Uncertainty propagation in observational references to climate model scales

Omar Bellprat*

Barcelona Supercomputing Centre (BSC), Earth Sciences, Barcelona, Spain

François Massonnet

Barcelona Supercomputing Centre (BSC), Earth Sciences, Barcelona, Spain.

Université catholique de Louvain, Louvain-la-Neuve, Belgium

Stefan Siegert

Exeter Climate Systems, University of Exeter, United Kingdom

Chloé Prodhomme, Daniel Macias Gómez

Barcelona Supercomputing Centre (BSC), Earth Sciences, Barcelona, Spain

Virginie Guemas

Barcelona Supercomputing Centre (BSC), Earth Sciences, Barcelona, Spain

Centre National de Recherche Meteorologique, Meteo-France, 42 avenue Gaspard Coriolis, Toulouse, France

Francisco Doblas-Reyes

Barcelona Supercomputing Center (BSC), Earth Sciences, Barcelona, Spain.

ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

Corresponding author address: Omar Bellprat, Barcelona Supercomputing Centre, Carrer de Jordi Girona, 29-31, 08034 Barcelona

E-mail: omar.bellprat@bsc.es

ABSTRACT

Climate model simulations and observational references of the Earth's climate are the two primary sources of information used for climate related decision-making. While uncertainties in climate models and observational references have been assessed thoroughly, it has remained difficult to integrate these, partly because of the lack of formal concepts on how to consider observational uncertainties in model-observation comparison. One of the difficulties dealing with observational uncertainty is its propagation to the space-time scales represented by the models. This is a challenge due to the correlation of observational errors in space and time. Here we present an approximation which allows to derive propagation factors to different model scales and apply these to uncertainty estimates provided by the Climate Change Initiative (CCI) sea-surface temperature (SST) dataset. The propagated uncertainty in SST observations is found to systematically lower seasonal forecast skill and to increase the uncertainty in verification of seasonal forecasts, an aspect that remains currently overlooked. Uncertainty in forecast quality assessment is dominated by the shortness of the satellite record. Expanding the record length of these datasets might hence reduce the verification uncertainties more than the efforts to reduce the observational uncertainties.

1. Introduction

The scientific community is taking action to confront the challenge of climate variability and change by understanding the physical basis and by providing estimates of the present and future climate. Climate model simulations and observational references are the two resulting sources of information that support stakeholders and policymakers. The quantification of uncertainties in both sources of information is crucial and large efforts are devoted quantifying these (Flato et al. 2013; Hartmann et al. 2013).

Climate model uncertainties are typically assessed by comparing simulated and observed conditions of the past climate (Reichler and Kim 2008). The agreement between models and observations is instrumental in gaining confidence into simulated climates which have not yet been observed (Knutti 2008). This holds particularly for near-term climate predictions such as subseasonal to seasonal predictions where retrospective predictions can be verified (Doblas-Reyes et al. 2013). Accurate observational references of the Earth's climate are therefore indispensable to quantify model uncertainties, yet observations are subject to uncertainties as well. While the uncertainties related to the limited statistical sample in model-observation comparison is usually reported (e.g. for seasonal forecasting Doblas-Reyes et al. 2013; Ferro 2014; Scaife et al. 2014; Siegert et al. 2016b) uncertainties in the observational references remain weakly explored. This tendency pertains to the climate modelling community in general (as highlighted in Gómez-Navarro et al. 2012; Addor and Fischer 2015; Massonnet et al. 2016; Mudryk et al. 2017) despite the large efforts that have gone into quantifying uncertainties in observational references (Kennedy 2014; Povey and Grainger 2015; Merchant et al. 2017)

Like climate models, observational references rely on a number of structural and parametric choices in the design and calibration of the algorithm used to generate the data sets (Thorne et al.

2005; Liu et al. 2015) and are therefore an approximation of the theoretical true climate (Massonnet et al. 2016). Data sets report the resulting uncertainties typically by characterizing the dispersion of the error distribution between the measured and the theoretical true value (Merchant et al. 2014; Liu et al. 2015). One of the challenges in including these uncertainty estimates in the assessment of model simulations is the aggregation to the space-time averages, motivated by the mismatch in observational and model grids and data frequency. Measurement errors are correlated in time and space due to for instance the background atmospheric or oceanic conditions that prevail locally in time and in space (Povey and Grainger 2015). Therefore, the information about uncertainty has to be propagated taking into account the expected correlation structure of the observational errors. The lack of knowledge of correlation length scales but also the missing methodological concepts to efficiently propagate uncertainties remain key obstacles to estimating uncertainties at model scales. Past studies have therefore used alternative data sets to estimate observational uncertainties (Stoffelen 1998; Reichler and Kim 2008), however, this approach ignores the uncertainty estimates actually reported in the data sets. Providing methodologies of uncertainty propagation to climate model scales is therefore an opportunity to bridge the modelling and observational data communities.

The European Space Agency (ESA) Climate Change Initiative (CCI) has placed a special focus on estimating uncertainties in climate data records (Merchant et al. 2017). This is an important contribution towards mutual uncertainty assessment of models and observations. This study aims to support this practice by illustrating simple ways to propagate uncertainties to scales used in seasonal forecast verification of the El Niño Southern Oscillation (ENSO) relying on the CCI sea-surface temperature (SST) gap-free analysis (L4 product) (Merchant et al. 2014). The propagated observational uncertainties are subsequently confronted to two other uncertainties present in the context of forecast verification: the limited ensemble size and the limited record length of the datasets. The comparison allows to understand how important the observational uncertainty is in the practice of seasonal forecast verification. Finally, an estimate of the systematic reduction in seasonal forecast skill due to observational uncertainty is provided, highlighting the fact that current practice underestimates the deterministic skill of forecasting systems.

2. Methods

a. Observational references and seasonal forecast verification

The role of observational uncertainty is explored in this study using the SST CCI gap-free analysis v1.1 (Merchant et al. 2014) and three alternative SST data sets which use different data and techniques to represent observed SSTs namely: the Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) data set v.1.1 (Rayner et al. 2003), the ERA-Interim re-analysis (Dee et al. 2011), and the Extended Reconstructed Sea Surface Temperature (ERSST) v.4 data set (Huang et al. 2015). The observational references are hereafter called ORs. HadISST uses in-situ data (Met Office Marine Data Bank (MDB) and Comprehensive Ocean-Atmosphere Data Set (ICOADS) release 2.5 and satellite data from Advanced Very High Resolution Radiometers (AVHRR) data. ERA-Interim is an atmospheric re-analysis product and uses SST data from different sources as described in Dee et al. (2011) which include both in-situe and satellite remotely sensed data. ERSST4 relies exclusively on in-situ (ICOADS) data. Finally, SST CCI relies on satellite remotely sensed data only blended from AVHRR and (A)ATSR (Advanced Along-Track Scanning Radiometers including ATSR1 and ATSR2). SST CCI and ERA-Interim (from 2009 onwards) use data from the near-realtime Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system (Donlon et al. 2012). The SST CCI product is the only OR that is both daily and provides an estimate of the observational uncertainty at its native resolution. Note that the uncertainty in the SST CCI gap-free product comprises of the observational error plus the error that arises from interpolation in space and time expressed as one standard deviation. In this study, we use the SST-CCI observational record, because a gap-free observational record appears to be most suitable for comparison with climate model data, which is typically gridded and without gaps. Other products, such as ERSSTv4, have uncertainty estimates which are however not explored in this study.

The observed SSTs are compared to seasonal coupled climate model predictions from the European Centre for Medium-range Weather Forecasts (ECMWF) forecasting system 4 (S4, Molteni et al. 2011). The hindcast considered spans the period 1981 - 2010 using 51 ensemble members with a horizontal resolution of $\sim$ 80 km in the atmosphere (T255) and with 1 degree resolution in the ocean. We focus on the El Niño Southern Oscillation (ENSO), which is the process that contributes most to seasonal predictability across the globe (Latif et al. 1998). The variability of ENSO is computed as the SST anomaly (with respect to the climatology 1981 - 2010) over the Niño3.4 region (170W - 120W; 5S - 5N, black box in Fig.1b). S4 is initialized every month and simulates the consecutive 7 months. Here, we only consider the prediction of summer months of the Northern Hemisphere (June-July-August, JJA) as they are the most difficult to predict from the predictions initialized in May (Barnston et al. 2012). The analysis is extended to global SSTs at a final stage.

Seasonal forecast skill is computed using the Pearson correlation of the ensemble mean prediction with the observations. Probabilistic properties that could be derived from the ensemble are omitted. The correlation is a popular skill metric of seasonal forecast quality (Doblas-Reyes et al. 2013; Scaife et al. 2014). It measures the linear relationship between the prediction and the observation across forecasts initialized at different dates and its square is equivalent to the re-calibrated mean square skill score (MSSS, Siegert et al. 2016a). This study focuses on the correlation coefficient only, keeping in mind that the observational uncertainty is equally relevant in probabilistic verification (Jolliffe 2017).

b. Propagation of uncertainties to climate model scales

The SST CCI analysis provides an estimate of the uncertainty at the resolution of the data (1/20 degree $\sim$ 6 km). This uncertainty at the grid point level has to be propagated to space-time averages used in the verification of seasonal predictions (typically monthly means and regional averages or coarser grid scales). In this study we are interested in the observational uncertainty of the average SST in the Niño3.4 region over a 30-day period. Since we can not expect observational errors to be uncorrelated in space and time, the usual formula to calculate the standard error of the mean does not apply. Instead, we have to take into account the finite correlation length (λ) and correlation time scale (τ) of the observational error.

Say we have an OR of the variable x with an accompanying observational uncertainty σ_x on a regular grid with grid spacing of Δx and Δt in space and time, respectively. We are consequently interested in the uncertainty $\sigma_{\bar{x}}$ of the space-time mean $\bar{x}$ in a configuration consisting of a domain with dimensions of M times N grid points and T time instances. We assume that the observational error $\varepsilon_{i,j,t}$ has an exponential correlation function

$$cor(\varepsilon_{i,j,t}, \varepsilon_{i',j',t'}) = exp\left(-\frac{\Delta x\sqrt{(i'-i)^2+(j'-j)^2}}{\lambda} - \frac{\Delta t|t'-t|}{\tau}\right) \tag{1}$$

while $i < M$, $j < N$, $t < T$ are indices of the data, such that the distances in space $\Delta x\sqrt{(i'-i)^2+(j'-j)^2}$ and time $\Delta t|t'-t|$ are scaled by the correlation lengths (Cressie 2015). The exponential function can be expanded for all possible distances (all possible values for i,j, and t) to form the covariance matrix Σ with dimension of all points in space and time (MNT). The uncertainty of $\bar{x}$ is consequently defined as,

$$\sigma_{\overline{x}} = \sqrt{w^T \Sigma w} \tag{2}$$

where w is the averaging vector with length of MNT values of $\frac{1}{MNT}$ or additional weighting values to account for the effective area of the grid points. The calculation of this expression requires enumeration over all pairs of grid points. The computational complexity of such an approach is $\mathscr{O}(M^2N^2T^2)$, which makes the calculation computationally unfeasible even for moderate domain sizes and time periods. To overcome the complexity, it is useful to assume a constant observational uncertainty within the domain ($\hat{\sigma}_x$). Since many points in space and time share the same distances (in space and time) one can formulate the following analytical solution (following the derivations described in Appendix A),

$$\sigma_{\overline{x}} = \frac{\hat{\sigma}_x}{MNT} \sqrt{(T + 2S_T)(MN + 2NS_M + 2MS_N + 4S_{MN})} \tag{3}$$

where the S terms describe the exponential decay in all dimensions,

$$\begin{aligned}
S_M &= \sum_{i=1}^{M-1} (M-i) e^{\frac{-i\Delta x}{\lambda}} \\
S_N &= \sum_{j=1}^{N-1} (N-j) e^{\frac{-j\Delta x}{\lambda}} \\
S_T &= \sum_{t=1}^{T-1} (T-t) e^{\frac{-t\Delta t}{\tau}} \\
S_{MN} &= \sum_{i=1}^{M-1} \sum_{j=1}^{N-1} (M-i)(N-j) e^{\frac{-\Delta x \sqrt{i^2+j^2}}{\lambda}}
\end{aligned}$$

The computational complexity is only $\mathscr{O}(M+N+T+MN)$ which allows us to efficiently propagate uncertainty to different length scales. An alternative approach is presented in Appendix B in case the assumption of a constant σ_x is weakly justified due to continental boundaries or strong inhomogeneity of σ_x in the space-time domain. The approach relies on generating random fields from Σ which are averaged for the space-time domain using a Monte Carlo approach. This solution is also sufficiently efficient to propagate observational uncertainty and explore the uncertainty related to the length and time scales. Note that the Monte-Carlo approach is orders of magnitude faster than the enumeration in equation 2 due to efficient algorithms based on Fourier transformations (Schlather et al. 2015).

It is useful to elaborate on equation 3 using practical examples for better understanding. Observational errors are traditionally classified into random and systematic errors (Povey and Grainger 2015). Errors such as sensor noise, which are uncorrelated in time and space, reduce with averaging with the square root of the sample size ($\sqrt{MNT}$) following the law of large numbers. Random errors are analogous to zero correlation scales ($\lambda = \tau = 0$) which yields zero for the S terms below the square root in equation 3 leaving $\sqrt{MNT}$ in the denominator. For locally systematic errors due to e.g. weather systems ($\lambda, \tau > 0$) the S-terms grow and therefore can be understood as the correction factor of the law of large numbers. If the errors are globally systematic due to e.g. errors in the retrieval algorithm, the length scales become infinitely large ($\lambda = \tau = \infty$) and the expression below square becomes $M^2N^2T^2$. The uncertainty does in this case not decrease $\sigma_{\overline{x}} = \hat{\sigma}_x$. The SST CCI provides the differentiated uncertainty components for non-gap filled data products (L3 products) with an accompanying tool for the propagation. In the gap-filled (L4) product these uncertainties can no longer be retained as the correlation structure is unknown after interpolation. In this case approximate length scales have to be used.

c. Inference of uncertainty from different observational references

An alternative way to determine the uncertainty in ORs is to infer it from the spread between available ORs for a given space-time mean (Martin et al. 2012). This can be done by assuming that different ORs are equally probable. This assumption is known to be flawed given that the quality of ORs differ (Massonnet et al. 2016). Martin et al. (2012) find that using ensembles of different SST products the resulting uncertainty is not robust (underestimated by a third in their analysis). However, this approach has been and remains the most adopted practice in the modelling community (e.g. Bellprat et al. 2012; Gómez-Navarro et al. 2012; Sunyer Pinya et al. 2013; Reichler and Kim 2008). It is therefore important to bridge to this practice. An advantage of the approach is that an ensemble of structurally different ORs allows to account for structural uncertainties in the retrieval algorithms (Thorne et al. 2005). The different ORs can consequently be understood as an ensemble of opportunity from which $\sigma_{\overline{x}}$ can be estimated. This approach fits naturally with data sets that systematically explore parameter choices using an ensemble approach (Morice et al. 2012). More sophisticated inference methods include parameters that account for structural differences in the ORs and estimate $\sigma_{\overline{x}}$ using the triple-collocation approach (Stoffelen 1998; Gruber et al. 2016) or Bayesian inference (Siegert et al. 2016b). In this study we use only the standard deviation between different ORs as a comparison for the uncertainty propagation.

3. Results

a. Uncertainty in the observed El Niño Southern Oscillation

The seasonal forecast capability of ECMWF S4 and the different ORs are summarized in figure 1. The time-series show the evolution of Niño3.4 SSTs for both the ensemble mean forecast (from which the correlation skill is determined) and the individual members. The time series length is constrained by the length of SST CCI, which spans the period 1992-2010. S4 has a high ensemble mean forecast skill shown here for the month of June (~ 0.9 correlation) and the ensemble range usually encompasses the estimates from the ORs. The ORs cluster and are closer to each other than to the model result, yet discrepancies between the ORs are visible.

The lower panel of figure 1 shows the observational uncertainty (σ_x) provided by SST CCI at a specific time instance (1st of June 2000). The variability of the spatial σ_x reaches one order of magnitude globally (not shown). Daily variations are negligible during the summer months but the uncertainty within the Niño3.4 region varies with a factor of three as denoted by the black box in figure 1b. Assuming constant uncertainty yields $\hat{\sigma}_x$=0.22 with a low standard deviation in space and time ($\pm$ 0.001 K) due to the temporal stability. The implications of the notable changes in the OR uncertainty in Niño3.4 is explored later in this section. In order to know $\sigma_{\overline{x}}$ for the monthly and spatial SST average in the Niño3.4 domain, we need to propagate $\hat{\sigma}_x$ to its space-time average.

The assumption of constant observational uncertainty greatly facilitates the propagation and allows to formulate the analytical solution as in equation 3. The solution suggests that the uncertainty propagates as a function of the ratio between the size of the space-time domain and the correlation length, independently of the data spacing ($\Delta x, \Delta t$), and the number of data points (MNT). This allows to present the propagation as a look-up graph (Fig.2) that is independent of the application. To describe this ratio we define spatial and temporal degrees of freedom (d.o.f) as the number of times that the correlation scale fits into the domain size. The spatial d.o.f is defined as $\frac{MN\Delta x^2}{\lambda^2}$ and the temporal d.o.f. as $\frac{T\Delta t}{\tau}$. A correlation time scale of 5 days is in this sense equal to 6 temporal d.o.f for a monthly average, while a length scale of 100 km would correspond to 100 d.o.f for a region of 1000km by 1000 km. The reader will note that spatial or temporal d.o.f should not be misinterpreted as effective sample sizes with which the standard deviation can be scaled. As shown in equation 3 the correction term is more complicated. To make the propagation general in the physical space, the graph is further shown for unit observational uncertainty ($\sigma_x = 1$). The resulting standard deviation of the space-time mean (y-axis) can consequently be understood as the propagation factor with which the average observational uncertainty ($\hat{\sigma}_x$) of the data needs to be multiplied.

The SST CCI reports correlation lengths for errors of 100 km in space and a time scale of one day for the locally systematic errors in single sensor L3 products. These represent scales associated with small synoptic systems and the coverage of the satellite (revisiting the same location every two days). We take here this estimate as a first guess, bearing in mind that these length scales do not take into account the uncertainty introduced from the interpolation in space and time. Taking the case of the monthly Niño3.4 domain the scales are equivalent to 30 temporal d.o.f. and 320 spatial d.o.f (the Niño3.4 regions covers 4000 km x 800 km). The resulting standard deviation of the space-time mean yields $\sigma_{\overline{x}}$= 0.007 K (the propagation factor is 0.03). This estimate is arguably too small and indicates that systematic uncertainties operating at larger scales are present. We consider therefore additionally scales associated with large synoptic systems of λ = 1000 km and τ= 10 days. The resulting estimate yields $\sigma_{\overline{x}}$ = 0.076 K.

The two estimates of monthly Niño3.4 SST uncertainties are compared in figure 3 with the standard deviations obtained from the four different ORs. The standard deviation from a sample of four points is highly uncertain and hence a distribution obtained from all individual years and the months (May-August) are shown as a histogram in figure 3. The propagated uncertainties from SST CCI are at the lower tail of uncertainty estimates, yet the estimate using large synoptic scales is consistent with the comparison of the different ORs for summer Niño3.4 SSTs (approximately $\sigma_{\overline{x}}$= 0.1 K). Differences between ORs can be substantially larger as seen in figure 3. Note that the two alternative estimates do not represent the same quantity as discussed in section 2c and are therefore not expected to agree entirely. The former is a self-consistent estimate of uncertainty in the SST CCI product, the latter is an estimate of the uncertainty collectively among the ORs. However, the comparison indicates that correlation scales associated with larger synoptic scales are reflecting the uncertainty of the Niño3.4 SSTs more realistically and might still underestimate the uncertainty Martin et al. (2012).

The propagated estimate assumes that the uncertainty is constant in space and time over the domain of interest, and that the spatial and temporal correlations decay exponentially with constant decorrelation parameters. The correlation function needs not necessarily to be exponential. The exponential function in equation 1 can be replaced by a different correlation function that is separable into the product of a temporal component and an isotropic spatial component with constant parameters. The assumption of constant observational variance used in figure 2 appears very restrictive, and seems to defeat the purpose of an observational data set that aims to resolve observational uncertainty in space and time. However, we have found by producing large samples from known distributions that the error due to the constant variance assumption is very small as long as the observational variance does not change too much over space and time in the domain of interest. In particular, we have analysed the observational error of Nino3.4 monthly average SST by sampling 1000 error fields 1) using the spatially and temporally varying observational error standard deviations provided in the data set (with much reduced spatial resolution), and 2) replacing all error standard deviations by their space-time mean, i.e. simulating under a constant error variance assumption. The analytical expression yields an observational error standard deviation of 0.0767 K. The 1000 simulated error fields with varying variances have standard deviation of 0.0766 K and the 1000 simulated error fields with constant variances have standard deviation of 0.0765 K. This result shows that analytical and simulated results agree when using 1000 Monte-Carlo simulations, and that the difference between varying and constant error variances is negligible (at least in this example).

b. Observational uncertainty in verification of seasonal sea-surface temperature forecasts

Having assessed the uncertainty in observed Niño3.4 SSTs, it is crucial to understand how important the uncertainty is in practice compared to other sources of uncertainty in forecast verification. There are three sources of uncertainties when dealing with the assessment of seasonal forecast skill: (1) a sample uncertainty due to the limited number of retrospective predictions or limited OR record length over which the skill is evaluated, (2) a sample uncertainty due to a limited ensemble size used to compute the ensemble-mean forecast often constrained by limited computational resources, (3) and an uncertainty due to the uncertainties in OR itself. Note that other uncertainties in the comparison of models and observations such as the unpredictable internal variability or the uncertainty due to model inadequacy (Notz 2015) are not uncertainties of the prediction skill, but part of the forecast error that the skill itself aims at measuring.

While uncertainties from (1) and (2) are commonly assessed (Ferro 2014; Scaife et al. 2014; Siegert et al. 2016b) the observational uncertainty remains an overlooked problem and formal concepts to include observational uncertainty in deterministic verification metrics are lacking (for probabilistic metrics approaches, different have been presented; Candille and Talagrand 2008; Jolliffe 2017). Here we explore impact of OR uncertainty on the correlation by generating an ensemble of observations. This is far from trivial (Povey and Grainger 2015) and proper ensemble generation is only possible at the level of the algorithm used to generate an ORs. However, at the user level the uncertainty estimate provided by CCI can be used to perturb the analysis using Gaussian random noise or using the different ORs as an ensemble of opportunity by resampling the ORs in each specfic year.

The impact of the observational uncertainty on the correlation skill of Niño3.4 SSTs is illustrated in figure 4 in comparison to the sampling uncertainties. The sample uncertainties are assessed by resampling the ensemble members of the forecast prior to computing the model ensemble mean and resampling the years in the verification period, both with replacement. An ensemble size of 10 members is used, which represents the typical ensemble size used in non-operational climate prediction hindcasts (Doblas-Reyes et al. 2013). The total uncertainty is estimated by sampling jointly all sources (1-3) using the alternative ORs as an estimate of the observational uncertainty. Note that the seamingly increased skill in July in comparison to June is an artifact of the limited period considered (1992 - 2010). For longer periods the forecast skill decreases monotonically as the model departs from the initialization date (May 1st).

The observational uncertainty (green area) contributes about 20% in the summer months and 50% in the first month after the initialisation with similar amplitudes for both observational ensemble approaches considered. The observational ensemble using the CCI uncertainty estimate tends to reduce the skill since adding observational error reduces the correlation (Massonnet et al. 2016). The total source of uncertainty increases with time and reaches a range of 0.7 - 0.95 correlation. The ensemble size uncertainty (orange area) remains overall small with 10 members as each member retains a strong signal over the Niño3.4 region. The record length of SST CCI is overall the largest source of uncertainty (blue area). Expanding the record length of SST CCI beyond the current 20 years might hence reduce the verification uncertainties more efficiently than current efforts to reduce the observational uncertainties for the Niño3.4 region. The sum of all three sources of uncertainties is clearly larger than the total uncertainty obtained by jointly sampling the uncertainty due to non-linear interactions of the terms. In the supplementary information (Fig. S1) we show that the qualitative conclusions drawn are also valid for varying ensemble sizes and record lengths.

The example gives a regionally limited perspective and the focus is expanded to a global view in figure 5 for the month of August by comparing the relative contribution of each uncertainty source with respect to the sum of all sources. The uncertainty related to the length of the SST record dominates almost everywhere except in the poles. The record length uncertainty is particularly large in regions of high interannual variability. The observational uncertainty, sampled using the CCI uncertainty estimate, is the dominant source of uncertainty over the polar regions and contributes also in various other regions up to 40%. The ensemble size uncertainty is the largest over the extratropical North Pacific and North Atlantic. The SSTs over these regions are primarily forced by the atmospheric flow at seasonal time scales (Cayan 1992) and therefore subject to the atmospheric internal variability which is large in the extratropical Northern Hemisphere. A large ensemble size is therefore required in this region to reduce the effect of the internal variability in the ensemble mean in this region (Scaife et al. 2014).

Finally, it is important to take into account that observational errors not only increase the verification uncertainty but also have systematic effects on the prediction skill. Uncertainties in a reference lower the correlation skill (Massonnet et al. 2016), similarly as a limited ensemble size leads to systematically lower correlation (Ferro 2014; Scaife et al. 2014). This reduction in correlation skill can be estimated by dividing the sample correlation by the correction for attenuation (Spearman 1904),

$$R = \frac{\sigma_o^2 - \sigma_x^2}{\sigma_o^2}, \tag{4}$$

where σ_o is the total interannual standard deviation of the ORs and σ_x the observational uncertainty. The reference variability is hence attenuated for the observational uncertainty without altering the co-variance between the model and the reference. Corrections for probabilistic measures have also recently been proposed (Ferro 2017). The resulting increase in the correlation skill of ECMWF S4 global SSTs is shown in figure 6. The skill increases in many regions up to 0.2 and beyond, in agreement with the regions where the uncertainty increases most (figure 5, first panel). In the poles and also regions in the southern Ocean the observational uncertainty is larger than the interannual variability of the OR and hence no attenuation can be calculated.

4. Discussion and conclusions

Just like climate model predictions, observational references (ORs) are subject to uncertainties. These uncertainties are usually disregarded in the verification of seasonal forecasts or the evaluation of climate models in general. The common assumption that limitations of the models dominate the observational uncertainty persists and the role of OR limitations is therefore often seen as minor. These assumptions are rarely assessed and individual studies suggest that observational uncertainties might be larger than anticipated (e.g. Addor and Fischer 2015; Prodhomme et al. 2016; Massonnet et al. 2016). Formal concepts of how to account for observational uncertainties provided by ORs in climate model evaluation are, however, still scarce.

In this study, we present a step forward to narrow this gap by presenting simple ways to propagate observational uncertainties to space-time means, a necessary step in forecast verification where the model and OR spatial and temporal resolution do not match each other. The solution described is independent of the data structure and is illustrated as a "look-up" graph from which propagated uncertainties can be readily estimated. The solution assumes a constant observational uncertainty in the region and under the period considered for the space-time average and an alternative Monte-Carlo simulation approach is suggested if this assumption is weakly justified. Propagated observational uncertainties from the SST CCI product are consistent with differences in different ORs over the Niño3.4 region, yet the latter tends to be larger. Using the different ORs as complementary estimates and the propagated SST CCI uncertainty we find that the observational uncertainty contributes fundamentally to the forecast skill assessment of seasonal predictions of SSTs. Particularly at high latitudes, the observational uncertainty can dominate over other sources of verification uncertainties. However, over most regions, the largest uncertainty in seasonal forecast quality originates from the limited period over which the hindcasts are evaluated.

The observational uncertainty is also shown to systematically reduce the correlation skill by up to 0.2 correlation and beyond. Accounting for the increased verification uncertainty and systematic underestimation of skill should become a future practice in order to fully understand the utility of a seasonal forecasts.

Acknowledgments. We thank C. Merchant (University of Reading) and N. Rayner (UK MetOffice) for their fundamental advise in understanding observational uncertainties. We are also indebted to C. Ferro (University of Exeter) for the indication on the attenuation of error in correlation skill. This work has been supported by EU Seventh Framework Programme FP7 projects SPECS (GA 308378), EUCLEIA (GA 607085) and the European Space Agency Living Planet Fellowship Programme under the project VERITAS-CCI.

References

Addor, N., and E. M. Fischer, 2015: The influence of natural variability and interpolation errors on bias characterization in RCM simulations. *Journal of Geophysical Research: Atmospheres*, **120**, D022 824.

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002-11: is our capability increasing? *Bulletin of the American Meteorological Society*, **93**, 631–651.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär, 2012: Exploring perturbed physics ensembles in a regional climate model. *Journal of Climate*, **25**, 4582–4599.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **134**, 959–971.

Cayan, D. R., 1992: Latent and Sensible Heat Flux Anomalies over the Northern Oceans: The Connection to Monthly Atmospheric Circulation. *Journal of Climate*, **5**, 354–369.

Cressie, N., 2015: *Statistics for spatial data*. John Wiley & Sons.

Dee, D., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, **4**, 245–268.

Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer, 2012: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sensing of the Environment*, **116**, 140–158.

Ferro, C., 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1917–1923.

Ferro, C., 2017: Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society, in review*.

Flato, G., and Coauthors, 2013: *Evaluation of Climate Models*, book section 9, 741–866. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Gómez-Navarro, J. J., J. P. Montávez, S. Jerez, P. Jiménez-Guerrero, and E. Zorita, 2012: What is the role of the observational dataset in the evaluation and scoring of climate models? *Geophysical Research Letters*, **39**, L054 206.

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner, 2016: Recent advances in (soil moisture) triple collocation analysis. *International Journal of Applied Earth Observation and Geoinformation*, **45**, 200–211.

Hartmann, D., and Coauthors, 2013: *Observations: Atmosphere and Surface*, chap. 2, 159–254. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Huang, B., and Coauthors, 2015: Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *Journal of Climate*, **28**, 911–930.

Jolliffe, I. T., 2017: Probability forecasts with observation error: what should be forecast? *Meteorological Applications*, **24**, 276–278.

Kennedy, J. J., 2014: A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Reviews of Geophysics*, **52**, 1–32.

Knutti, R., 2008: Should we believe model predictions of future climate change? *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, **366**, 4647–4664.

Latif, M., and Coauthors, 1998: A review of the predictability and prediction of ENSO. *Journal of Geophysical Research: Oceans*, **103**, 14 375–14 393.

Liu, W., and Coauthors, 2015: Extended reconstructed sea surface temperature version 4 (ERSST. v4): part II. Parametric and structural uncertainty estimations. *Journal of Climate*, **28**, 931–951.

Martin, M., and Coauthors, 2012: Group for High Resolution Sea Surface temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE). *Deep Sea Research Part II: Topical Studies in Oceanography*, **77**, 21 – 30.

Massonnet, F., O. Bellprat, V. Guemas, and F. J. Doblas-Reyes, 2016: Using climate models to estimate the quality of global observational data sets. *Science*, **354**, 452–455.

Merchant, C. J., and Coauthors, 2014: Sea surface temperature datasets for climate applications from phase 1 of the european space agency climate change initiative (sst cci). *Geoscience Data Journal*, **1**, 179–191.

Merchant, C. J., and Coauthors, 2017: Uncertainty information in climate data records from Earth observation. *Earth System Science Data Discussions*, **2017**, 1–28.

Molteni, F., and Coauthors, 2011: *The new ECMWF seasonal forecast system (System 4)*. European Centre for Medium-Range Weather Forecasts.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117**, D08 101.

Mudryk, L. R., P. J. Kushner, C. Derksen, and C. Thackeray, 2017: Snow cover response to temperature in observational and climate model ensembles. *Geophysical Research Letters*, **44**, 919–926.

Notz, D., 2015: How well must climate models agree with observations? *Phil. Trans. R. Soc. A*, **373**, 20140 164.

Povey, A. C., and R. G. Grainger, 2015: Known and unknown unknowns: uncertainty estimation in satellite remote sensing. *Atmospheric Measurement Techniques*, **8**, 4699–4718.

Prodhomme, C., L. Batt, F. Massonnet, P. Davini, O. Bellprat, V. Guemas, and F. J. Doblas-Reyes, 2016: Benefits of increasing the model resolution for the seasonal forecast quality in EC-Earth. *Journal of Climate*, **29**, 9141–9162.

Rayner, N., D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, **108**.

Reichler, T., and J. Kim, 2008: Uncertainties in the climate mean state of global observations, reanalyses, and the GFDL climate model. *Journal of Geophysical Research: Atmospheres*, **113**, D05 106.

Scaife, A., and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, **41**, 2514–2519.

Schlather, M., A. Malinowski, P. J. Menck, M. Oesting, and K. Strokorb, 2015: Analysis, Simulation and Prediction of Multivariate Random Fields with Package ”RandomFields”. *Journal of Statistical Software*, **63**, 1–25.

Siegert, S., D. B. Stephenson, O. Bellprat, M. Ménégoz, and F. J. Doblas-Reyes, 2016a: Detecting improvements in forecast correlation skill: Statistical testing and power analysis. *Monthly Weather Review*, **145**, 437–450.

Siegert, S., D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas, 2016b: A Bayesian Framework for Verification and Recalibration of Ensemble Forecasts: How Uncertain is NAO Predictability? *Journal of Climate*, **29**, 995–1012.

Spearman, C., 1904: The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, **15**, 72–101.

Stoffelen, A., 1998: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research*, **103**, 7755–7766.

Sunyer Pinya, M. A., H. J. D. Sørup, O. B. Christensen, H. Madsen, D. Rosbjerg, P. S. Mikkelsen, and K. Arnbjerg-Nielsen, 2013: On the importance of observational data properties when assessing regional climate model performance of extreme precipitation. *Hydrology and Earth System Sciences*, **17**, 4323–4337.

Thorne, P. W., D. E. Parker, J. R. Christy, C. A. Mears, P. W. Thorne, D. E. Parker, J. R. Christy, and C. A. Mears, 2005: Uncertainties in climate trends: Lessons from Upper-Air Temperature Records. *Bulletin of the American Meteorological Society*, **86**, 1437–1442.

LIST OF FIGURES

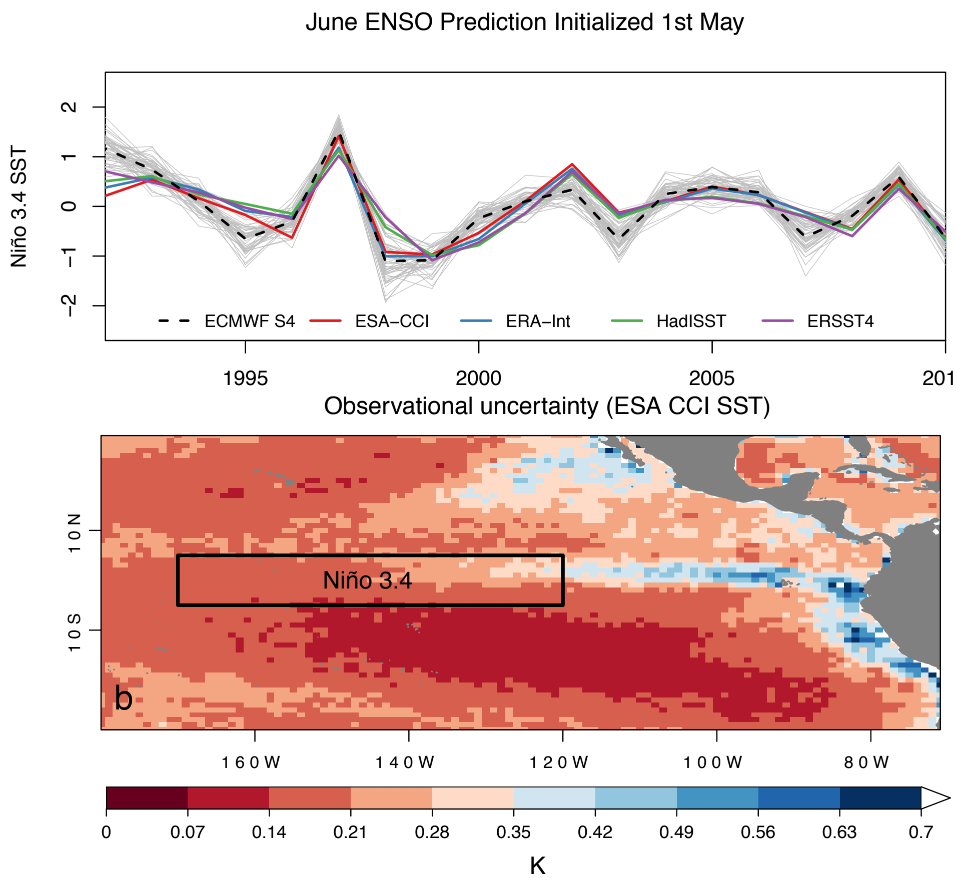

FIG. 1. a) June observations (solid lines) and seasonal forecast of ECMWF System 4 initialized in 1st May (dashed line shows the ensemble mean, gray lines the individual members) of Niño3.4 sea-surface temperature (SST) anomalies with respect to the climatology of 1992 - 2010. The time-series are shown only for the period where ESA SST CCIs is available. (b) Observational uncertainty (one standard deviation) of SST in the Niño3.4 region for the 1st June 2000.

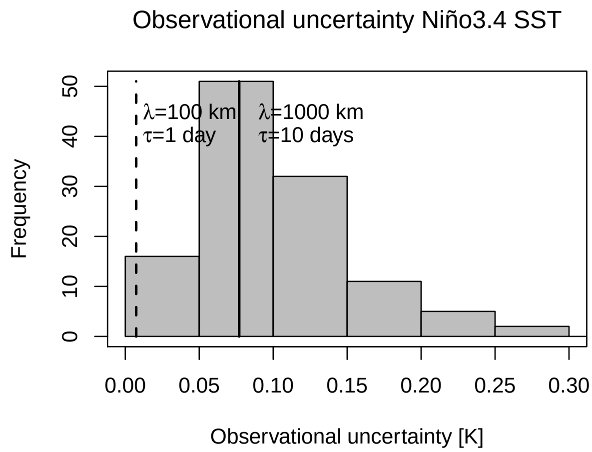

FIG. 2. Uncertainty propagation to space-time averages as a function of the correlation scales in space (x-axis) and time (different lines) for unit observational uncertainty σ_x=1. The correlation scales are expressed as degrees of freedom (d.o.f.) by computing the number of times the correlation scale fits in the space-time domain. The propagation is consequently independent of the data spacing and the number of data points. The aspect ratio of the spatial domain impacts the propagation. The mean distance between all possible pair of points in a square is smaller than in a strongly rectangular region as for instance the Niño3.4 region with aspect ratio of region of 1:5. The observational uncertainty therefore decreases stronger in non-rectangular regions as denoted by the different aspect ratios. The standard deviation of the space-time average serves as a propagation factor with which the observational uncertainty provided by the OR has to be multiplied. For example for 5 spatial and temporal d.o.f. the observational uncertainty reduces by a factor of 0.5. Mind the logarithmic scales of the axes.

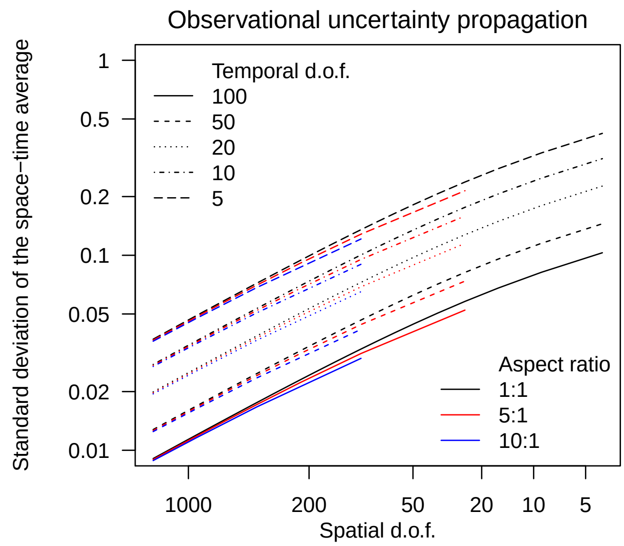

FIG. 3. Observational uncertainty of monthly Niño3.4 SSTs as propagated from SST CCI uncertainty estimates using the approach depicted in figure 2 and length scales associated with small (dashed line) and large (solid line) synoptic scales. The histogram shows the standard deviation between the four ORs in all years of the period 1981-2010 (only three ORs prior to 1992) during the months May - August as a comparison of observational uncertainty inferred from the data itself.

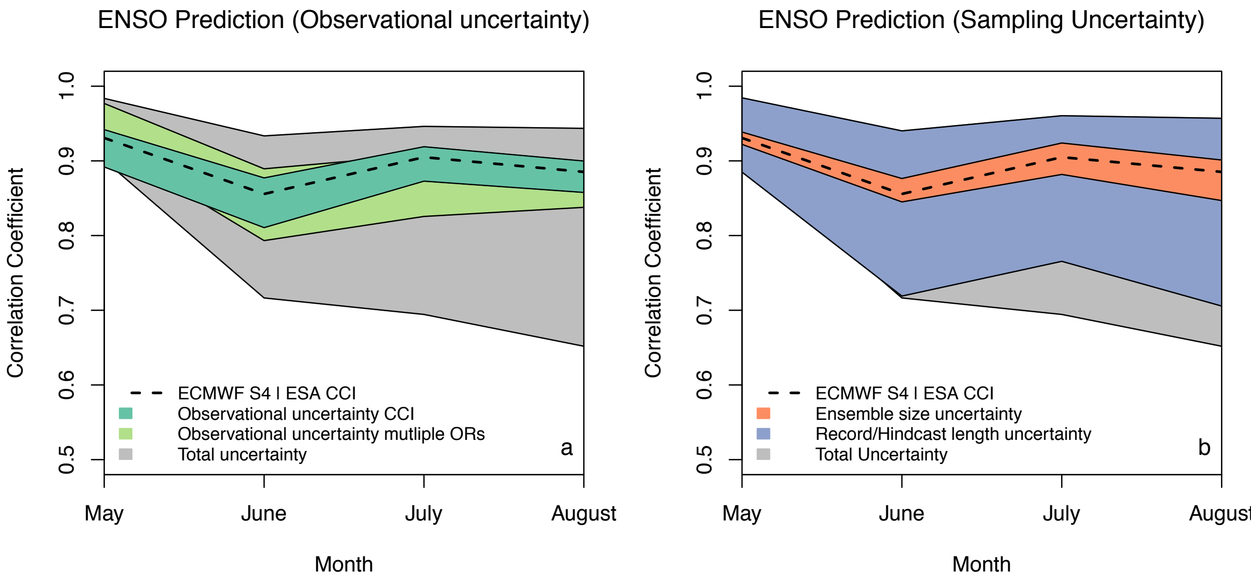

FIG. 4. Sub-seasonal to seasonal forecast skill of ECMWF S4 (10 members) with respect to SST CCI (dashed line). The areas show the 5-95% percentile range of the bootstrapped (10^6 samples) uncertainty sources around the sample correlation skill for (a) the uncertainty in the observations assessed using the SST CCI propagated uncertainty (λ = 1000 km and τ= 10 days) and the ensemble of different ORs and for (b) the sample uncertainty due to a limited ensemble size and record length of the SST CCI dataset. The grey area shows the total uncertainty obtained by resampling all sources at the same time.

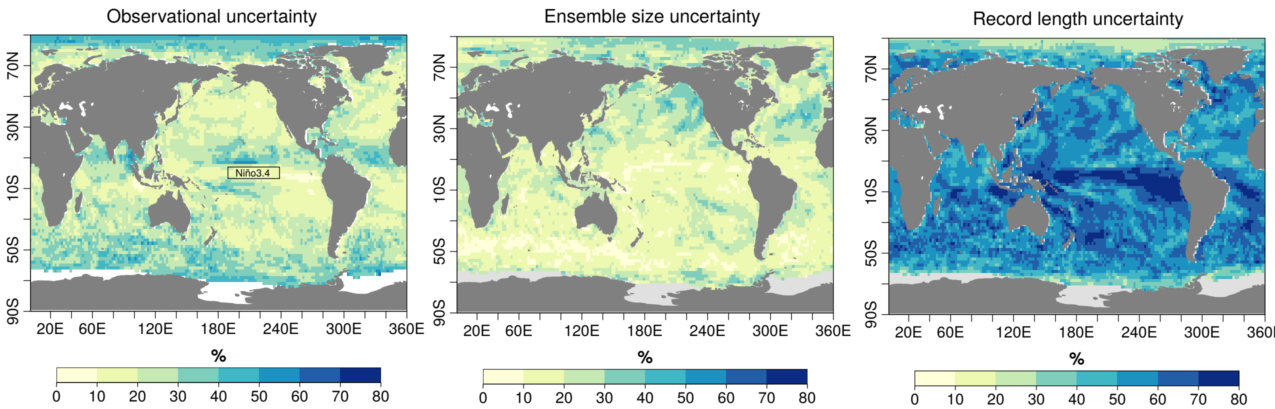

FIG. 5. Relative contribution of each source of uncertainty with respect to the sum of all sources. The relative contribution is calculated by the variance of the correlation after resampling one source divided by the sum of variances of all sources (instead of the total uncertainy due to interaction of the individual terms).

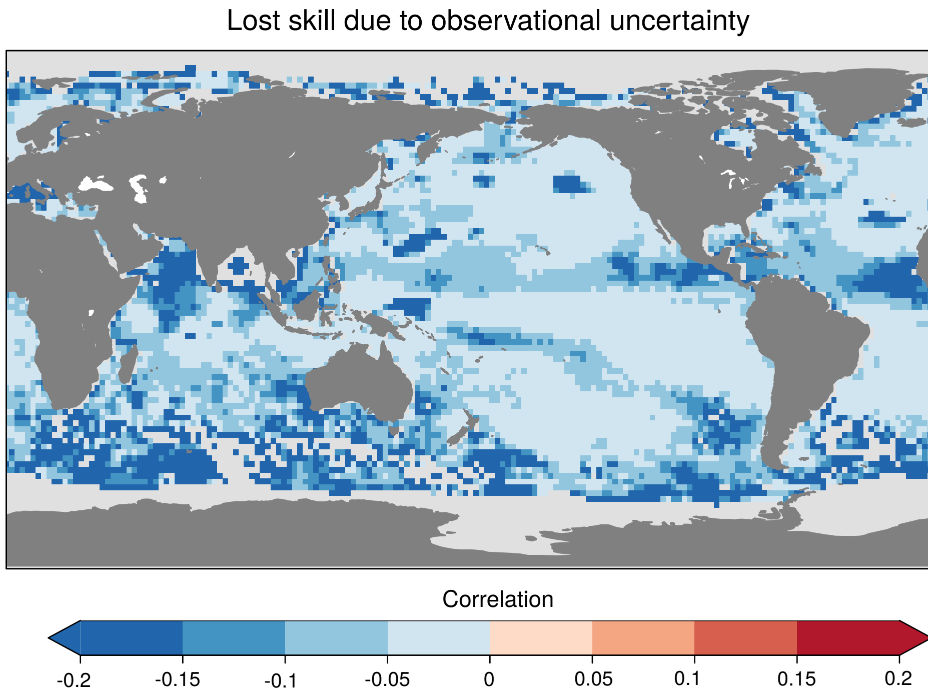

FIG. 6. Reduction of correlation skill in ECMWF S4 due to the observational uncertainty for the prediction of the month of August (initialized in 1st of May) estimated using the correction for attenuation (Spearman 1904). The observational uncertainty is estimated by propagating SST CCI uncertainties to monthly means in each grid-point. Grid-points in gray denote areas where the observational uncertainty is larger than the interannual variability of the SST CCI and where as a consequence no correction for attenuation can be calculated.