11}, x_{11}) & \operatorname{cov}(x_{12}, x_{11}) & \cdots & \operatorname{cov}(x_{33}, x_{11}) \\ \operatorname{cov}(x_{11}, x_{12}) & \operatorname{cov}(x_{12}, x_{12}) & \cdots & \operatorname{cov}(x_{33}, x_{12}) \\ \vdots & & \ddots & \\ \operatorname{cov}(x_{11}, x_{33}) & \operatorname{cov}(x_{12}, x_{33}) & \cdots & \operatorname{cov}(x_{33}, x_{33}) \end{bmatrix}.$$

The mean can be any real-valued nine-dimensional vector. Theoretically, the covariance can be any valid 9×9 covariance matrix; that is, it must only satisfy the conditions of being symmetric and positive semidefinite. However, in standard practice Gaussian random fields are not arbitrary Gaussian distributions; they have spatial structure. Figure 3 shows how this spatial structure can be encoded in the covariance matrix. For a single dimension, the structure manifests as a concentration of positive covariance near the diagonal, indicating that nearby points should be closely related. The structure for two dimensions is harder to interpret, due to the two dimensions being flattened into a single index, but the interpretation is similar: any point is most closely related to its close neighbors in two dimensions.

The covariance is generally taken to have one of a number of standard forms known as kernel functions (or covariance functions, or covariance kernels), which are known to form valid covariance matrices. Kernel functions are discussed in detail in Rasmussen and Williams (2006). Using kernels to describe covariance matrices dramatically simplifies covariance estimation. Instead of having to estimate N^4 parameters for a two-dimensional $N \times N$ space, it is only necessary to learn the kernel parameters, of which there are usually just one or two. We make use of this convenient estimation property in section 4b(2). The effect of kernel parameters is illustrated in Fig. 3, which shows the effect of using different length scale parameters.

Kernels are often selected to encode properties which are believed to hold for the system under consideration. The standard assumptions are:

1) Locality: Covariance decreases with distance, which that is nearby points are more closely related than distant points.
2) Isotropy: The covariance of two points only depends on the distance between them, not on the direction.
3) Spatial stationarity: The covariance of two points depends only on the relative distance between them, not on the absolute location of either point.

These assumptions apply to all spatial models. If we are considering spatiotemporal models, we can add a fourth assumption:

4) Temporal stationarity: The covariance between two points does not change with time.

While these assumptions considerably simplify the analysis, their validity in atmospheric science is far from guaranteed. For instance, global teleconnections would break the locality assumption, continental versus marine air masses would break the isotropy assumption, orographic forcing would break the spatial stationarity

[1] It should be noted that many variables of importance in atmospheric science are not close to Gaussian distributed, notably moist variables such as precipitation and cloud coverage. Applying this method to such variables will likely result in an unrealistic distribution. If the deviation from Gaussianity is relatively minor at small scales, this may still be an acceptable approximation. If the deviation is more significant, it may be possible to apply this method by either normalizing the data using a Box–Cox transformation (Box and Cox 1964) or by viewing the field as a transformation of a latent Gaussian process (Kleiber et al. 2012).

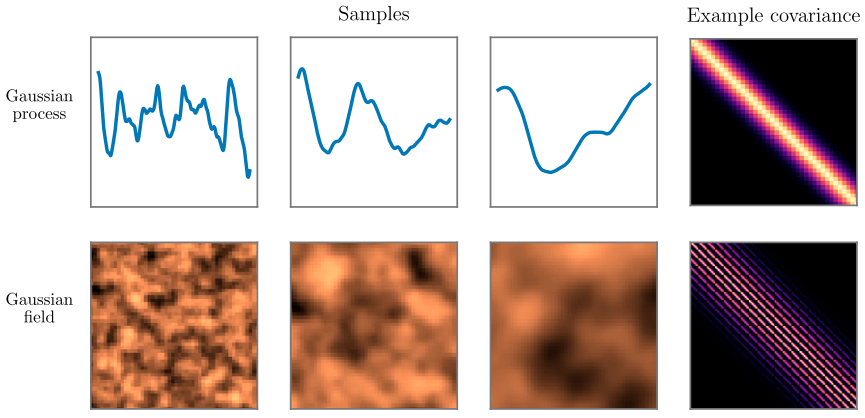


FIG. 3. Examples of 1D Gaussian processes and 2D Gaussian random fields. The rightmost column shows the structure of the covariance matrix for each dimension. Other columns show samples drawn using different kernel length scale parameters.

assumption, and seasonality would violate the temporal stationarity assumption. Fronts would be another example, as they can introduce sharp discontinuities which would be excessively smoothed over by standard kernels. However, these assumptions can often be lessened or removed by using more flexible models for the covariance, at the cost of increasing complexity.

b. *An adaptive GRF model for downscaling*

The challenge of downscaling to convective scales is that the covariance structure is highly variable. Motivated by this problem, in this paper we use a time-varying covariance function which can adapt to these changing conditions.

By fitting to each time step separately, we reject assumption four, that is we allow relationships between points to change with time. In contrast, since we only target a small area in the present work, the effect of different regimes should be minimal within a given time step. The structure of the downscaling problem itself works to justify the assumption of locality, which need only be assumed conditional on the coarse-grained field.

The setup of our model is as follows. To capture the dependence of the covariance structure on weather conditions, a separate covariance model will be required for each time step, estimated from low-resolution data. We can then condition on the low-resolution field to give our final model. The general framework is illustrated in Fig. 4.

This approach gives us the advantages of GRFs discussed previously but additionally should allow us to capture time varying properties of changing weather regimes. Note as well the absence of a training stage—the length scale, variance, and covariance of the high-resolution fields are estimated from that of the low-resolution field, allowing us to generate high-resolution fields.

1) CONDITIONING AND SAMPLING

Conditioning a Gaussian random field on a subset of its values, or on a linear combination, reduces to closed-form matrix algebra by means of the relations:

$$\boldsymbol{\mu}_{t|o} = \boldsymbol{\mu}_t + \mathbf{C}_{t,o}^{\mathrm{T}}\mathbf{C}_o^{-1}(\mathrm{obs} - \boldsymbol{\mu}_o), \tag{1}$$

$$\mathbf{C}_{t|o} = \mathbf{C}_t - (\mathbf{C}_{t,o}^{\mathrm{T}}\mathbf{C}_o^{-1}\mathbf{C}_{t,o}), \tag{2}$$

where a subscript o denotes observed variables (variables on which we condition), and a subscript t denotes target variables (for which we obtain a conditional distribution). A proof of these relations is given in appendix A. Here, $\mathbf{C}_{t|o}$ is the covariance of the target variables conditional on the observations, $\mathbf{C}_t$ is the unconditioned covariance of the target field, $\mathbf{C}_o$ is the covariance of the observed variables, and $\mathbf{C}_{t,o}$ is the joint covariance of the target and observed fields. Likewise, $\boldsymbol{\mu}_{t|o}$ is the conditional mean of the target field, $\boldsymbol{\mu}_t$ is the unconditioned mean, $\boldsymbol{\mu}_o$ is the expected value of the observed variables, and obs is the true observation.

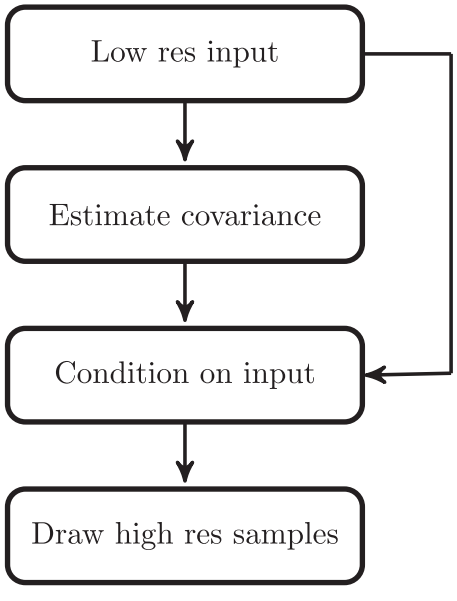


FIG. 4. General framework for adaptive probabilistic downscaling. Information from the low-resolution input field is used to estimate the covariance; the estimated covariance and input are then combined to obtain a conditional distribution. This is carried out independently for each time step, eliminating the assumption of temporal stationarity.

We can apply this to our downscaling problem by noting that our coarse-graining operation can be represented by a matrix multiplication. For a coarse-graining matrix $\mathbf{A}$, we have the linear constraint $\mathbf{Ax} = \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the coarse-grained field. Following Rue and Held (2005), the joint distribution of $\mathbf{x}$ and $\mathbf{Ax}$ is given by

$$\mathbb{E}\begin{pmatrix}\mathbf{x}\\ \mathbf{Ax}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_t\\ \boldsymbol{\mu}_o\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_t\\ \mathbf{A}\boldsymbol{\mu}_t\end{pmatrix},$$

$$\operatorname{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{Ax}\end{pmatrix} = \begin{pmatrix}\mathbf{C}_t & \mathbf{C}_{o,t}\\ \mathbf{C}_{t,o} & \mathbf{C}_o\end{pmatrix} = \begin{pmatrix}\mathbf{C}_t & \mathbf{C}_t\mathbf{A}^{\mathrm{T}}\\ \mathbf{AC}_t & \mathbf{AC}_t\mathbf{A}^{\mathrm{T}}\end{pmatrix}. \quad (3)$$

Using the relations of Eq. (2), the distribution $p(\mathbf{x}|\mathbf{Ax} = \bar{\mathbf{x}})$ can then be found by computing

$$\boldsymbol{\mu}_{t|o} = \boldsymbol{\mu}_t - \mathbf{C}_t\mathbf{A}^{\mathrm{T}}(\mathbf{AC}_t\mathbf{A}^{\mathrm{T}})^{-1}(\bar{\mathbf{x}} - \mathbf{A}\boldsymbol{\mu}_t), \quad (4)$$

$$\mathbf{C}_{t|o} = \mathbf{C}_t - \mathbf{C}_t\mathbf{A}^{\mathrm{T}}(\mathbf{AC}_t\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{AC}_t. \quad (5)$$

Proof of these identities is given in appendix A. Now, $\boldsymbol{\mu}_{t|o}$ and $\mathbf{C}_{t|o}$ are the mean and covariance of the conditional GRF given by conditioning on a low-resolution "observation." This forms the basis for our probabilistic downscaling model. This conditioning relies on having an estimate of the covariance of the high-resolution field, given by $\mathbf{C}_t$. Estimating $\mathbf{C}_t$ is the subject of the following section.

Having obtained our conditional distribution, we can draw samples using the standard method for sampling a Gaussian distribution: if $\mathbf{w}$ is a vector containing white noise (Gaussian with zero mean and unit variance) then

$$\mathbf{s} = \boldsymbol{\mu}_{t|o} + \mathbf{C}_{t|o}^{1/2}\mathbf{w} \quad (6)$$

is a sample from our conditional distribution, where for a matrix $\mathbf{M}$ the symbol $\mathbf{M}^{1/2}$ denotes a matrix with the property that $\mathbf{M} = \mathbf{M}^{1/2}\mathbf{M}^{(1/2)\mathrm{T}}$.

Example of spatial conditioning

Since our approach to conditioning is different to the standard approach used in the Gaussian process and GRF literature we will demonstrate using a one-dimensional example.

Consider a sequence of observations which we will use to condition a Gaussian process model (Fig. 5). Depending on what we know about the system we are trying to model, we might interpret this sequence of observations as point values or as spatial averages.

The effects of this choice of interpretation are shown in Fig. 6. The prior mean, variance, and covariance are the same in each case (specifically we use zero mean, unit variance, and squared exponential covariance with length scale ten). However, the two conditional distributions have quite different properties. Conditioning on point values leads to a "pinched" marginal variance with much higher variance at a distance from the observations. By contrast, conditioning on spatial averages leads to a more uniform level of variance.

While the sample fields in Fig. 6c are constrained to pass through the observed points, the constraint on the sample fields shown in Fig. 6d is less easy to see. However, Fig. 7 makes this constraint easy to see. It shows that the spatial averages of the sample fields conditioned on spatial averages are all in agreement with the original coarse-grained observations, while those conditioned on point values show substantial variability. This property is enforced directly by conditioning on these spatial averages, as described in Eq. (5). Instead of conditioning on a specific point value, we have effectively conditioned on a low-resolution version of the field—just as required for our downscaling task.

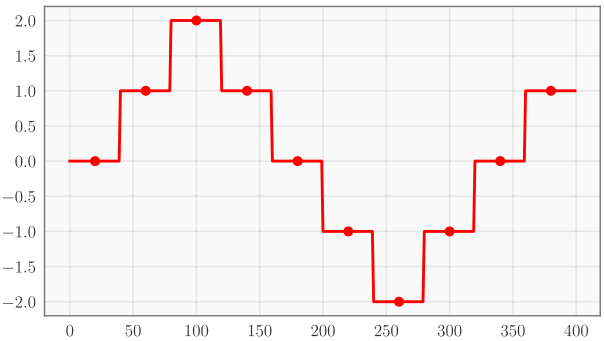


FIG. 5. Observed values. Dots indicate an interpretation as point observations, and horizontal lines indicate an interpretation as spatial averages.

As a further illustration, notice how the properties of the two distributions are encoded in the posterior covariance matrices shown in Fig. 8, particularly in the distribution of their negative values. Conditioning on a point observation introduces a localized positive/negative "dipole" forcing smoothness of the samples through the observation. On the other hand, conditioning on a spatial average induces a wider-scale negative correlation between the points in the area being averaged. This is in contrast to the prior covariance, which consists of only positive values which peak along the diagonal (not shown).

All the essential properties of this example hold true in two or more dimensions; we have chosen here to illustrate the idea with one dimension for ease of interpretation. The rest of the paper is focused primarily on the two-dimensional case.

2) COVARIANCE ESTIMATION

To perform the conditioning step of Eq. (5) we need to have an estimate of the high-resolution covariance $\mathbf{C}$. As discussed above, since this estimated covariance is what controls the behavior of our model at small scales, it needs to be estimated separately for each time step. Fortunately, by making use of a parameterized covariance kernel we can reduce this to a low-dimensional optimization problem.

We have chosen to consider the Matern covariance due to its flexibility in modeling smoothness properties (Stein 1999). We treat the shape parameter ν as a hyperparameter, leaving us to fit two parameters for each input: the length scale ℓ and the variance σ^2.

The length scale ℓ we estimate using maximum likelihood estimation, where the likelihood of a parameter is defined to be

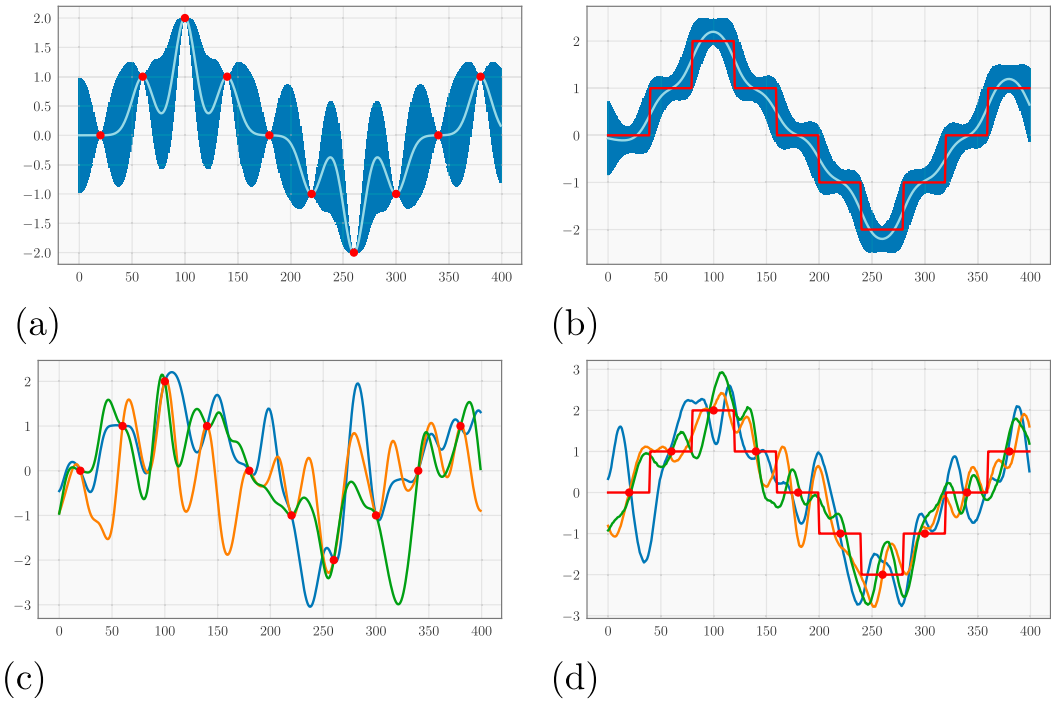


FIG. 6. The distribution obtained by conditioning on (a) point values and (b) spatial averages. (c),(d) Sample fields are shown beneath the corresponding distributions. The distributional means are shown in pale blue, and the marginal variances are in darker blue.

the joint density of the observed data as a function of the parameter only, with the data held fixed. The density for a GRF is just the standard Gaussian density, where we take the mean to be zero and the covariance $\mathbf{K}(\theta)$ to be determined by its parameters θ:

$$p(\mathbf{x}_i|\theta) = \frac{1}{\sqrt{2\pi|\mathbf{K}|}} \exp\left(-\frac{1}{2}\mathbf{x}_i^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{x}_i\right), \tag{7}$$

thus the log likelihood is given by

$$\log p(\mathbf{x}|\theta|) = -\frac{1}{2}\left[n\log(2\pi) + n\log|\mathbf{K}| + \sum_{i=1}^{n}\mathbf{x}_i^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{x}_i\right]. \tag{8}$$

By maximizing the log likelihood using standard optimization methods, we can find the kernel parameter values which best fit the observed data. Thus we approximate the true target covariance $\mathbf{C}_t$ of Eq. (5) by a parameterized version $\mathbf{K}(\theta)$.

In the conditional downscaling task we do not observe samples of the target field directly, but only after convolution with a filter matrix $\mathbf{A}$. Therefore, the relevant covariance matrix is not the one computed directly from the kernel $\mathbf{K}(\theta)$, but the filtered covariance $\mathbf{AKA}^{\mathrm{T}}$. The log likelihood is then given by

$$\log p(\mathbf{x}|\theta) = -\frac{1}{2}\left[n\log(2\pi) + n\log|\mathbf{K}||\mathbf{A}|^2 + \sum_{i=1}^{n}\mathbf{x}_i^{\mathrm{T}}(\mathbf{AKA}^{\mathrm{T}})^{-1}\mathbf{x}_i\right], \tag{9}$$

which gives an analogous optimization.

The optimization of the variance can be separated from the optimization of the other parameters. This is possible because the dependence of $\mathbf{K}$ on the variance σ is just $\mathbf{K}_\sigma = \sigma\mathbf{K}_1$. We then have $\mathbf{K}_\sigma^{-1} = (1/\sigma)\mathbf{K}_1^{-1}$ and $\partial\mathbf{K}_\sigma/\partial\theta_j = \sigma(\partial\mathbf{K}_1/\partial\theta_j)$. Any dependence on σ will therefore drop out of the gradient equation, so that the extremal values of any other parameters are independent of the variance. The variance parameter can then be optimized in a separate step.

Taken together, the methods described in this section allow us to estimate the full covariance matrix for a high-resolution field given a coarse-grained version of that field. We can then

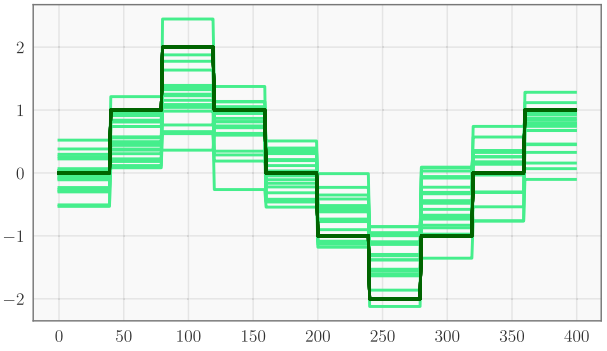


FIG. 7. Spatial averages of samples from the distribution conditioned on spatial averages (dark green) and point observations (light green).

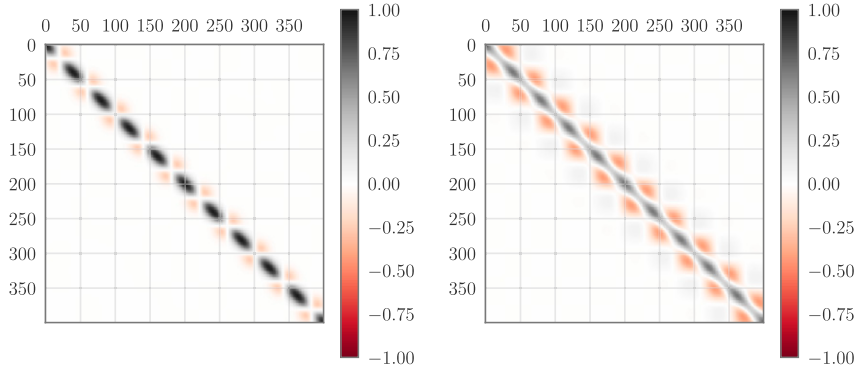


FIG. 8. Posterior covariance matrices after conditioning on (a) point observations and (b) spatial averages.

apply the conditioning and sampling steps of section 4b(1) to generate plausible reconstructions.

5. Verification

The question of how to best evaluate probabilistic downscaling models is a challenging one, especially when the target structure is chaotic. This reflects a known difficulty in verifying high-resolution physics-based models; although there are sound reasons to believe that increased resolution improves the realism of the model, this improvement is often not reflected by standard verification scores (Sansom 2015). In particular, point-based verification metrics such as mean-squared error (MSE) have the undesirable property that insignificant displacement of features is heavily penalized due to the double-penalty problem: the displacement is treated as a missing feature, and a separate anomalous feature. As has been observed in the context of machine learning, such scores tend to reward blurry predictions and penalize specificity (Mathieu et al. 2015).

To address these difficulties, we consider two additional deterministic verification metrics besides the standard MSE. These are the spectral and neighborhood Wasserstein scores, which are both spatial verification metrics applied to samples drawn from the probabilistic model. We also consider the distribution properties of stochastic methods by means of the continuous ranked probability score (CRPS).

a. Spectral score

The purpose of the spectral score is to evaluate the quality of the generated fields in frequency space as opposed to physical space. The generated and target fields are first transformed into Fourier space using the two-dimensional Fourier transform. The one-dimensional power spectral density (PSD) is then given by the squared amplitude component integrated over circles of given radius. For a given radius φ we have

$$\mathrm{PSD}_{\varphi}(\mathbf{x}) = \int_{\sqrt{\omega^2+\sigma^2}=\varphi} \hat{\mathbf{x}}^2(\omega,\sigma)\, d\tan^{-1}\left(\frac{\sigma}{\omega}\right). \tag{10}$$

where the integral is taken over the angle $\theta = \tan^{-1}(\sigma/\omega)$ and $\hat{\mathbf{x}}$ denotes the two-dimensional Fourier transform of $\mathbf{x}$. We will use the notation $\mathrm{PSD}(\mathbf{x})$ to denote the power spectral density as a function of φ, so that $\mathrm{PSD}(\mathbf{x})(\varphi) = \mathrm{PSD}_{\varphi}(\mathbf{x})$.

The generated and target fields can then be compared using any of the standard histogram distance measures. As in Martinez et al. (2016), we here use the 1-Wasserstein distance between their PSDs. The p-Wasserstein distance is defined to be

$$W_p(\mu,\nu) = \left[\inf_{\pi\in\Pi(\mu,\nu)} \int_{M\times M} d(x,y)^p\, d\pi(x,y)\right]^{1/p}, \tag{11}$$

where μ and ν are probability measures on M, and $\Pi(\mu, \nu)$ is the collection of all measures on $M \times M$ with marginals μ and ν (Villani 2008). Where $p = 1$, the Wasserstein distance is the same thing as the Earth mover's distance, or optimal transport, which describes the cost of transforming one distribution into another (Arjovsky et al. 2017). This is particularly appropriate for comparing power spectra, as the cost of small frequency displacements is correspondingly small. The spectral score is then given by

$$W_1[\mathrm{PSD}(\mathbf{x}), \mathrm{PSD}(\mathbf{x}^*)], \tag{12}$$

where $\mathbf{x}$ and $\mathbf{x}^*$ are the target and generated fields, respectively, and we have abused notation to write $\mathrm{PSD}(\mathbf{x})$ to denote the image of that function.

b. Neighborhood Wasserstein score

Point-based verification methods have the undesirable property that slight displacement of features is heavily penalized due to the double-penalty problem: the displacement is treated as a missing feature, and a separate anomalous feature. In contrast, spatial verification methods are designed to avoid, or at least characterize, this effect. The main subcategories of spatial verification include scale-separation approaches (e.g., Casati et al. (2004)), feature-based methods (e.g., Davis et al. (2006)), and neighborhood methods. Neighborhood methods are especially well-suited to our

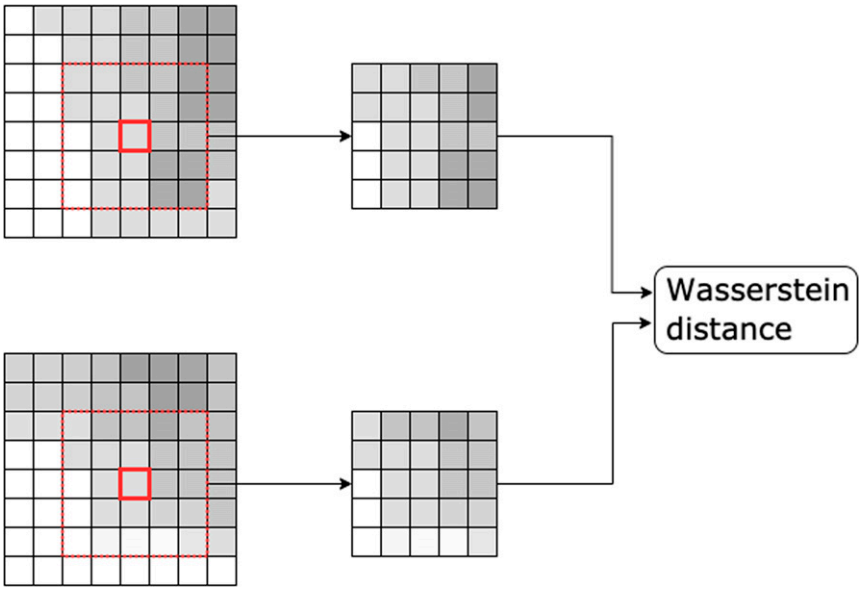


FIG. 9. Schematic showing calculation of the neighborhood Wasserstein score.

purposes. Unlike feature-based methods they can be easily applied to continuous fields (e.g., Rezacova et al. (2007)). Scale-separation methods are designed to assess the scale-dependence of errors, but do not in themselves address the double-penalty problem as do neighborhood methods (Yu et al. 2020).

The assumption underlying neighborhood verification methods is that a slightly displaced forecast may still be more useful than one which fails to capture the relevant features (Ebert 2008). Instead of comparing fields point by point, these methods consider a neighborhood surrounding each grid cell and some measure of similarity of the real and predicted neighborhoods. The canonical example is the fractions skill score (Roberts and Lean 2008), which compares the number of grid cells exceeding a given threshold within a neighborhood; other variants are reviewed in Ebert (2008).

When considering a continuous valued field for which there is no obvious threshold of special interest, it is more natural to compare the full distribution of values within a neighborhood than an exceedance count. One possible score of this type was proposed by Rezacova et al. (2007), who use the mean squared distance of vectors containing the ordered values from the two fields. However, there are a number of possible measures of the distance between distributions; several possibilities are discussed in Bellemare et al. (2018), Ramdas et al. (2015) and Feydy et al. (2019). We here propose a spatial score which uses the 1-Wasserstein distance, as described in the previous section [Eq. (12)], rewarding fields which have similar distributions over small neighborhoods. We will call this the neighborhood Wasserstein score.

Note that we compute the 1-Wasserstein distance between the histogram of values within the neighborhoods. Thus, the calculation is invariant to the spatial distribution of values within a particular neighborhood. Computationally, calculating the 1-Wasserstein distance is equivalent to sorting the flattened list of values within the neighborhoods to be compared, and summing the unsigned differences of the sorted lists. A proof of this equivalence is given in appendix B.

To compute this score, it is first necessary to choose a neighborhood size, which sets the side length of a square centered at the target point. For each point, a neighborhood is extracted from the two fields being compared. The score for that point is the Wasserstein distance between the observed and target neighborhoods. The overall score is then the mean of the scores for each point:

$$\mathrm{NWass}_k(X,Y) = \frac{1}{N_k}\sum_{n=1}^{N_k} W_1(X_n^k, Y_n^k),$$

where there are N_k neighborhoods of size k, and X_n^k is the nth neighborhood of size k in X. A schematic illustration is shown in Fig. 9.

For any neighborhood score, it is important to ensure that the choice of neighborhood size is appropriate to the application. In our experiment, we chose a neighborhood size of 4×4 grid cells, or approximately 10 km. This is slightly below the 15 km used in Golding et al. (2016), which justifies that our choice of neighborhood size is not unreasonably large.

c. *Continuous ranked probability score*

While the spectral and neighborhood Wasserstein scores are able to compare the spatial structure of the forecast and target fields, they are still fundamentally deterministic scores and do not take the full distribution of stochastic models into account. We therefore need to complement these scores with a probabilistic metric to get a full picture of the performance.

For this purpose, we have selected the continuous ranked probability score (CRPS). For a probabilistic forecast, the

TABLE 1. Correlation coefficients for wet bulb potential temperature at 850 hPa and low-type cloud coverage. Low-pass filter is a Gaussian filter with a sigma coefficient of 128, which is subtracted from the original field to give the high-pass filter.

	Full	Low pass	High pass
corr(wbpt, cloud)	0.32	0.27	0.38

CRPS is defined as the integrated squared difference between the forecast CDF F and the observed CDF:

$$\mathrm{CRPS}(F,x)\int_{-\infty}^{\infty}[F(y)-1(y\geq x)]^2\,dy,$$

where $1\,(y \geq x)$ is an indicator function for the observed value x exceeding the threshold y (Gneiting et al. 2007).

6. Application to wet bulb potential temperature

In this section, we test the performance of the conditional downscaling model on fields from the MOGREPS-U.K. model produced by the Met Office (Golding et al. 2016; Hagelin et al. 2017). We have chosen to work with wet bulb potential temperature at 850 hPa for two reasons. First, it has substantial small-scale chaotic structure, making it a meaningful target for probabilistic downscaling. Indeed, Table 1 shows that WBPT is correlated with low-type cloud coverage, particularly at smaller scales. However, unlike cloud coverage its values are approximately Gaussian distributed (Fig. 10), making it a more suitable target for our model.

Our dataset consists of wet bulb potential temperature data at 850 hPa from the operational 2016 MOGREPS-U.K. model. Dates used are summarized in Table 2. All data is taken from the unperturbed zeroth ensemble member at the 3-h time step of the 1500 UTC model run.

To test our model we have chosen to focus on areas of low orography, as this is where stochastic variability is more likely to dominate over orographically driven variability at small scales. We consider three locations for our dataset, one in southeast England (51.3°–52.2° latitude, from −1.0° to 0.6° longitude), one in the Midlands (51.5°–52.4° latitude, from −2.0° to −0.4° longitude), and one in Ireland (52.2°–53.2° latitude, from −8.4° to −6.8° longitude).

a. Benchmarks

To gauge the effectiveness of using an adaptive covariance, we contrast two versions of our GRF model: one baseline version using a stationary covariance estimated from the full dataset, and one using a seperate covariance estimate for each individual field. We refer to these models as GRF-S (stationary) and GRF-T (time-adaptive). We also consider a third version, identical to GRF-T except that the input variables are treated as point values instead of spatial averages (GRF-T pt); this is to test our assertion that spatial conditioning is more appropriate for our downscaling problem.

We compare our model against five benchmarks. Two are deterministic, and utilize only the low-resolution input. These are the synthetic low-resolution field upsampled to match the resolution of the target field (lres); and the bicubic interpolation of the low-resolution field (bicub). The third is a data-driven deterministic model, namely, an ElasticNet regression. Finally, we compare our approach to two stochastic data-driven models based on the cascade decomposition approach.

1) ELASTICNET MODEL

For our deterministic data-driven benchmark, we have chosen to use a linear regression mapping the low-resolution covariates to the high-resolution target variables. Specifically, we consider a mapping $Y = f(X)$ from the full low-resolution input X to the full high-resolution output Y. Since we are dealing with a large number of dimensions and a relatively modest sized training set, it is important to assess the need for model regularization. There are three widely adopted methods for regularizing linear models: the Lasso model uses an L1 penalty to encourage sparsity in the model weights; Ridge regression uses an L2 penalty to encourage smaller weights; and ElasticNet encompasses both cases by incorporating both an L1 and an L2 penalty (Friedman et al. 2001). We have chosen the ElasticNet model as it is the most general.

We tuned the model hyperparameters using fivefold cross validation with a maximum of 10 000 iterations. Following the convention used in the SciKit-Learn library (Pedregosa et al. 2011), we define L1-ratio to be the relative weighting of the

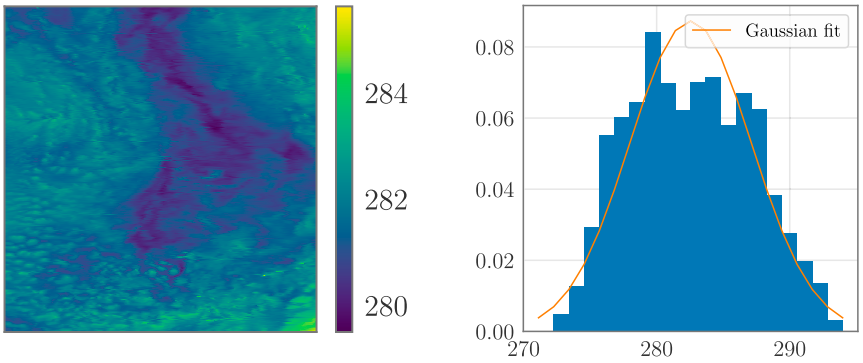


FIG. 10. (left) An example WBPT field. (right) Histogram of WBPT values for the period under consideration with best fit Gaussian PDF shown in orange.

TABLE 2. Data used in development and evaluation sets. The dev set is used to fit the shape hyperparameter for the Matern kernel, and the dev2 set is used to for training and hyperparameter tuning benchmark models.

	Years	Months	Days	Hours	Lead times	Members
dev	2016	1–12	1, 11, 21	15	3	0
dev2	2016	1–12	[1–29]\{5, 15, 25}	15	3	0
eval	2016	1–12	5, 15, 25	15	3	0

L1 to L2 penalty, and alpha to be a scalar factor multiplying both penalties. Overall, the combination of alpha = 0.01 and L1-ratio = 1 performed best for both the mesoscale and synoptic-scale experiments. This means that in both cases the model performed best when only the L1 penalty was applied, corresponding to a Lasso regression.

Having selected the optimal hyperparameters, we fit our two models on the full training dataset dev2 using scikit-learn, using a maximum of 100 000 iterations.

2) FRACTAL CASCADE MODEL

An alternative approach to stochastic downscaling is the fractal cascade approach. This was initiated by the work of Lovejoy and Mandelbrot (1985) in the context of modeling rainfall, and is based on the assumption of a simple power scaling law connecting the larger and smaller scales. By using this law to extend the Fourier spectrum of the input field, it is possible to generate fields having appropriate small-scale structure. In this work, we use the RainFARM (Rebora et al. 2006) algorithm as implemented in the PySTEPS (Pulkkinen et al. 2019) package, which is based on these principles.

To apply the RainFARM algorithm to our approximately Gaussian distributed data, we need to first transform the data to more closely match the distribution of rainfall. We do this by first standardising the input field by subtracting its mean and dividing by the standard deviation, and then exponentiate the standardized field:

$$X' = \exp\left[\frac{X - \text{mean}(X)}{\text{std}(X)}\right].$$

The RainFARM algorithm uses a parameter called alpha, defining the slope used to extend the power spectrum (specifically, it defines the slop of the log power spectrum). There is an option to determine this parameter directly from the input field; however, we found this method gave suboptimal results on our dataset. Instead, we fit the alpha parameter by optimizing results for the power spectral score over the full training set dev2. Normalized training set scores for integral values of alpha are shown in Fig. 11; higher values of alpha lead to improved MSE scores in general, while the PSD and neighborhood Wasserstein scores are minimal around alpha = 4. Optimizing the PSD score using Scikit-learn gave an optimal value for alpha of 3.54 over the southeast England region, 3.49 for the Midlands, and 3.50 for the Ireland region, in the synoptic experiment. For the mesoscale experiment, the corresponding values are 3.49, 3.53, and 3.51.

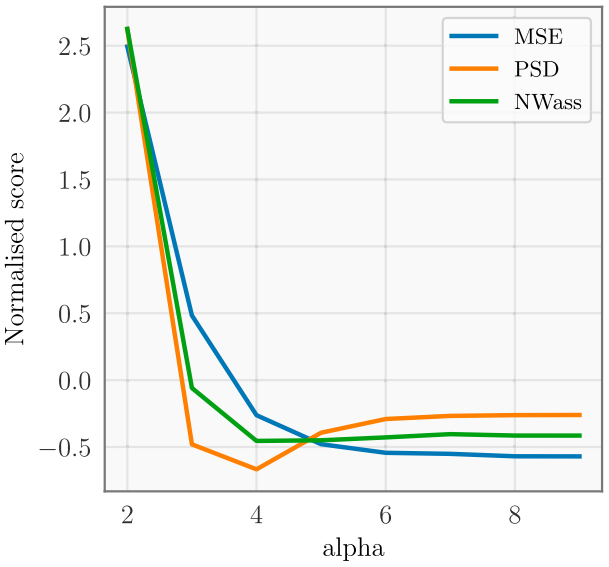


FIG. 11. Normalized training set scores for integral values of alpha for southeast England region.

3) MULTIPLICATIVE CASCADE MODEL

Another kind of cascade model, also developed for rainfall disaggregation, is known as a multiplicative cascade. Instead of working with the Fourier spectrum, this approach uses stochastic weights to redistribute aggregated rainfall values among the four component subgrid cells at the next resolution, repeating this process until the desired resolution is reached. The idea was originally explored by Perica and Foufoula-Georgiou (1996), who treated it as a form of fractal cascade by using a power scaling law to determine the parameters for the distributions generating the stochastic weights. However, multiplicative cascade models need not be fractal cascades. For our implementation, we follow the classical equal-volume area model described in Schleiss (2020), which uses an empirical approach to fit the disaggregation parameters.

Since our task requires three levels of disaggregation the model requires six parameters, corresponding to the horizontal and vertical variability at each level. We follow Schleiss (2020) in using a symmetric logit-normal distribution to model the weights, leaving the standard deviations to be estimated from the training data. In the original model, Schleiss (2020) allows the standard deviation parameters to depend on the rainfall intensity as a polynomial. We follow the same approach, using a polynomial regression to determine the standard deviations as a function of the aggregate field value at the source level. We allow a maximal polynomial degree of five, and use a hard cutoff at zero to prevent nonsensical negative standard deviation values.

b. Results

We compare these GRF models with our five benchmarks using the verification scores described in section 5 and the mean squared error. The results aggregated over the three regions are shown in Table 3.

TABLE 3. Results for aggregated over all three regions. Deterministic scores for GRF samples report an average of 20 samples. Neighborhood Wasserstein score is reported for a square neighborhood of size-length four. The best results for each score in each experiment are shown in bold.

Convolution	Model	MSE	PSD Wass	NWass (4)	CRPS
4×4	lres	0.017	0.72	0.030	—
	bicub	0.012	1.08	0.033	—
	elasticnet	0.012	0.97	0.034	—
	rainfarm	0.019	0.69	0.038	0.056
	cascade	0.035	1.68	0.036	0.066
	GRF-S samples	0.032	0.84	0.034	0.051
	GRF-T-pt samples	0.018	0.89	0.034	0.053
	GRF-T mean	**0.009**	1.07	**0.028**	—
	GRF-T samples	0.018	**0.68**	0.030	**0.046**
8×8	lres	0.040	1.06	0.050	—
	bicub	0.029	1.30	0.049	—
	elasticnet	0.030	1.14	0.049	—
	rainfarm	0.059	**0.79**	0.062	0.092
	cascade	0.078	2.07	0.051	0.102
	GRF-S samples	0.055	1.00	0.043	0.080
	GRF-T-pt samples	0.039	1.14	0.048	0.082
	GRF-T mean	**0.025**	1.32	0.039	—
	GRF-T samples	0.041	0.99	**0.038**	**0.076**

Considering first the CRPS score, our GRF-T model has consistently the best performance for both the mesoscale and synoptic experiments. The other GRF models also perform well, out-performing the cascade models; the multiplicative cascade has the weakest performance overall.

For the MSE score, the GRF-T mean scores consistently best, with bicubic interpolation and ElasticNet regression also performing well. As expected, the stochastic models have generally higher MSE scores than the deterministic models, particularly the multiplicative cascade model.

The comparison is less clear for the neighborhood Wasserstein and PSD scores, with the relative performance varying depending on the experiment scale. The RainFARM model performs well according to the PSD score, outperforming the other models in the synoptic experiment. The GRF-T model also performs well, as does the original low-resolution input. For the neighborhood Wasserstein score, the GRF-T model performs well, as does the GRF-T mean, with the mean performing better in the mesoscale experiment and the samples in the synoptic experiment.

A comparison of the model performance on both the neighborhood Wasserstein and PSD scores together gives a clearer picture (Figs. 12 and 13). The models are plotted against the two scores on perpendicular axes, so that models which perform better on both scores appear toward the bottom left corner. The GRF-T model is near to the bottom left in both the synoptic and mesoscale experiments. In the mesoscale case, the GRF-T model and GRF-T mean form a Pareto front, with the GRF-T

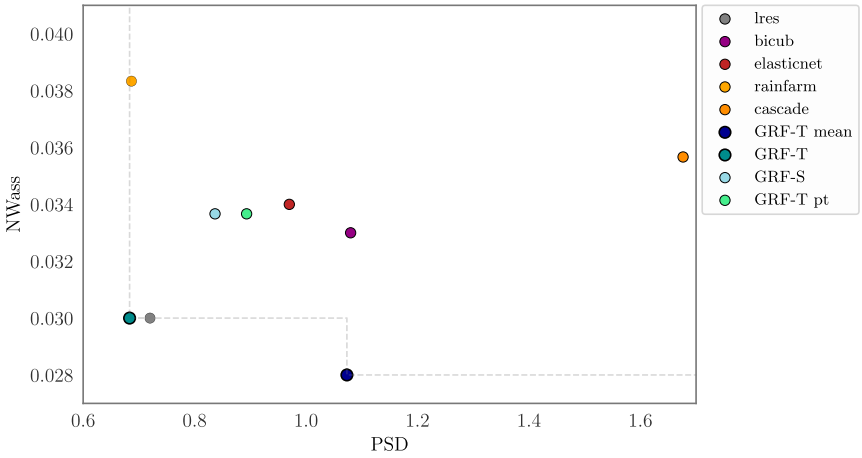


FIG. 12. Scatterplot of PSD and neighborhood Wasserstein scores for mesoscale experiment (aggregate over all locations). Dashed gray line denotes Pareto front; Pareto optimal points are drawn with bold outlines.

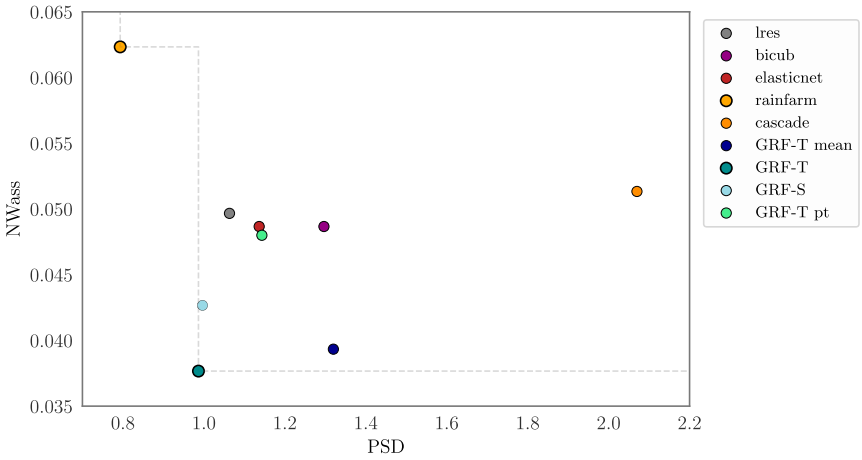


FIG. 13. Scatterplot of PSD and neighborhood Wasserstein scores for synoptic experiment (aggregate over all locations). Dashed gray line denotes Pareto front; Pareto optimal points are drawn with bold outlines.

mean scoring best for the neighborhood Wasserstein score but less well for the PSD score. On the other hand, for the synoptic experiment the GRF-T and RainFARM models form a Pareto front, with RainFARM scoring best for the PSD score but worst for the neighborhood Wasserstein score.

The disparity between the NWass and PSD scores is perhaps surprising, since both are methods for spatial verification and might be expected to favor the same models. In the case of RainFARM, this disparity may be due in part to its approach of randomizing the phase components of the Fourier spectrum to produce stochastic samples. Since it is the phase information which encodes visual structure (Taylor 2003) this phase randomization may have the effect of obliterating predictable structure in the generated fields. This would be penalized by location-based scores such as MSE and NWass, but not by the PSD score which is determined by the spectrum of the field as a whole.

Sample outputs for the models and benchmarks can be found in Figs. 14, 15, 16, 17, 18, and 19. The subfigures show the samples in which the GRF samples perform best, median and worst compared with bicubic interpolation with respect to the neighborhood Wasserstein score. In the best cases, the GRF-T model introduces variability which is missing from the bicubic interpolation, as well as better capturing the predictable component where it extends beyond the value range of the coarse-grained observation. In the worst cases, the GRF-T model can introduce spurious variability (e.g., Figure 17c).

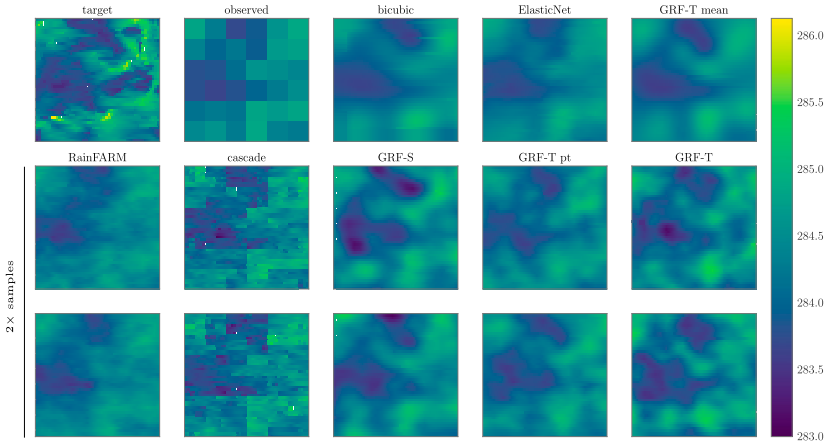
Overall, taking account of all the metrics, our GRF-T model performs very well compared to the benchmarks. The fact that the model scores best on the CRPS metric gives some assurance that the model is well calibrated in a probabilistic sense, and that the GRF-T mean predictions have the lowest MSE losses suggests that the model is able to make effective use of the information in the low-resolution fields to improve deterministic predictions. Figures 12 and 13 show that the stochastic predictions have good spatial characteristics, scoring well on both the PSD and NWass metrics. The benefit of these stochastic predictions is clear in the synoptic experiment; in the mesoscale case the benefit is more marginal, as most of the variability seems to be already captured in the low-resolution field.

To better understand the performance of the GRF-T model, it may be instructive to look at the optimal values for length scale and variance selected for each test sample. Taken together, these values fully determine the covariance matrix used to generate the stochastic samples. The selected values for the synoptic experiment are shown in Fig. 20, and for the mesoscale experiment in Fig. 21. There is a notable general tendency for the selected values to lie on a curve, with lower length scale values tending to correspond to higher variance values. In addition, the values selected in winter months are generally seen to have longer length scales; this is what we would expect if shorter length scales correspond to convective weather. While the broad structure is similar for both the synoptic and mesoscale models, there is a noticeable difference in the values selected for the length scale parameter, with those for the synoptic experiment being around twice as high as in the mesoscale experiment. Not too surprisingly, this suggests a limited ability of the tmodel to distinguish length scales which are substantially lower than the resolution of the input field. There may be scope to improve this further by incorporating different sources of information into the length scale estimation step.

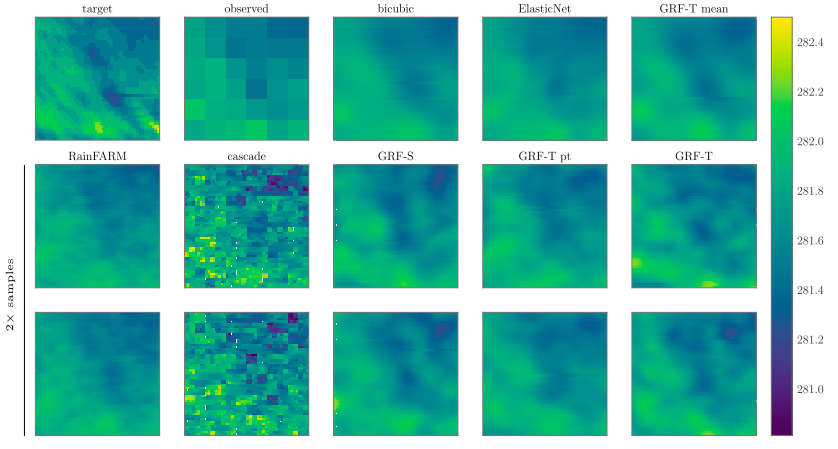
7. Conclusions

We have argued that downscaling to convective scales calls for a stochastic approach due to inherent chaotic variability at these scales. We have proposed a model for variables which are approximately Gaussian distributed, based on a spatial statistical method called Gaussian random fields. An advantage of our model is that it only requires a single input field; not being data driven, it does not require a training dataset, only the estimation of model parameters.

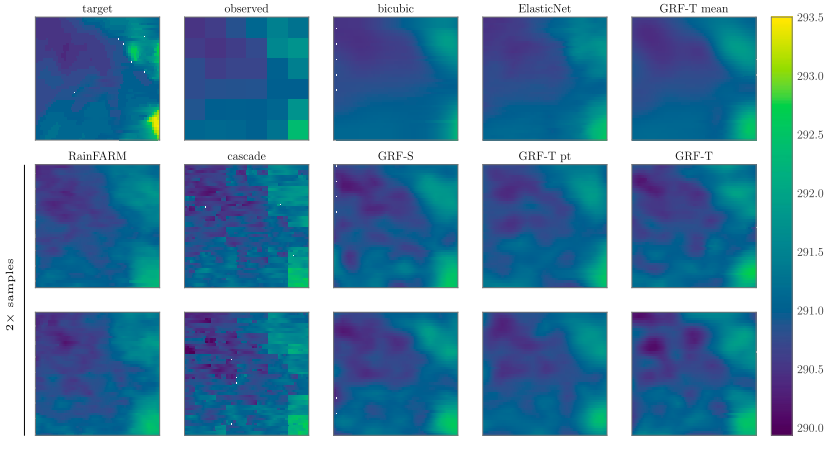
We have argued that special attention needs to be paid to verification of stochastic downscaling models, as standard



(a)

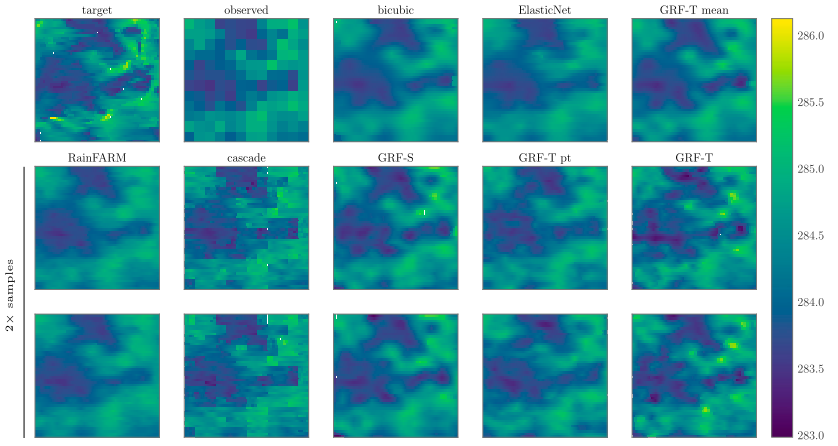


(b)

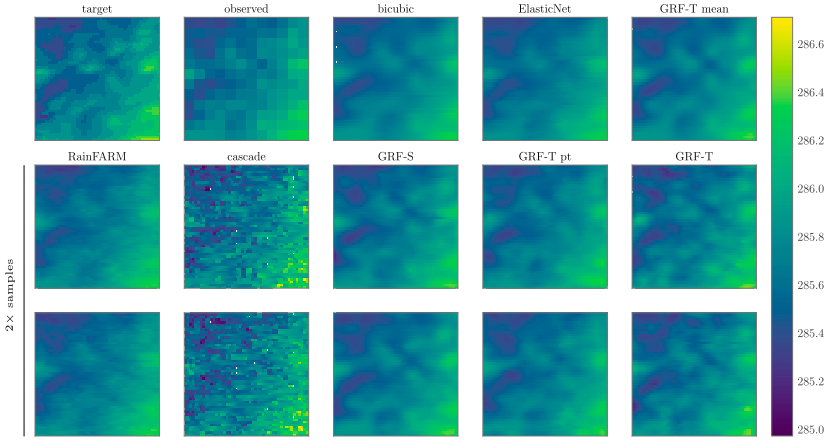


(c)

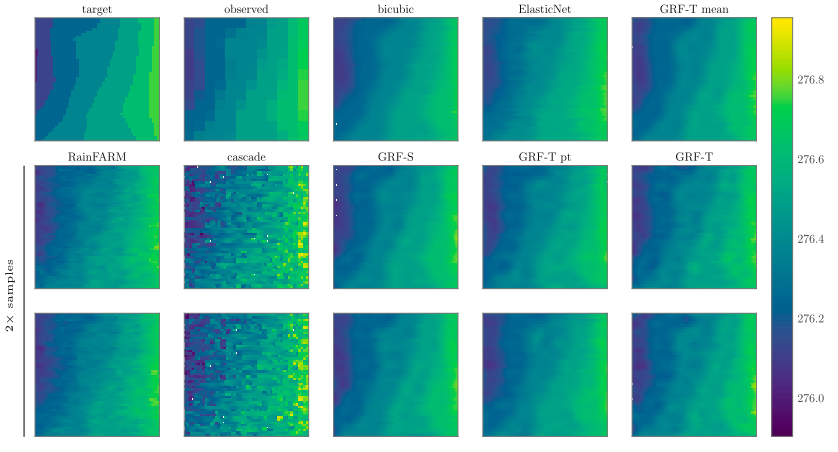
FIG. 14. Example outputs for the synoptic GRF-T model and benchmarks for southeast England region. (a) Best, (b) median, and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.



(a)

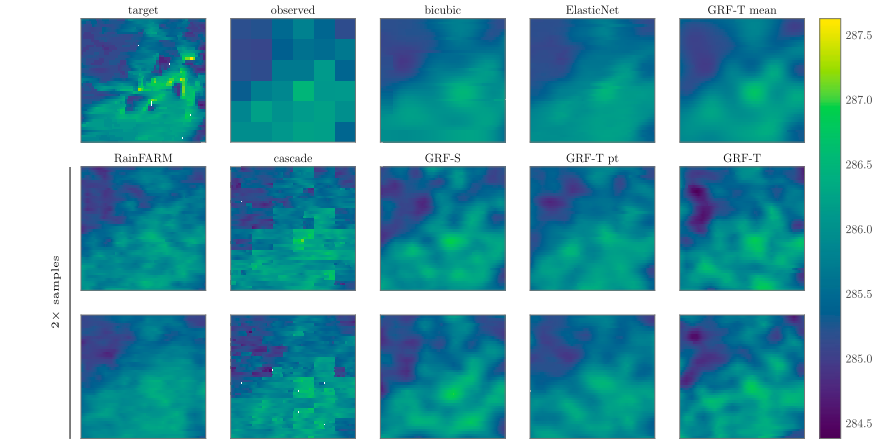


(b)

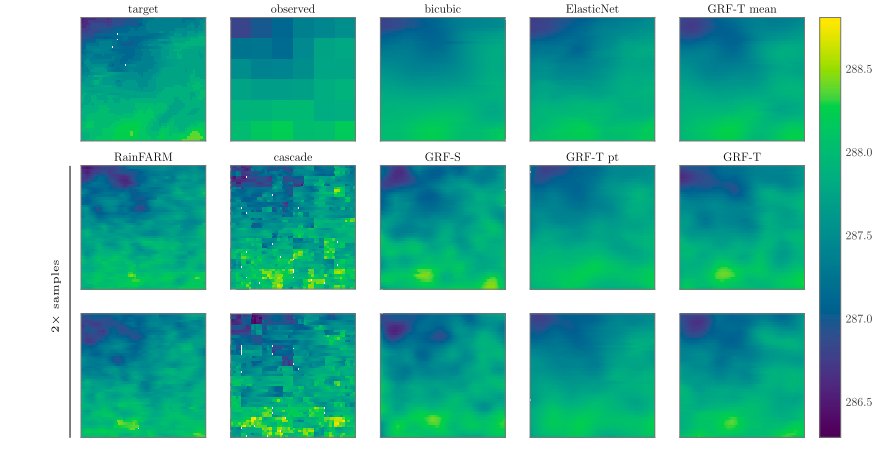


(c)

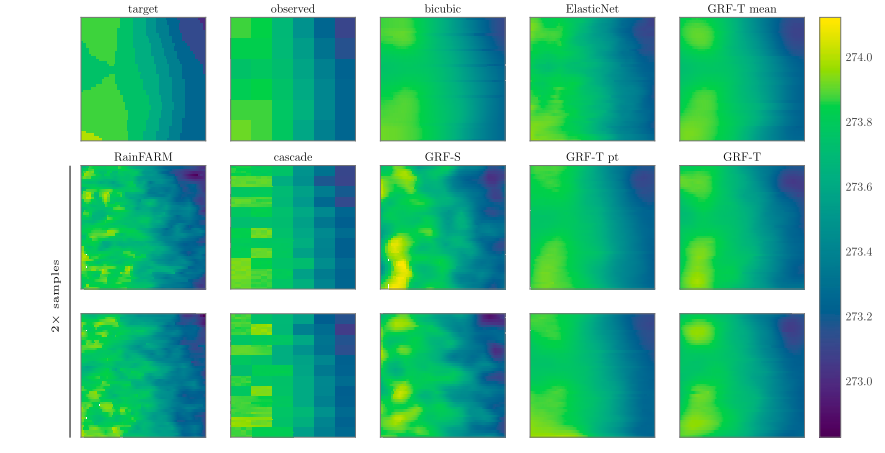
FIG. 15. Example outputs for the mesoscale GRF-T model and benchmarks for southeast England region. (a) Best, (b) median, and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.



(a)

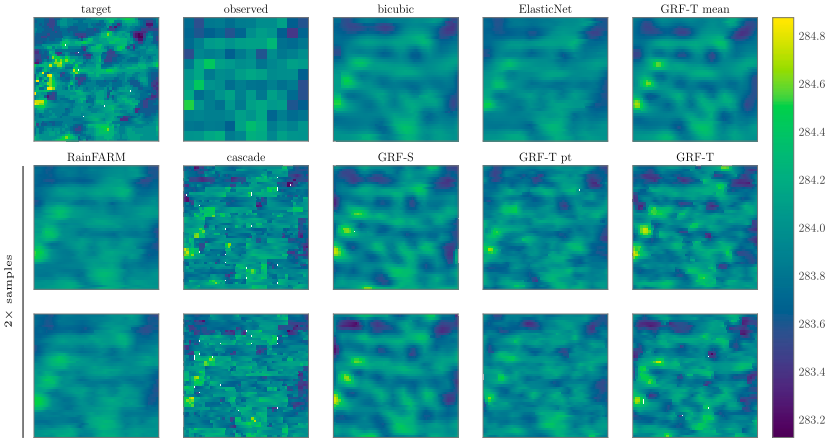


(b)

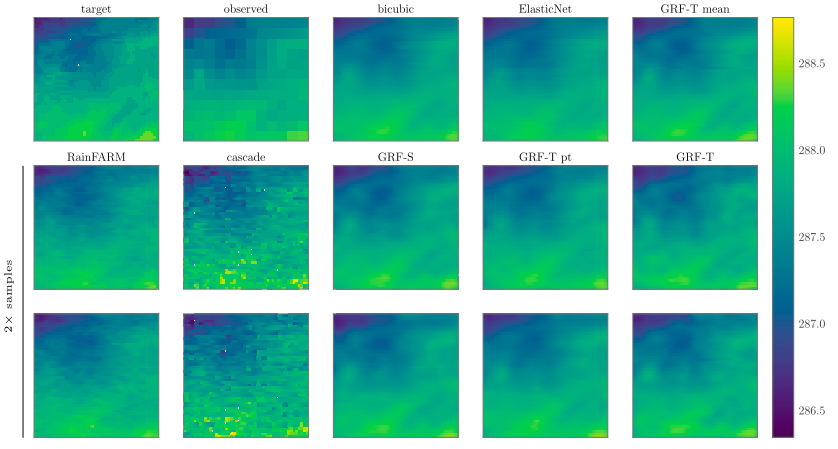


(c)

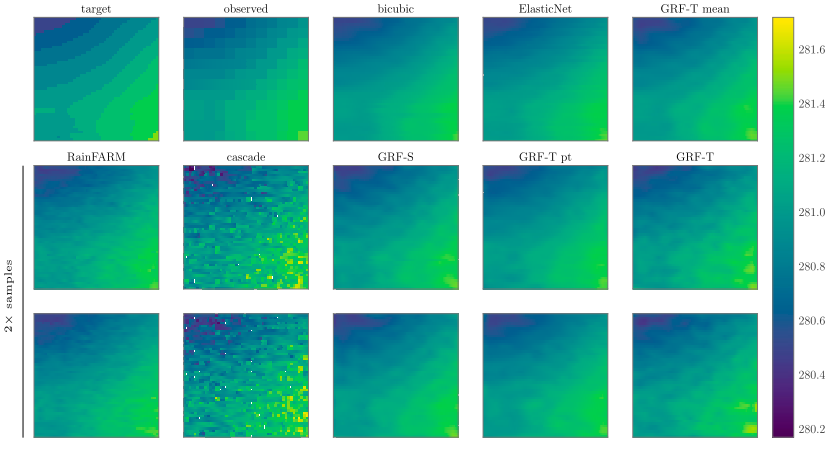
FIG. 16. Example outputs for the synoptic GRF-T model and benchmarks for Midlands region. (a) Best, (b) median and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.



(a)

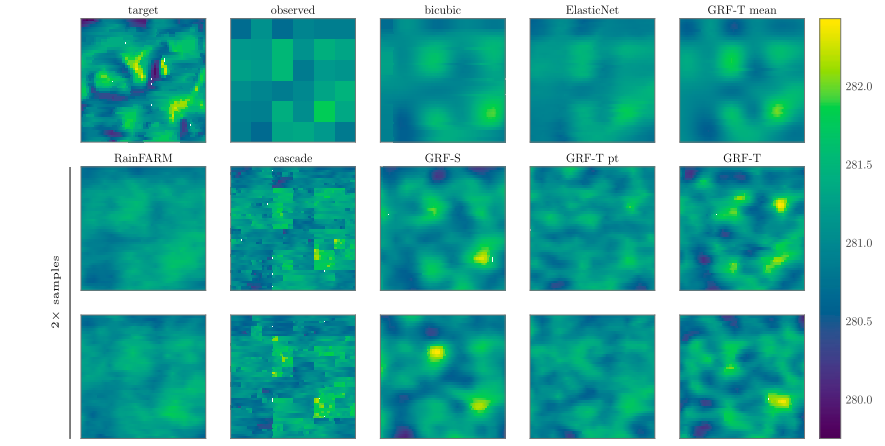


(b)

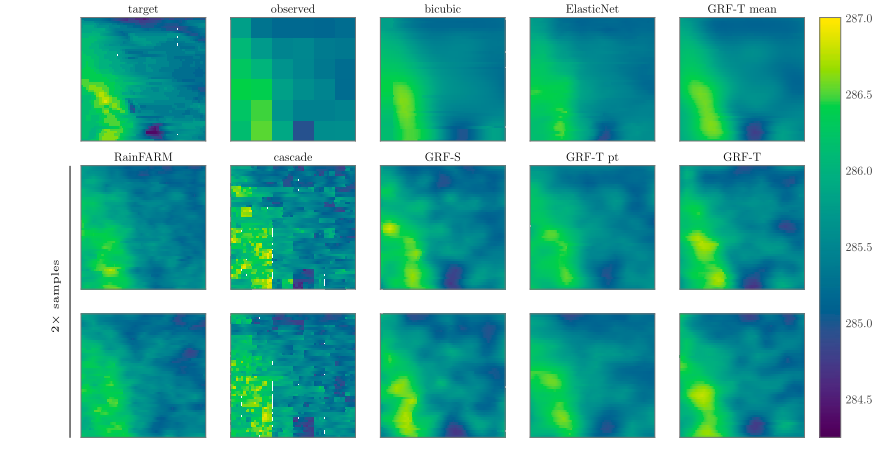


(c)

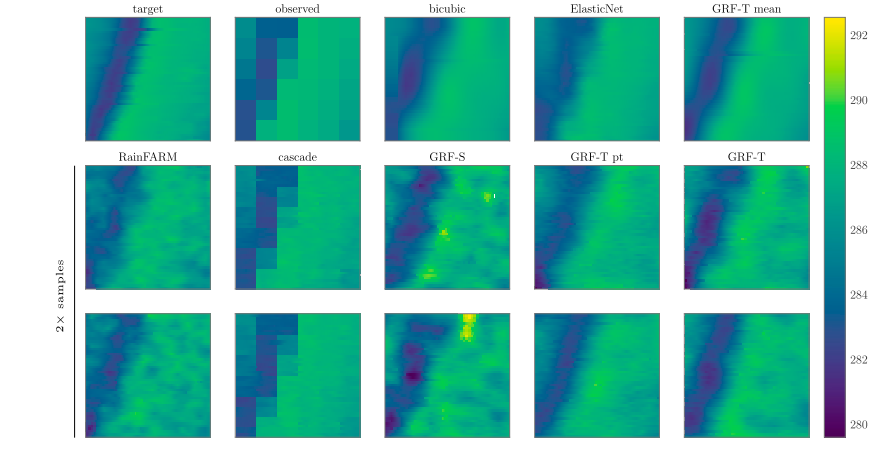
FIG. 17. Example outputs for the mesoscale GRF-T model and benchmarks for Midlands region. (a) Best, (b) median, and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.



(a)

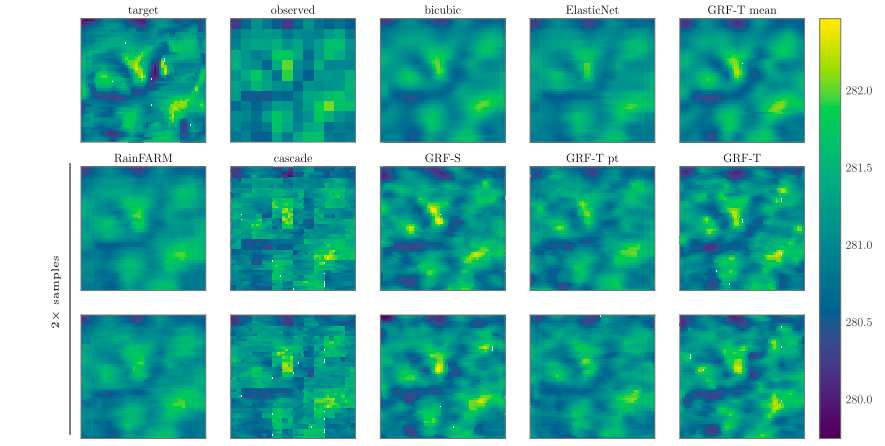


(b)

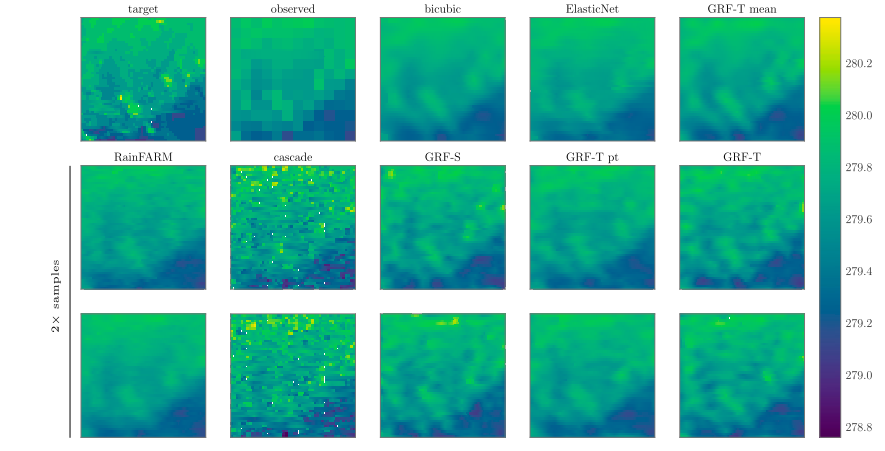


(c)

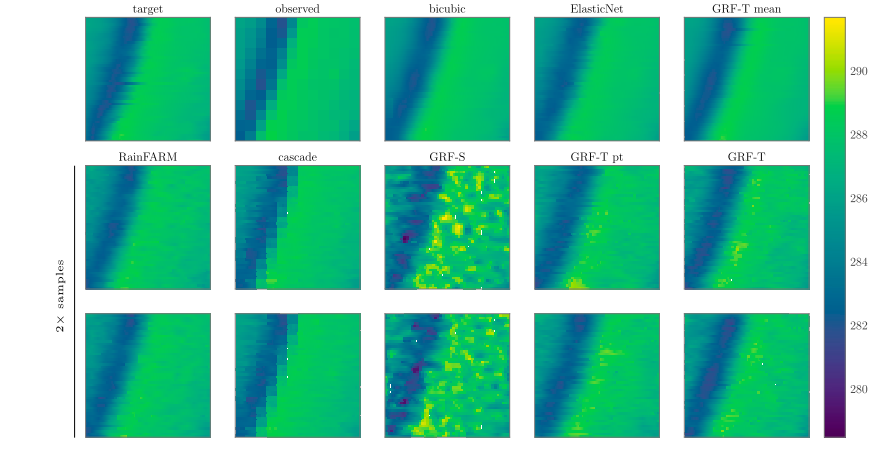
FIG. 18. Example outputs for the synoptic GRF-T model and benchmarks for Ireland region. (a) Best, (b) median, and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.



(a)



(b)



(c)

FIG. 19. Example outputs for the mesoscale GRF-T model and benchmarks for Ireland region. (a) Best, (b) median, and (c) worst performance of GRF-T compared to bicubic interpolation in terms of neighborhood Wasserstein score. Area shown is approximately 106 km^2.

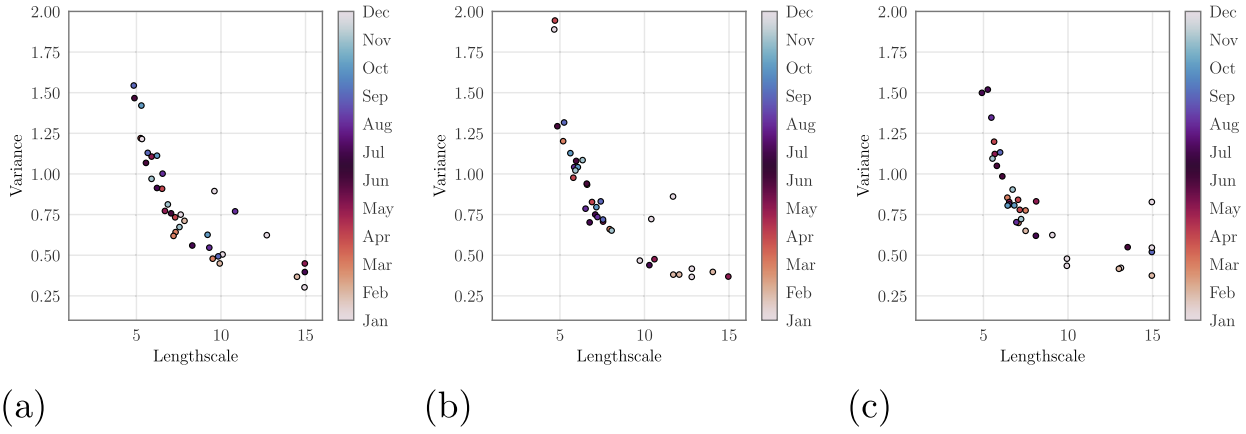


FIG. 20. Length scale and variance parameters selected by the GRF-T model in the synoptic experiment for (a) southeast England, (b) Midlands, and (c) Ireland regions. Colors denote the sample month. For legibility, two outlier points are omitted from (c) with coordinates (4, 2.4) and (13.5, 3).

scores such as MSE can be misleading when considered in isolation. In this work, we have therefore considered a suite of verification scores which includes point-based, probabilistic, and spatial scores (including one which is new to the literature). This multifaceted approach to verification can greatly improve progress in probabilistic and stochastic environmental modeling.

Overall, our model compares well to the benchmarks we have considered. It has the lowest ensemble CRPS in all cases, which demonstrates that it produces skillful forecasts in a probabilistic sense. The model mean has the lowest MSE score of all the models, including the data-driven deterministic ElasticNet model. Finally, the model scores extremely well when considering the neighborhood Wasserstein and PSD scores together (Figs. 12 and 13), although it is not uniformly the best when considering either score individually.

To test the effectiveness of different aspects of our model we have compared against two alternative GRF models, one using a time-stationary covariance, and one using point conditioning instead of spatial conditioning. Our GRF-T model outperforms both alternatives, giving the best performance in the PSD, neighborhood Wasserstein, and CRPS scores. Thus, it appears that both the spatial conditioning approach and the adaptive covariance help to improve the model performance.

With regards to future work, there are various extensions to our model which could be considered. For example, our model can be extended by considering more general methods of estimating the high-resolution covariance without any change to the conditioning step. A clear avenue for future work is therefore to replace our simple method of covariance estimation with one which can account for spatial nonstationarity.

Going further in this direction, the isotropic assumption used in our covariance model is likely to be unjustified for convective-scale weather, since factors such as synoptic drivers and atmospheric waves can introduce a major directional component. In principle, this can be easily dealt with by considering a linear transformation of the input space defined by a positive semidefinite matrix $\mathbf{M}$ which captures the directional dependence, so that the distance r^2 between points $\mathbf{x}$ and

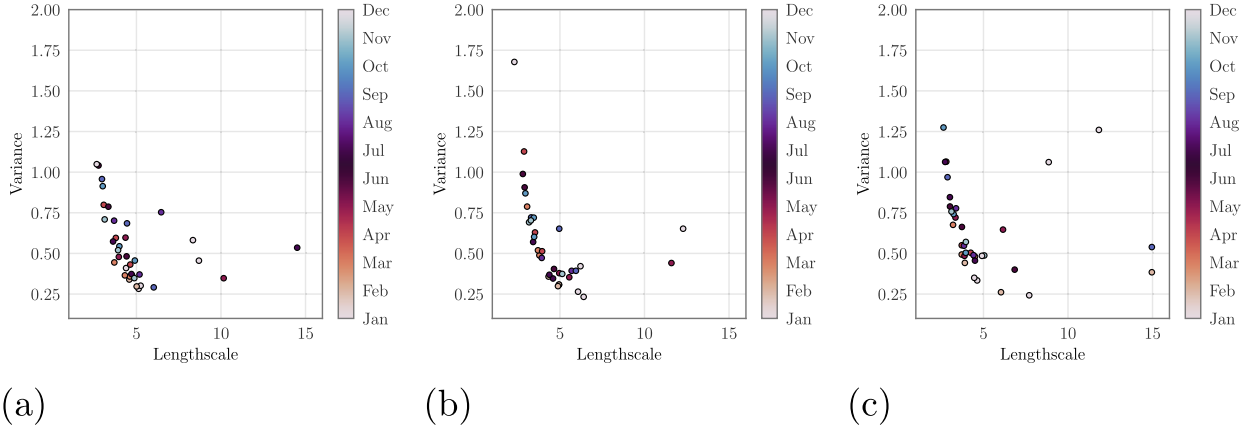


FIG. 21. Length scale and variance parameters selected by the GRF-T model in the mesoscale experiment for (a) southeast England, (b) Midlands, and (c) Ireland regions. Colors denote the sample month. For legibility, one outlier point is omitted from (c) with coordinates (14.5, 4.1).

$\mathbf{x}'$ is given by $r^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^{\mathrm{T}}\mathbf{M}(\mathbf{x} - \mathbf{x}')$ (Rasmussen and Williams 2006). In addition, the present approach of estimating the covariance for each time-step independently could be replaced by one which makes use of temporal continuity, such as an autoregressive model or a Kalman filter (e.g., Carron et al. (2016)).

Besides the assumptions used for covariance estimation, a core limitation of our current method is its assumption of Gaussianity. Another branch of future work could consider extensions to non-Gaussian variables, particularly cloud cover and rainfall.

Data availability statement. The data used for this study are available from the Met Office archives. Contact the corresponding author for access.

APPENDIX A

Conditioning a Gaussian Density

Theorem 1. If $\mathbf{x}_o$ and $\mathbf{x}_t$ follow a joint Gaussian distribution so that

$$\begin{bmatrix}\mathbf{x}_t\\ \mathbf{x}_o\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_t\\ \boldsymbol{\mu}_o\end{bmatrix}, \begin{bmatrix}\mathbf{C}_t & \mathbf{C}_{o,t}\\ \mathbf{C}_{t,o} & \mathbf{C}_o\end{bmatrix}\right),$$

then the conditional distribution $p(\mathbf{x}_t|\mathbf{x}_o)$ is Gaussian with mean $\boldsymbol{\mu}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_o)$ and covariance $\mathbf{C}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}\mathbf{C}_{t,o}$.

Proof. Our exposition here follows Do (2008). We have the conditional density formula:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_o) &= \frac{p(\mathbf{x}_t, \mathbf{x}_o)}{p(\mathbf{x}_o)} = \frac{1}{Z}\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}_t - \boldsymbol{\mu}_t\\ \mathbf{x}_o - \boldsymbol{\mu}_o\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{C}_t & \mathbf{C}_{o,t}\\ \mathbf{C}_{t,o} & \mathbf{C}_o\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}_t - \boldsymbol{\mu}_t\\ \mathbf{x}_o - \boldsymbol{\mu}_o\end{bmatrix}\right)\\ &= \frac{1}{Z}\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}_t - \boldsymbol{\mu}_t\\ \mathbf{x}_o - \boldsymbol{\mu}_o\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{P}_t & \mathbf{P}_{o,t}\\ \mathbf{P}_{t,o} & \mathbf{P}_o\end{bmatrix}\begin{bmatrix}\mathbf{x}_t - \boldsymbol{\mu}_t\\ \mathbf{x}_o - \boldsymbol{\mu}_o\end{bmatrix}\right), \end{aligned} \tag{A1}$$

where we have extracted all terms not depending on $\mathbf{x}_t$ into a normalization constant Z and used the precision matrix notation for the inverse covariance, $\mathbf{P} = \mathbf{C}^{-1}$.

Expanding the matrix product gives us

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_o) = \frac{1}{Z}\exp\Big\{-\frac{1}{2}[&(\mathbf{x}_t - \mu_t)^{\mathrm{T}}\mathbf{P}_t(\mathbf{x}_t - \boldsymbol{\mu}_t)\\ &+ (\mathbf{x}_o - \boldsymbol{\mu}_o)^{\mathrm{T}}\mathbf{P}_{t,o}(\mathbf{x}_t - \boldsymbol{\mu}_t) + (\mathbf{x}_t - \boldsymbol{\mu}_t)^{\mathrm{T}}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)\\ &+ (\mathbf{x}_o - \boldsymbol{\mu}_o)^{\mathrm{T}}\mathbf{P}_o(\mathbf{x}_o - \boldsymbol{\mu}_o)]\Big\}. \end{aligned} \tag{A2}$$

Now we can complete the square to give a quadratic term in $\mathbf{x}_t$ plus a constant term. We use the formula:

$$\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + 2\mathbf{z}^{\mathrm{T}}\mathbf{b} + \mathbf{c} = (\mathbf{z} + \mathbf{A}^{-1}\mathbf{b})^{\mathrm{T}}\mathbf{A}(\mathbf{z} + \mathbf{A}^{-1}\mathbf{b}) + \mathbf{c} - \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b} \tag{A3}$$

for a symmetric matrix $\mathbf{A}$, and substitute

$$\begin{aligned} \mathbf{z} &= (\mathbf{x}_t - \boldsymbol{\mu}_t),\\ \mathbf{A} &= \mathbf{P}_t,\\ \mathbf{b} &= \mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o),\\ \mathbf{c} &= (\mathbf{x}_o - \boldsymbol{\mu}_o)^{\mathrm{T}}\mathbf{P}_o(\mathbf{x}_o - \boldsymbol{\mu}_o), \end{aligned} \tag{A4}$$

to give

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_o) = \frac{1}{Z}\exp\Big(-\frac{1}{2}\{&[\mathbf{x}_t - \boldsymbol{\mu}_t + \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)]^{\mathrm{T}}\mathbf{P}_t[\mathbf{x}_t - \boldsymbol{\mu}_t + \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)]\\ &+ (\mathbf{x}_o - \boldsymbol{\mu}_o)^{\mathrm{T}}\mathbf{P}_o(\mathbf{x}_o - \boldsymbol{\mu}_o) - (\mathbf{x}_o - \boldsymbol{\mu}_o)^{\mathrm{T}}\mathbf{P}_{t,o}\mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)\}\Big). \end{aligned} \tag{A5}$$

We can then move the constant term not depending on $\mathbf{x}_t$ out of the exponential into the normalization term:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_o) = \frac{1}{Z'}\exp\Big\{-\frac{1}{2}&[\mathbf{x}_t - \boldsymbol{\mu}_t + \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)]^{\mathrm{T}}\\ &\times \mathbf{P}_t[\mathbf{x}_t - \boldsymbol{\mu}_t + \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)]\Big\}. \end{aligned} \tag{A6}$$

This gives a Gaussian density in terms of the precision $\mathbf{P}$, with mean $\boldsymbol{\mu}_t - \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o)$ and covariance $\mathbf{P}_t^{-1}$. Now we would like to put this in terms of the covariance $\mathbf{C}$. We can do so using the matrix inversion formula by noting that

$$\begin{bmatrix}\mathbf{C}_t & \mathbf{C}_{o,t}\\ \mathbf{C}_{t,o} & \mathbf{C}_o\end{bmatrix} = \begin{bmatrix}(\mathbf{P}_t - \mathbf{P}_{o,t}\mathbf{P}_o^{-1}\mathbf{P}_{t,o})^{-1} & -\mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{P}_o - \mathbf{P}_{t,o}\mathbf{P}_t^{-1}\mathbf{P}_{o,t})^{-1}\\ -(\mathbf{P}_o - \mathbf{P}_{t,o}\mathbf{P}_t^{-1}\mathbf{P}_{o,t})^{-1}\mathbf{P}_{t,o}\mathbf{P}_t & (\mathbf{P}_o - \mathbf{P}_{t,o}\mathbf{P}_t^{-1}\mathbf{P}_{o,t})^{-1}\end{bmatrix} \tag{A7}$$

and so

$$\boldsymbol{\mu}_{t|o} = \boldsymbol{\mu}_t - \mathbf{P}_t^{-1}\mathbf{P}_{o,t}(\mathbf{x}_o - \boldsymbol{\mu}_o) = \boldsymbol{\mu}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_o). \quad \text{(A8)}$$

On the other hand, we have

$$\begin{bmatrix} \mathbf{P}_t & \mathbf{P}_{o,t} \\ \mathbf{P}_{t,o} & \mathbf{P}_o \end{bmatrix} = \begin{bmatrix} (\mathbf{C}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}\mathbf{C}_{t,o})^{-1} & -\mathbf{C}_t^{-1}\mathbf{C}_{o,t}(\mathbf{C}_o - \mathbf{C}_{t,o}\mathbf{C}_t^{-1}\mathbf{C}_{o,t})^{-1} \\ -(\mathbf{C}_o - \mathbf{C}_{t,o}\mathbf{C}_t^{-1}\mathbf{C}_{o,t})^{-1}\mathbf{C}_{t,o}\mathbf{C}_t & (\mathbf{C}_o - \mathbf{C}_{t,o}\mathbf{C}_t^{-1}\mathbf{C}_{o,t})^{-1} \end{bmatrix}, \quad \text{(A9)}$$

which gives us

$$\mathbf{C}_{t|o} = \mathbf{P}_t^{-1} = \mathbf{C}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}\mathbf{C}_{t,o} \quad \text{(A10)}$$

as required.

Theorem 2. For a continuous Gaussian process with covariance kernel k, the conditional density of a point value $\mathbf{x}_t$ conditioned on a spatial average $\mathbf{x}_o = \left[1/\left(\int_T 1\, d\tau\right)\right]\int_T \mathbf{x}_\tau\, d\tau$ is given by

$$\begin{aligned} \boldsymbol{\mu}_{t|o} &= \mathbf{C}_{o,t}\mathbf{C}_o^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_o), \\ \mathbf{C}_{t|o} &= \mathbf{C}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}\mathbf{C}_{t,o}, \end{aligned} \quad \text{(A11)}$$

with

$$\begin{aligned} C_{(t)i,j} &= k(x_i, x_j), \\ C_{(o,t)i,j} &= C_{(t,o)j,i} = \int_{T_j} k(x_i, x_{\tau_j})\, d\tau_j, \\ C_{(o)i,j} &= \iint_{T_i T_j} k(x_{\tau_i}, x_{\tau_j})\, d\tau_i\, d\tau_j, \end{aligned} \quad \text{(A12)}$$

and

$$\boldsymbol{\mu}_o = \mathbb{E}(\mathbf{x}_o) = \int_T \mathbb{E}(\mathbf{x}_\tau)\, d\tau. \quad \text{(A13)}$$

Proof. Conditioning on a spatial average is closely related to conditioning on an integral of the underlying process. In fact, the average is equivalent to an integral multiplied by a scaling factor $\left[1/\left(\int_T 1\, d\tau\right)\right]\int_T \mathbf{x}_\tau\, d\tau$. For the remainder of this section, we take $I = \int_T \mathbf{x}_\tau\, d\tau$ to be such a rescaled integral so that $\int_T 1\, d\tau = 1$.

In what follows, we will need to make reference to the space of samples $\mathcal{H}$ of the Gaussian process. We will take $s \in \mathcal{H}$ to be any possible sample of the process, and $p(s)$ to be the probability density of the sample s.

Considering first the mean, we can use this terminology to write the expected value of the observed spatial integral I as

$$\mathbb{E}(I) = \int_{\mathcal{H}} I(s)p(s)\, ds, \quad \text{(A14)}$$

where $I(s)$ is the integral I taken over a sample s. Here, we have rewritten the expectation as a weighted integral over all possible samples of our process.

We can rewrite this equation by noting that the probability density of process samples is symmetrically weighted around the mean, that is,

$$p[\mathbb{E}(s) + r] = p[\mathbb{E}(s) - r]. \quad \text{(A15)}$$

On the other hand, the value of the spatial integral I follows the relation:

$$\begin{aligned} I[\mathbb{E}(s) + r] &= I[\mathbb{E}(s)] + I(r) \\ I[\mathbb{E}(s) - r] &= I[\mathbb{E}(s)] - I(r). \end{aligned} \quad \text{(A16)}$$

When we consider Eq. (A14) the opposite-signed terms involving r then cancel, leaving

$$\begin{aligned} \mathbb{E}(I) &= \int_{\mathcal{H}} I(s)p(s)\, ds = \int_{\mathcal{H}} I[\mathbb{E}(s)]p(s)\, ds \\ &= I[\mathbb{E}(s)]\int_{\mathcal{H}} p(s)\, ds = I[\mathbb{E}(s)]. \end{aligned} \quad \text{(A17)}$$

For the covariance, we need to consider both the covariance between an integral and a point, and the covariance between two integrals (the covariance between two points being the standard form). We begin with the covariance between a point $\mathbf{x}$ and an integral $I = \int_T \mathbf{x}_\tau\, d\tau$. We have

$$\operatorname{cov}(\mathbf{x}, I) = \mathbb{E}(\mathbf{x}I) - \mathbb{E}(\mathbf{x})\mathbb{E}(I). \quad \text{(A18)}$$

Considering the term $\mathbb{E}(\mathbf{x}I)$ we have

$$\begin{aligned} \mathbb{E}(\mathbf{x}I) &= \mathbb{E}\left(\mathbf{x}\int_T \mathbf{x}_\tau\, d\tau\right) \\ &= \mathbb{E}\left(\int_T \mathbf{x}\mathbf{x}_\tau\, d\tau\right) \\ &= \int_{\mathcal{H}}\int_T \mathbf{x}\mathbf{x}_\tau\, d\tau\, p(s)\, ds \\ &= \int_{\mathcal{H}}\int_T \mathbf{x}\mathbf{x}_\tau\, p(s)\, d\tau\, ds. \end{aligned} \quad \text{(A19)}$$

Now, $\mathbf{x}\mathbf{x}_\tau$ is continuous everywhere in $\mathcal{H} \times T$ where s is continuous. But if the process we are considering uses a standard kernel function such as the Matern or squared exponential kernel, its sample paths are always continuous (Paciorek 2003). Thus the integrand is a continuous function and we can change the order of integration to give

$$\begin{aligned} \mathbb{E}(\mathbf{x}I) &= \int_{\mathcal{H}}\int_T \mathbf{x}\mathbf{x}_\tau\, p(s)\, d\tau\, ds \\ &= \int_T\int_{\mathcal{H}} \mathbf{x}\mathbf{x}_\tau\, p(s)\, ds\, d\tau \\ &= \int_T \mathbb{E}(\mathbf{x}\mathbf{x}_\tau)\, d\tau \\ &= \int_T \operatorname{cov}(\mathbf{x}, \mathbf{x}_\tau) + \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x}_\tau)\, d\tau \\ &= \int_T \operatorname{cov}(\mathbf{x}, \mathbf{x}_\tau)\, d\tau + \mathbb{E}(\mathbf{x})\mathbb{E}(I) \end{aligned} \quad \text{(A20)}$$

by (A14), and so

$$\begin{aligned}\operatorname{cov}(\mathbf{x}, I) &= \int_T \operatorname{cov}(\mathbf{x}, \mathbf{x}_\tau)\, d\tau + \mathbb{E}(\mathbf{x})\mathbb{E}(I) - \mathbb{E}(\mathbf{x})\mathbb{E}(I) \\ &= \int_T \operatorname{cov}(\mathbf{x}, \mathbf{x}_\tau)\, d\tau \\ &= \int_T k(\mathbf{x}, \mathbf{x}_\tau)\, d\tau \end{aligned} \tag{A21}$$

for covariance kernel k.

We now consider the covariance between two integrals, $I_1 = \int_{T_1} x_{\tau_1}\, d\tau_1$ and $I_2 = \int_{T_2} x_{\tau_2}\, d\tau_2$. As before, we have

$$\operatorname{cov}(I_1, I_2) = \mathbb{E}(I_1 I_2) - \mathbb{E}(I_1)\mathbb{E}(I_2), \tag{A22}$$

and considering the first term gives us

$$\mathbb{E}(I_1 I_2) = \mathbb{E}\left(\int_{T_1} x_{\tau_1} d\tau_1 \int_{T_2} x_{\tau_2} d\tau_2\right). \tag{A23}$$

Since the domains of integration are independent and the integrands are continuous by the argument above, we can rewrite this as a double integral:

$$\mathbb{E}(I_1 I_2) = \mathbb{E}\left(\int_{T_1}\int_{T_2} x_{\tau_1} x_{\tau_2}\, d\tau_1\, d\tau_2\right). \tag{A24}$$

Now the argument of (A17) gives

$$\begin{aligned}\mathbb{E}(I_1 I_2) &= \int_{T_1}\int_{T_2} \mathbb{E}(x_{\tau_1} x_{\tau_2})\, d\tau_1\, d\tau_2 \\ &= \int_{T_1}\int_{T_2} \operatorname{cov}(x_{\tau_1}, x_{\tau_2}) + \mathbb{E}(x_{\tau_1})\mathbb{E}(x_{\tau_2})\, d\tau_1\, d\tau_2 \\ &= \int_{T_1}\int_{T_2} \operatorname{cov}(x_{\tau_1}, x_{\tau_2})\, d\tau_1\, d\tau_2 + \mathbb{E}(I_1)\mathbb{E}(I_2), \end{aligned} \tag{A25}$$

so that

$$\begin{aligned}\operatorname{cov}(I_1, I_2) &= \int_{T_1}\int_{T_2} \operatorname{cov}(x_{\tau_1}, x_{\tau_2})\, d\tau_1\, d\tau_2 \\ &= \int_{T_1}\int_{T_2} k(x_{\tau_1}, x_{\tau_2})\, d\tau_1\, d\tau_2 \end{aligned} \tag{A26}$$

for covariance kernel k.

We can combine this with our equations from the previous section, namely,

$$\begin{aligned}\boldsymbol{\mu}_{t|o} &= \mathbf{C}_{o,t}\mathbf{C}_o^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_o), \\ \mathbf{C}_{t|o} &= \mathbf{C}_t - \mathbf{C}_{o,t}\mathbf{C}_o^{-1}\mathbf{C}_{t,o}, \end{aligned} \tag{A27}$$

with integrals $I = \int_T \mathbf{x}_\tau\, d\tau$ playing the role of the observed variables $\mathbf{x}_o$, and target variables $\mathbf{x}_t$ remaining discrete as before. We now have

$$\begin{aligned} C_{(t)i,j} &= k(x_i, x_j), \\ C_{(o,t)i,j} &= C_{(t,o)j,i} = \int_{T_j} k(x_i, x_{\tau_j})\, d\tau_j, \\ C_{(o)i,j} &= \iint_{T_i T_j} k(x_{\tau_i}, x_{\tau_j})\, d\tau_i\, d\tau_j, \end{aligned} \tag{A28}$$

and

$$\boldsymbol{\mu}_o = \mathbb{E}(\mathbf{x}_o) = \int_T \mathbb{E}(\mathbf{x}_\tau)\, dt. \tag{A29}$$

In the discrete case, this corresponds to the relations in Eq. (5).

APPENDIX B

Computation of 1-Wasserstein Distance

By using the Earth movers's interpretation of the 1-Wasserstein distance, we know that it can be found by summing the distance moved by each point. Each value in the first list must map uniquely onto a value of the second list. So, for any chosen ordering on the first list, we can find an ordering of the second list such that the summed difference between the ordered lists is equal to the 1-Wasserstein distance, and specifically this ordering is the one which minimizes the summed difference. Since we have the freedom to choose an ordering on the first list, let us assume it to be sorted in increasing order. Suppose we have an ordering on the second list. Clearly, any pair of two points in the second list are either in increasing order or they are swapped. We will proceed by showing that exchanging two swapped points will never increase the total difference, it must either reduce the difference or leave it unchanged. Thus, ordering the second list in increasing order must give a minimal value for the summed difference of the two lists, giving the desired result.

It remains to prove that exchanging two swapped points of the second list will never increase the total summed distance. We will do so by enumerating cases. We take A to be the ordered list and B to define our mapping. The total difference is given by $\sum_i |a_i - b_i|$, with i an index on A and B.

Consider switching a pair b_i with b_j in B with $b_j < b_i$, and let a_i and a_j be the corresponding elements of A with $a_i < a_j$. There are three possible cases we need to consider. One of the following must hold:

1) $a_i \le a_j \le b_j < a_i$
2) $a_i \le b_j \le a_j \le b_i$
3) $a_i \le b_j \le b_i \le a_j$
4) $b_j \le b_i \le a_i \le a_j$
5) $b_i \le a_i \le b_i \le a_j$
6) $b_j \le a_i \le a_j \le b_i$

Case 1 (a symmetrical argument holds for case 3):

$$\begin{aligned}|b_i - a_i| + |b_j - a_j| &= (b_i - a_i) + (b_j - a_j) \\ &= [(b_i - b_j) + (b_j - a_j) + (a_j - a_i)] + (b_j - a_j) \\ &= [(a_j - a_i) + (b_j - a_j)] + [(b_j - a_j) + (b_i - b_j)] \\ &= (b_j - a_i) + (b_i - a_j) = |b_j - a_i| + |b_i - a_j| \end{aligned}$$

Case 2 (a symmetrical argument holds for case 4):

$$\begin{aligned}|b_i - a_i| + |b_j - a_j| &= (b_i - a_i) + (a_j - b_j) \\ &= [(b_i - a_j) + (a_j - b_j) + (b_j - a_i)] + (a_j - b_j) \\ &\ge (b_j - a_i) + (b_i - a_j) \\ &= |b_j - a_i| + |b_i - a_j| \end{aligned}$$

Case 3 (a symmetrical argument holds for case 6):

$$\begin{aligned}|b_i - a_i| + |b_j - a_j| &= (b_i - a_i) + (a_j - b_j)\\ &= [(b_i - b_j) + (b_j - a_i)] + [(a_j - b_i) - (b_i - b_j)]\\ &\geq (b_j - a_i) + (a_j - b_i)\\ &= |b_j - a_i| + |b_i - a_j|\end{aligned}$$

Thus, in all cases exchanging the elements b_i and b_j either reduces or leaves unchanged the total difference. This concludes our argument.

REFERENCES

Allcroft, D. J., and C. A. Glasbey, 2003: A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *J. Roy. Stat. Soc.*, **52C**, 487–498, https://doi.org/10.1111/1467-9876.00419.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment. Part I. *J. Atmos. Sci.*, **31**, 674–701, https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2.

Arjovsky, M., S. Chintala, and L. Bottou, 2017: Wasserstein generative adversarial networks. *Proc. 34th Int. Conf. on Machine Learning*, Vol. 70, 214–223, https://proceedings.mlr.press/v70/arjovsky17a.html.

Barratt, S., and R. Sharma, 2018: A note on the inception score. arXiv:1801.01973.

Bellemare, M. G., I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, 2018: The Cramer distance as a solution to biased Wasserstein gradients. arXiv:1705.10743.

Bordoy, R., and P. Burlando, 2014: Stochastic downscaling of climate model precipitation outputs in orographically complex regions: 2. Downscaling methodology. *Water Resour. Res.*, **50**, 562–579, https://doi.org/10.1002/wrcr.20443.

Box, G. E. P., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26B**, 211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Carron, A., M. Todescato, R. Carli, L. Schenato, and G. Pillonetto, 2016: Machine learning meets Kalman filtering. *IEEE 55th Conf. on Decision and Control (CDC). 2016*, Las Vegas, NV, IEEE, 4594–4599, https://doi.org/10.1109/CDC.2016.7798968.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154, https://doi.org/10.1017/S1350482704001239.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Do, C. B., 2008: More on multivariate Gaussians. Stanford University, 11 pp., http://cs229.stanford.edu/section/more_on_gaussians.pdf.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

Feydy, J., T. Séjourné, F.-X. Vialard, S. Amari, A. Trouvé, and G. Peyré, 2019: Interpolating between optimal transport and MMD using Sinkhorn divergences. *22nd Int. Conf. on Artificial Intelligence and Statistics*, PMLR, 2681–2690.

Flack, D. L. A., R. S. Plant, S. L. Gray, H. W. Lean, C. Keil, and G. C. Craig, 2016: Characterisation of convective regimes over the British Isles: Convection over the British Isles. *Quart. J. Roy. Meteor. Soc.*, **142**, 1541–1553, https://doi.org/10.1002/qj.2758.

Friedman, J., and Coauthors, 2001: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 1. Springer, 533 pp.

Gagnon, P., A. N. Rousseau, A. Mailhot, and D. Caya, 2012: Spatial disaggregation of mean areal rainfall using Gibbs sampling. *J. Hydrometeor.*, **13**, 324–337, https://doi.org/10.1175/JHM-D-11-034.1.

Ge, Y., and Coauthors, 2019: Principles and methods of scaling geospatial Earth science data. *Earth-Sci. Rev.*, **197**, 102897, https://doi.org/10.1016/j.earscirev.2019.102897.

Ghosh, S., 2010: SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *J. Geophys. Res.*, **115**, D22102, https://doi.org/10.1029/2009JD013548.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Golding, B., N. Roberts, G. Leoncini, K. Mylne, and R. Swinbank, 2016: MOGREPS-UK convection-permitting ensemble products for surface water flood forecasting: Rationale and first results. *J. Hydrometeor.*, **17**, 1383–1406, https://doi.org/10.1175/JHM-D-15-0083.1.

Gotway, C. A., and L. J. Young, 2002: Combining incompatible spatial data. *J. Amer. Stat. Assoc.*, **97**, 632–648, https://doi.org/10.1198/016214502760047140.

Gregory, D., and P. R. Rowntree, 1990: A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Wea. Rev.*, **118**, 1483–1506, https://doi.org/10.1175/1520-0493(1990)118<1483:AMFCSW>2.0.CO;2.

Gupta, V. K., and E. C. Waymire, 1993: A statistical analysis of mesoscale rainfall as a random cascade. *J. Appl. Meteor.*, **32**, 251–267, https://doi.org/10.1175/1520-0450(1993)032<0251:ASAOMR>2.0.CO;2.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

Hessami, M., P. Gachon, T. B. Ouarda, and A. St-Hilaire, 2008: Automated regression-based statistical downscaling tool. *Environ. Modell. Software*, **23**, 813–834, https://doi.org/10.1016/j.envsoft.2007.10.004.

Hewer, R., P. Friederichs, A. Hense, and M. Schlather, 2017: A Matérn-based multivariate Gaussian random process for a consistent model of the horizontal wind components and related variables. *J. Atmos. Sci.*, **74**, 3833–3845, https://doi.org/10.1175/JAS-D-16-0369.1.

Hu, M., and Y. Huang, 2020: atakrig: An R package for multivariate area-to-area and area-to-point kriging predictions. *Comput. Geosci.*, **139**, 104471, https://doi.org/10.1016/j.cageo.2020.104471.

Kelsall, J., and J. Wakefield, 2002: Modeling spatial variation in disease risk. *J. Amer. Stat. Assoc.*, **97**, 692–701, https://doi.org/10.1198/016214502388618438.

Kleiber, W., R. W. Katz, and B. Rajagopalan, 2012: Daily spatio-temporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.*, **48**, W01523, https://doi.org/10.1029/2011WR011105.

Kyriakidis, P. C., 2004: A geostatistical framework for area-to-point spatial interpolation. *Geogr. Anal.*, **36**, 259–289, https://doi.org/10.1111/j.1538-4632.2004.tb01135.x.

Li, X. Y., L. M. Zhang, and J. H. Li, 2016: Using conditioned random field to characterize the variability of geologic profiles. *J. Geotech. Geoenviron. Eng.*, **142**, 04015096, https://doi.org/10.1061/(ASCE)GT.1943-5606.0001428.

Liu, Y., H. Qin, Z. Zhang, S. Pei, C. Wang, X. Yu, Z. Jiang, and J. Zhou, 2019: Ensemble spatiotemporal forecasting of solar irradiation using variational Bayesian convolutional gate recurrent unit network. *Appl. Energy*, **253**, 113596, https://doi.org/10.1016/j.apenergy.2019.113596.

Lovejoy, S., and B. B. Mandelbrot, 1985: Fractal properties of rain, and a fractal model. *Tellus*, **37A**, 209–232, https://doi.org/10.1111/j.1600-0870.1985.tb00423.x.

Martinez, M., M. Tapaswi, and R. Stiefelhagen, 2016: A closed-form gradient for the 1D Earth mover's distance for spectral deep learning on biological data. *ICML 2016 Workshop on Computational Biology (CompBio@ICML16)*, New York, NY, 5 pp.

Martinez-Beneito, M. A., 2013: A general modelling framework for multivariate disease mapping. *Biometrika*, **100**, 539–553, https://doi.org/10.1093/biomet/ast023.

Mathieu, M., C. Couprie, and Y. LeCun, 2015: Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440.

Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain, 2015: A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Stat.*, **24**, 579–599, https://doi.org/10.1080/10618600.2014.914946.

Paciorek, C., 2003: Nonstationary Gaussian processes for regression and spatial modelling. Ph.D. thesis, Carnegie Mellon University, 236 pp.

Pardo-Iguzquiza, E., M. Chica-Olmo, and P. M. Atkinson, 2006: Downscaling cokriging for image sharpening. *Remote Sens. Environ.*, **102**, 86–98, https://doi.org/10.1016/j.rse.2006.02.014.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Perica, S., and E. Foufoula-Georgiou, 1996: Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. *J. Geophys. Res.*, **101**, 26 347–26 361, https://doi.org/10.1029/96JD01870.

Pulkkinen, S., D. Nerini, A. A. Perez Hortal, C. Velasco-Forero, A. Seed, U. Germann, and L. Foresti, 2019: Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geosci. Model Dev.*, **12**, 4185–4219, https://doi.org/10.5194/gmd-12-4185-2019.

Ramdas, A., N. Garcia, and M. Cuturi, 2015: On Wasserstein two sample testing and related families of nonparametric tests. *Entropy*, **19**, 47, https://doi.org/10.3390/e19020047.

Rasmussen, C. E., and C. K. I. Williams, 2006: *Gaussian Processes for Machine Learning.* Vol. 2. The MIT Press, 266 pp.

Raut, B. A., M. J. Reeder, C. Jakob, and A. W. Seed, 2019: Stochastic space-time downscaling of rainfall using event-based multiplicative cascade simulations. *J. Geophys. Res. Atmos.*, **124**, 3889–3902, https://doi.org/10.1029/2018JD029343.

Rebora, N., L. Ferraris, J. von Hardenberg, and A. Provenzale, 2006: RainFARM: Rainfall downscaling by a filtered autoregressive model. *J. Hydrometeor.*, **7**, 724–738, https://doi.org/10.1175/JHM517.1.

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmos. Res.*, **83**, 211–224, https://doi.org/10.1016/j.atmosres.2005.08.011.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rue, H., and L. Held, 2005: *Gaussian Markov Random Fields: Theory and Applications.* Chapman & Hall, 280 pp.

Sansom, P., 2015: Advances in forecast verification. *Weather*, **70**, https://doi.org/10.1002/wea.2445.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The Hydrological Ensemble Prediction Experiment. *Bull. Amer. Meteor. Soc.*, **88**, 1541–1547, https://doi.org/10.1175/BAMS-88-10-1541.

Schirrmann, M., R. Herbst, P. Wagner, and R. Gebbers, 2012: Area-to-point kriging of soil phosphorus composite samples. *Commun. Soil Sci. Plant Anal.*, **43**, 1024–1041, https://doi.org/10.1080/00103624.2012.656166.

Schleiss, M., 2020: A new discrete multiplicative random cascade model for downscaling intermittent rainfall fields. *Hydrol. Earth Syst. Sci.*, **24**, 3699–3723, https://doi.org/10.5194/hess-24-3699-2020.

Schoof, J. T., and S. C. Pryor, 2001: Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *Int. J. Climatol.*, **21**, 773–790, https://doi.org/10.1002/joc.655.

Stein, M. L., 1999: *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science and Business Media, 263 pp.

Strang, G., 1993: Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc.*, **28**, 288–306, https://doi.org/10.1090/S0273-0979-1993-00390-2.

Taylor, G., 2003: The phase problem. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 1881–1890, https://doi.org/10.1107/S0907444903017815.

Van Den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus*, **46A**, 314–324, https://doi.org/10.3402/tellusa.v46i3.15481.

Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, 2017: DeepSD: Generating high resolution climate change projections through single image super-resolution. *KDD'17: Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, ACM Press, 1663–1672, https://doi.org/10.1145/3097983.3098004.

——, ——, and A. Ganguly, 2019: Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theor. Appl. Climatol.*, **137**, 557–570, https://doi.org/10.1007/s00704-018-2613-3.

Villani, C., 2008: *Optimal Transport: Old and New*. Springer, 998 pp.

Wikle, C. K., R. F. Milliff, D. Nychka, and L. M. Berliner, 2001: Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds. *J. Amer. Stat. Assoc.*, **96**, 382–397, https://doi.org/10.1198/016214501753168109.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, and L. O. Mearns, 2004: Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting Material of the Intergovernmental Panel on Climate Change, 27 pp., https://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf.

Wood, A. W., 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, https://doi.org/10.1029/2001JD000659.

Wytock, M., and J. Z. Kolter, 2013: Large-scale probabilistic forecasting in energy systems using sparse Gaussian conditional random fields. *52nd IEEE Conf. on Decision and Control*, Firenze, Italy, IEEE, 1019–1024, https://doi.org/10.1109/CDC.2013.6760016.

Yu, B., K. Zhu, M. Xue, and B. Zhou, 2020: Using new neighborhood-based intensity-scale verification metrics to evaluate WRF precipitation forecasts at 4 and 12 km grid spacings. *Atmos. Res.*, **246**, 105117, https://doi.org/10.1016/j.atmosres.2020.105117.

Zhang, B., P. Dehghanian, and M. Kezunovic, 2016: Spatial-temporal solar power forecast through use of Gaussian conditional random fields. *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Boston, MA, IEEE, 1–5, https://doi.org/10.1109/PESGM.2016.7741503.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate*, **12**, 2474–2489, https://doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.