Detecting Improvements in Forecast Correlation Skill: Statistical Testing and Power Analysis

STEFAN SIEGERT

Exeter Climate Systems, University of Exeter, Exeter, United Kingdom

OMAR BELLPRAT AND MARTIN MÉNÉGOZ

Earth Sciences Department, Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS), Barcelona, Spain

DAVID B. STEPHENSON

Exeter Climate Systems, University of Exeter, Exeter, United Kingdom

FRANCISCO J. DOBLAS-REYES

ICREA, and Earth Sciences Department, Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS), Barcelona, Spain

(Manuscript received 25 January 2016, in final form 2 September 2016)

ABSTRACT

The skill of weather and climate forecast systems is often assessed by calculating the correlation coefficient between past forecasts and their verifying observations. Improvements in forecast skill can thus be quantified by correlation differences. The uncertainty in the correlation difference needs to be assessed to judge whether the observed difference constitutes a genuine improvement, or is compatible with random sampling variations. A widely used statistical test for correlation difference is known to be unsuitable, because it assumes that the competing forecasting systems are independent. In this paper, appropriate statistical methods are reviewed to assess correlation differences when the competing forecasting systems are strongly correlated with one another. The methods are used to compare correlation skill between seasonal temperature forecasts that differ in initialization scheme and model resolution. A simple power analysis framework is proposed to estimate the probability of correctly detecting skill improvements, and to determine the minimum number of samples required to reliably detect improvements. The proposed statistical test has a higher power of detecting improvements than the traditional test. The main examples suggest that sample sizes of climate hindcasts should be increased to about 40 years to ensure sufficiently high power. It is found that seasonal temperature forecasts are significantly improved by using realistic land surface initial conditions.

1. Introduction

Hindcast experiments are routinely generated to detect systematic biases of forecast systems, and to assess forecast quality. Hindcast data from a competing forecast system are often available, from either a low-resolution version of the same forecast system, the system of a competing forecast institution, or a simple statistical benchmark forecast. It is then of interest to address the question whether the forecast system at hand offers an improvement over the competitor. A very common measure of forecast skill is the (Pearson product moment) correlation coefficient between forecast and observations. To answer the question of whether the new forecast offers an improvement over a competitor, the difference in the correlation coefficient could be considered. Furthermore, in order to assess the robustness of an observed difference in correlation, some measure of uncertainty must be calculated.

As pointed out by Jolliffe (2007): "The value of a verification measure on its own is of little use; it also needs some quantification of the uncertainty associated

Corresponding author address: Stefan Siegert, Exeter Climate Systems, University of Exeter, Exeter EX4 4QF, United Kingdom. E-mail: s.siegert@exeter.ac.uk

DOI: 10.1175/MWR-D-16-0037.1

with the observed value'' (p. 637). Uncertainty quantification is important to distinguish genuine improvements in forecast skill from random sampling variability due to the finite hindcast samples. Jolliffe (2007) presents various statistical methods to quantify uncertainty in forecast skill and differences in forecast skill. DelSole and Tippett (2014) show that commonly used statistical tests for comparing skill of climate forecasts make the questionable assumption that the competing forecasts are independent. They show that this assumption can invalidate the test results, and suggest suitable alternatives.

The present paper complements Jolliffe (2007) and DelSole and Tippett (2014) by reviewing statistical methods that are directly applicable to testing for differences in correlation forecast skill, and by emphasizing the power of statistical tests to detect skill improvements. Section 2 briefly reviews the correlation coefficient, statistical hypothesis testing, and confidence intervals. Section 3 describes the most currently used hypothesis test for quantifying uncertainty of a correlation difference. A hypothesis test by Steiger (1980), and an approximate method to calculate confidence intervals by Zou (2007) are suggested as more appropriate methods for comparing correlation coefficients of two forecasts for the same set of observations. In section 4, the different statistical methods are applied to datasets of seasonal near-surface air temperature forecasts. The analyses provide detailed examples of how the different test statistics are calculated in practice. It is shown that the alternative tests indicate significant improvements in forecast skill where the traditional test does not. In section 5, the differences between the tests are assessed by analyzing their type-I error rates (the probability of falsely detecting an improvement) and their power (the probability of correctly detecting an improvement). It is shown that the traditional test can have a too low type-I error rate, and that the alternative test has higher power and thus increases the chance of detecting genuine improvements in forecast skill. Section 6 compares predictions of climate indices (ENSO and NAO) using model versions with different resolutions. Section 7 concludes the paper with a discussion and additional remarks.

2. Basic concepts

Assume two forecast systems—system A and system B—both of which make predictions about the same observable Y. A hindcast dataset generated by A and B for the same observation Y consists of a series of triplets $\{a_t, b_t, y_t\}$, where $t = 1, \ldots, n$. The Pearson product-moment correlation coefficient between the forecasts generated by system A and the observations is denoted by

$$r_{ay} = \frac{\sum_{t=1}^{n}(a_t - \overline{a})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{n}(a_t - \overline{a})^2}\sqrt{\sum_{t=1}^{n}(y_t - \overline{y})^2}}, \tag{1}$$

where $\overline{a}$ and $\overline{y}$ are the sample averages of $a_1, \ldots, a_n$ and of $y_1, \ldots, y_n$, respectively (Wilks 2011, section 3.5.2). There are at least 13 possible interpretations of correlation (Rodgers and Nicewander 1988). For forecast verification, one of the most relevant interpretations is that the squared correlation is equal to the mean squared skill score of the linearly recalibrated forecasts; a derivation can be found in the appendix, and also in Murphy and Epstein (1989). Because of the implicit linear recalibration, the correlation coefficient is insensitive to systematic biases. To also account for biases, the uncentered anomaly correlation can be used (Wilks 2011, section 8.6.4.) instead of Eq. (1), but then the statistical tests of section 3 do not apply.

To quantify uncertainty of the correlation coefficient, it is often assumed that the hindcast dataset is a random sample from an infinite population of forecasts and observations. The sample correlations r_{ay} and r_{by} are interpreted as imprecise, noisy measurements of the unknown population correlation coefficients ρ_{ay} and ρ_{by}. When the population correlations of two forecast systems are to be compared, two questions are often of interest: "Is there any improvement?" and "How big is the improvement?" The question of whether or not there is an improvement is a *testing problem* that can be addressed by significance testing. The question of how big an improvement is, is an *estimation problem* that can be addressed by confidence intervals.

Tests for improvements in correlation skill assume null hypotheses such $H_0: \rho_{ay} = 0$ (no improvement over zero skill) or $H_0: \rho_{by} = \rho_{ay}$ (no improvement of system B over system A). To test H_0, a test statistic T is calculated, which is a function of the hindcast data, and whose sampling distribution is known if H_0 is true. Based on the observed value of T, say $\hat{T}$, the p value is calculated (i.e., the chance of observing a value of T that is more extreme than $\hat{T}$ when H_0 is true). A low p value such as $p < 0.05$ implies that the observed $\hat{T}$ is a relatively unlikely value if H_0 were true, which is interpreted as evidence against H_0. Confidence intervals are used as an interval estimate of the magnitude of the unknown population correlation ρ_{ay}, or correlation difference $\rho_{by} - \rho_{ay}$. A 95% confidence interval has a nominal frequency of 95% of covering the unknown population quantity (i.e., if confidence intervals were calculated repeatedly for data drawn from the population, the interval would cover the population value 95% of the time). (The usual disclaimer applies: the p value is not the probability that H_0 is true, and the confidence coefficient

of 95% is not the probability that the confidence interval covers the population value.) Confidence intervals can also be used for hypothesis testing. If a confidence interval fails to overlap a value of interest such as 0, this can be taken as sufficient evidence to reject the null hypothesis $H_0: \rho_{ay} = 0$, say. If we decide to reject H_0 if the p value is smaller than 0.05, we accept a 5% chance of mistakenly rejecting a true H_0 (i.e., of committing a type-I error). The same chance of a type-I error results if we rejected H_0 whenever the 95% confidence interval does not overlap zero. Failure to reject a false H_0 is known as a type-II error, and the probability of correctly rejecting a false H_0 is called the power. In forecast verification, statistical power of a test quantifies our ability to correctly detect improvements in forecast skill. For deeper treatment of the concepts outlined in this section we refer the reader to the statistical climatology literature, especially Von Storch and Zwiers (2001) and Wilks (2011).

3. Methods

In this section we summarize statistical methods for hypothesis testing and confidence intervals of correlation coefficients. In the atmospheric sciences literature, methods to quantify uncertainty in a single correlation coefficient are well known. However, statistical methods to calculate hypothesis tests and confidence intervals for the difference between correlation coefficients are less well known.

a. *Hypothesis tests and confidence intervals for a single correlation coefficient*

Under the null hypothesis of zero correlation, $H_0: \rho_{ay} = 0$, the test statistic

$$T_0 = r_{ay}\left(\frac{n-2}{1-r_{ay}^2}\right)^{1/2} \tag{2}$$

has a Student's t distribution with $n - 2$ degrees of freedom (Von Storch and Zwiers 2001, section 8.2.3). The test assumes that the hindcast data are independently and identically normally distributed. The term T_0 tends to fall far into the upper or lower tail of the t distribution if r_{ay} is close to $+1$ or -1, respectively. A two-sided test at significance level α would thus reject the null hypothesis $H_0: \rho_{ay} = 0$ if T_0 is either smaller than the $(\alpha/2)$ quantile or larger than the $(1 - \alpha/2)$ quantile of the t distribution.

Confidence intervals for a correlation coefficient can be calculated based on the Fisher transformation of r_{ay} (also called the z transform), defined by

$$z_{ay} = \frac{1}{2}\log\left(\frac{1+r_{ay}}{1-r_{ay}}\right) = \operatorname{atanh}(r_{ay}) \tag{3}$$

(Von Storch and Zwiers 2001, section 8.2.3). For data that are identically and independently normally distributed, the Fisher transformation of r_{ay} is approximately normally distributed with mean $\operatorname{atanh}(\rho_{ay})$ and variance $(n - 3)^{-1}$. A 95% confidence interval for r_{ay} is thus given by $[l, u]$, where

$$l = \tanh\left(z_{ay} + \frac{Z_{0.025}}{\sqrt{n-3}}\right), \tag{4a}$$

$$u = \tanh\left(z_{ay} + \frac{Z_{0.975}}{\sqrt{n-3}}\right), \tag{4b}$$

and where Z_p denotes the p quantile of the standard normal distribution (e.g., $Z_{0.025} = -1.96$).

The Fisher transformation can be used to assess the difference between two independent correlation coefficients. Under the null hypothesis $\rho_{by} = \rho_{ay}$, z_{ay} and z_{by} have the same normal distribution, with variance $(n - 3)^{-1}$. Under the assumption that z_{ay} and z_{by} are statistically independent, their difference $z_{by} - z_{ay}$ has a normal distribution with mean zero and variance $2(n - 3)^{-1}$. This leads to a hypothesis test of the null hypothesis $\rho_{by} - \rho_{ay} = 0$ where the test statistic

$$T_1 = (z_{by} - z_{ay})\sqrt{\frac{n-3}{2}} \tag{5}$$

has a standard normal distribution. This test is presented in Jolliffe and Stephenson (2012, section 5.4.4) and has been used in the climate literature to assess correlation differences between forecasting systems; examples include Keenlyside et al. (2008), Du et al. (2012), Doblas-Reyes et al. (2013b), and Pepler et al. (2015). We will show in section 5 that the test based on T_1 has a serious shortcoming, namely, the assumption of independence between z_{ay} and z_{by}. If two forecasts are made for the same observation, they are likely to be correlated with one another, and therefore any statistics that depend on the forecasts, such as correlations r_{ay} and r_{by} (or their Fisher transformations), are likely to be correlated as well; see also DelSole and Tippett (2014). We will next review alternative methods for uncertainty quantification of correlation differences that improve the test based on T_1 by taking into account the correlation between forecasts.

b. *Testing and estimating the difference of two overlapping correlations*

Two correlation coefficients that share a common variable such as r_{ay} and r_{by} are said to be *overlapping* (Zou 2007). The following test presented by Steiger (1980) [based on results from Williams (1959)] tests equality of overlapping correlations (i.e.,

H_0: $\rho_{by} - \rho_{ay} = 0$), taking into account that the forecasts generated by systems A and B can be correlated with one another. We define the auxiliary quantity R as

$$R = (1 - r_{ay}^2 - r_{by}^2 - r_{ab}^2) + (2r_{ay}r_{by}r_{ab}), \tag{6}$$

where r_{ab} is the correlation between the forecasts generated by systems A and B for the same observations. (Here R is the determinant of the 3×3 sample correlation matrix of forecasts and observations.) The test statistic

$$T_2 = (r_{by} - r_{ay}) \sqrt{\frac{(n-1)(1+r_{ab})}{2\left(\frac{n-1}{n-3}\right)R + \frac{1}{4}(r_{ay} + r_{by})^2(1 - r_{ab})^3}} \tag{7}$$

has a Student's t distribution with $n - 3$ degrees of freedom under the null hypothesis of zero correlation difference. If $r_{by} - r_{ay} > 0$ (i.e., if forecast B has higher correlation than forecast A), the test statistic T_2 becomes large, and will fall far into the upper tail of the corresponding t distribution. A one-sided test at significance level α would thus compare T_2 to the $(1 - \alpha)$ quantile of the t distribution with $n - 3$ degrees of freedom, and reject the null hypothesis of zero correlation difference if T_2 exceeds this critical value. Such a one-sided test can be used to test whether forecast system B offers improved forecasts compared to forecast system A.

Zou (2007) provides an approximate method to calculate confidence intervals for a difference between two overlapping correlation coefficients. First calculate the auxiliary quantity:

$$c_{ab} = \frac{\left(r_{ab} - \frac{1}{2}r_{ay}r_{by}\right)(1 - r_{ay}^2 - r_{by}^2 - r_{ab}^2) + r_{ab}^3}{(1 - r_{ay}^2)(1 - r_{by}^2)}, \tag{8}$$

which approximates the correlation between r_{ay} and r_{by}. Then calculate $(1 - \alpha) \times 100\%$ confidence intervals (l_a, u_a) for r_{ay} and (l_b, u_b) for r_{by}, using Eq. (4). An approximate $(1 - \alpha) \times 100\%$ confidence interval (L, U) for the correlation difference $r_{by} - r_{ay}$ is then given by

$$\begin{aligned} L &= (r_{by} - r_{ay}) - \sqrt{(r_{by} - l_b)^2 + (u_a - r_{ay})^2 - 2c_{ab}(r_{by} - l_b)(u_a - r_{ay})}, \\ U &= (r_{by} - r_{ay}) + \sqrt{(u_b - r_{by})^2 + (r_{ay} - l_a)^2 - 2c_{ab}(u_b - r_{by})(r_{ay} - l_a)}. \end{aligned} \tag{9}$$

Note that high values of c_{ab} lead to narrow confidence intervals. The methods by Steiger (1980) and Zou (2007) are approximations: they assume that the data are normally distributed, and that the sample size is sufficiently large.

4. Application to seasonal near-surface air temperature forecasts

a. Description of the data

A comparison of seasonal climate forecasts serves here as a practical example. Forecasts of average summer (JJA) near-surface air temperatures are initialized on 1 May for the $n = 17$ yr from 1993 to 2009. Such small sample sizes are caused by computational constraints and limited observation data, and lead to large uncertainties in verification measures (Siegert et al. 2016).

The hindcast experiment addresses the effect of initializing the land surface conditions. A more realistic initialization of the land surface conditions is expected to have particular impact on prediction of summer temperatures over landmasses. One forecast, denoted forecast A, was generated by using the same climatological land surface conditions to initialize the forecast in each year. We computed the climatology of surface parameters (soil moisture and temperature at all soil levels, and the albedo, depth, density, and temperature of the snow layer) by taking their 1993–2009 averages in a window of 10 days centered around the initialization date 1 May, using data from the ERA-Interim/Land global reanalysis dataset (Balsamo et al. 2015). The 10-day window ensures a robust estimate of the climatology. The other set of forecasts, denoted forecast B, were initialized with the actual land surface parameters on the initialization date in the respective year, taken from the ERA-Interim/Land dataset. All model hindcasts were carried out with the global climate system model EC-Earth3 (Hazeleger et al. 2012), which has been widely used for studying intraseasonal to multiannual predictability and climate projections (Doblas-Reyes et al. 2013a). Hindcasts are initialized with reanalysis data from Global Ocean Reanalysis and Simulations, version 1 (GLORYS2v1) for the ocean (Ferry et al. 2012), ERA-Interim reanalysis data for the atmosphere (Dee et al. 2011), ERA-Interim/Land data for the land surface (Balsamo et al. 2015), and sea ice initial conditions from Guemas et al. (2014). Each prediction is calculated as the mean over 10 ensemble members initialized by atmospheric singular vectors.

TABLE 1. Region specifications. The regions are also indicated in Fig. 2.

Region	Label	Coordinates of region corners
Central Europe	CEU	(45°N, 10°W) (48°N, 10°W) (61.32°N, 40°E) (45°N, 40°E)
Eastern Asia	EAS	(20°N, 100°E) (50°N, 100°E) (50°N, 145°E) (20°N, 145°E)
Northeastern Brazil	NEB	(20°S, 34°W) (20°S, 50°W) (0°, 50°W) (0°, 34°W)
Western Africa	WAF	(11.365°S, 20°W) (15°N, 20°W) (15°N, 25°E) (11.365°S, 25°E)

Surface temperature data from the ERA-Interim reanalysis were used as verifying observations.

We evaluate differences in correlation skill at each land grid point individually, and also for area averages. The area averages are calculated for four regions defined in the SREX special report of the IPCC (IPCC 2012). The region specifications are given in Table 1. These four regions are either in semiarid climates, where the land surface–atmosphere interactions play an important role for the energy balance, or for which land surface–atmosphere couplings were previously reported in the literature (Koster et al. 2004; Zhang et al. 2011; Bellprat et al. 2013). The time series plots of the area-weighted temperature averages for the four regions are shown in Fig. 1. It can be noted that forecasts and observations have negligible serial correlation.

b. Correlation analysis

We first provide a detailed example of how the various test statistics, p values, and confidence limits of section 3 are calculated. We use the time series of the central European (CEU) region (Fig. 1a) for illustration. The sample correlations of the forecasts with the observations are $r_{ay} = 0.56$ and $r_{by} = 0.80$. The sample size is $n = 17$, and sample correlation between the forecasts is $r_{ab} = 0.62$. The Fisher transformations of the correlations r_{ay} and r_{by} are $z_{ay} = 0.63$ and $z_{by} = 1.10$, from which we calculate the central 95% confidence intervals $(l_a, u_a) = (0.11, 0.82)$ and $(l_b, u_b) = (0.52, 0.92)$. The hypothesis test of the null hypothesis $\rho_{by} - \rho_{ay} = 0$, without accounting for correlation between the forecasts, yields a test statistic of $T_1 = 1.23$, which has a p value of 0.11 under the standard normal distribution. That is, if the null hypothesis of zero correlation difference were true (and if the forecasts were uncorrelated), 11% of all sample values of the test statistic T_1 would be at least as large as the observed value of 1.23. For the t test of Steiger (1980), which accounts for the correlation between the forecasts, we obtain a value of the test statistic of $T_2 = 1.69$. The p value under the t distribution with $n - 3 = 14$ degrees of freedom is 0.06, that is, about 6% of all values of the test statistic T_2 would exceed the observed 1.69 if the null hypothesis of zero correlation difference were true. The confidence interval of the correlation difference, based on the method by Zou (2007) is equal to $(L, U) = (-0.05, 0.65)$.

Table 2 summarizes the correlation analysis of the four time series of Fig. 1. In all examples, the t test based on the test statistic T_2 [Eq. (7)], which accounts for correlation between forecasts, yields lower p values than the test based on T_1, which ignores correlation between forecasts. The effect of accounting for high correlation between forecasts is best illustrated in the analysis of region western Africa (WAF). The correlation difference between the forecasts is very small at 0.06. The test statistic T_1 yields a p value of 0.37, indicating that the observed value of T_1 is compatible with the null hypothesis of zero difference. But the correlation between the forecasts is very large at 0.98. For the test based on T_2, which does account for correlation between forecasts, the p value is very small, leading to rejection of H_0 at the 5% significance level. A given correlation difference is deemed to be more significant, the more strongly the forecasts are correlated with each other. Note, however, that the two forecasts for WAF are very similar to each other, so it is important to distinguish between statistical and practical significance of the results. It is further worthwhile to note that the correlation differences in regions CEU and eastern Asia (EAS) are very different, but the corresponding p values are very similar. On the other hand, the correlation differences in regions EAS and northeastern Brazil (NEB) are very similar, but the p values are very different. Last, we note that as a result of the soil moisture–temperature feedback (dry/wet conditions lead to warmer/colder temperatures), the variance of the forecast system B is slightly higher than the variance of forecast A in all four regions.

Figure 2 shows correlation coefficients r_{ay}, r_{by}, and r_{ab} on individual grid points over land. Upon visual inspection, the correlations of forecast B with the observations seem to be higher than for forecast A. Further, the plot of r_{ab} shows that the two forecasts are highly positively correlated in most regions, which shows that the underlying assumption of the test statistic T_1 is not justified most of the time. There are, however, some regions where the correlation between forecasts is actually close to zero, or even negative. Forecasts seem to be less correlated with each other in the regions where they show low correlation skill (e.g., northeast Asia). In regions where there is high correlation skill in both forecasts

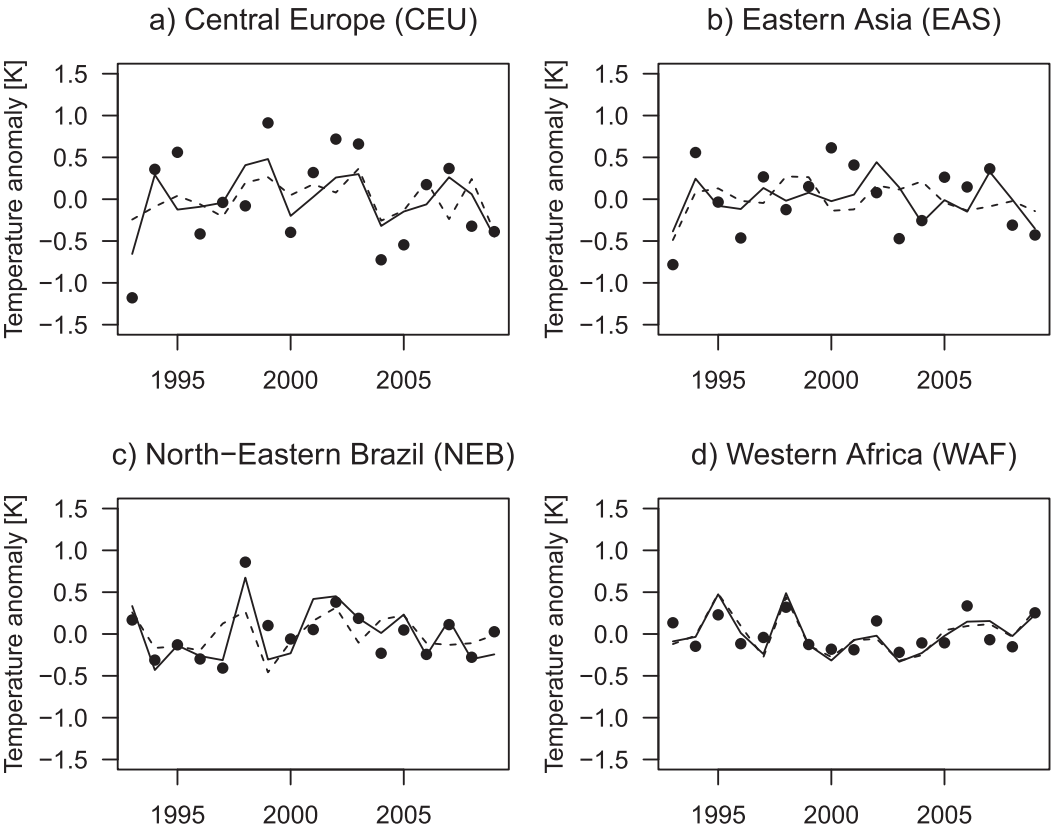

FIG. 1. Time series plots of area-averaged temperature anomalies for the four regions: observations (circles), forecast A initialized with climatological land surface conditions (dashed lines), and forecast B initialized with realistic land surface conditions (solid lines).

($r_{ay} > 0$ and $r_{by} > 0$), the correlation between the forecasts r_{ab} also tends to be high (e.g., central Africa).

The correlation differences at individual grid points are shown in the top panel of Fig. 3. Stippled points indicate grid points where the one-sided test based on the test statistic T_2 yields a p value smaller than 0.05, and, therefore, rejects the null hypothesis at the 5% significance level. As expected, these points appear mainly in regions where the correlation difference $r_{by} - r_{ay}$ is large, or where the correlation between forecasts r_{ab} is large. The bottom panels of Fig. 3 show that the same correlation difference can be deemed significant by the test based on T_2 but not significant when T_1 is used, and vice versa. In general, the test based on T_2 leads to more rejections of the null hypothesis. More than twice as many points are marked as significant in the bottom-right panel of Fig. 3 than in the bottom-left panel.

We comment on field significance, following the procedure first proposed by Livezey and Chen (1983). There is a total of $n = 6964$ land grid points in the top

TABLE 2. Table summarizing correlation coefficients, hypothesis tests, and confidence intervals for the data shown in Fig. 1.

Region		CEU	EAS	NEB	WAF
Sample correlations	r_{ay}	0.56	0.17	0.41	0.69
	r_{by}	0.80	0.58	0.83	0.75
	r_{ab}	0.62	0.41	0.72	0.98
	$r_{by} - r_{ay}$	0.24	0.42	0.41	0.06
One-sided test of H_0: $\rho_{by} - \rho_{ay} = 0$ using T_1	$\hat{T}_1$	1.23	1.30	1.99	0.33
	p value	0.109	0.097	0.023	0.371
One-sided test of H_0: $\rho_{by} - \rho_{ay} = 0$ using T_2	$\hat{T}_2$	1.69	1.72	4.07	2.12
	p value	0.057	0.053	< 0.001	0.026
95% confidence interval for $\rho_{by} - \rho_{ay}$	L	−0.05	−0.07	0.15	−0.09
	U	0.65	0.89	0.85	0.29

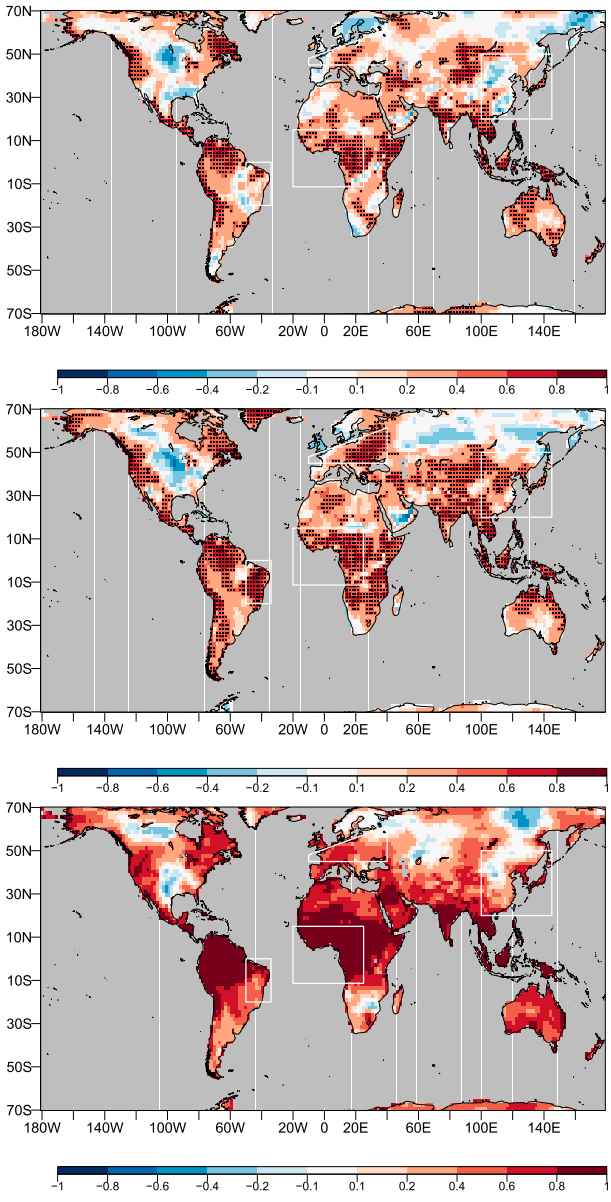

FIG. 2. Correlation maps. (top) Correlation r_{ay} between forecast A and observations. (middle) Correlation r_{by} between forecast B and observations. (bottom) Correlation r_{ab} between forecast A and forecast B. Black dots in the (top) and (middle) indicate points where the p value of a one-sided test based on the test statistic T_0 is less than 0.05. The white polygons indicate the four regions of Table 1.

panel of Fig. 3, and $k = 443$ grid points are significant (i.e., a fraction of $k/n = 0.0636$). If the null hypothesis were true on every single grid point, and if the tests on the individual grid points were independent, k would follow a binomial distribution with size n and success probability 0.05. The chance of observing a value at least as large as $k = 443$ under this binomial distribution is $\approx 2 \times 10^{-7}$ (i.e., highly unlikely). But the individual tests are not independent due to spatial correlation; the effective number of independent grid points is less than n. In order for a fraction $k/n = 0.0636$ to have a larger than 5% chance of occurring under the null hypothesis, the number of independent tests n would have to be smaller than 725. A detailed estimation of the spatial degrees of freedom is outside the scope of this study, but we can provide a rough estimate based on visual inspection. The decorrelation length of the data is about 10 grid cells, which suggests that the map consists of independent circular regions, each consisting of about 80 grid cells. Our estimate of the effective degrees of freedom is thus about $n/80 \approx 90$, which is much smaller than 725. A field significance test thus would not reject the global null hypothesis that the correlation difference is zero everywhere.

According to Table 2, the p values of the two tests based on T_1 and T_2 differ—the p value of T_2 is always smaller than that based on T_1. Furthermore, similar correlation differences in different regions do not imply similar p values. It happens that the p value is smaller than 0.05, but the 95% confidence interval overlaps zero. There are thus situations where one test deems the difference in sample correlation to be statistically significant, while another test does not. In the following section we analyze in more detail the difference between the various statistical tests using simulated data.

5. Type-I error rate and power analysis

The present section addresses two important questions concerning statistical tests of correlation forecast skill:

1) If forecasts A and B had equal skill (i.e., $\rho_{ay} = \rho_{by}$), how frequently does a given statistical test (falsely) reject the null hypothesis of zero correlation difference?
2) If forecast B were more skillful than forecast A (i.e., $\rho_{by} > \rho_{ay}$), how often does a given statistical test (correctly) reject the null hypothesis of zero correlation difference?

In the first question, a rejection of H_0 is clearly undesired, and constitutes a type-I error. If H_0 is true, the 95% confidence interval should fail to include the value of zero correlation difference on average 5% of the time. Similarly, on average 5% of all one-sided p values should be smaller than 0.05 if H_0 is true. Statistical tests based on p values and confidence intervals should, by definition, have a type-I error rate equal to the nominal significance level (e.g., 5%). But since the statistical methods of section 3 involve approximations and parametric assumptions about the data, the actual rate of

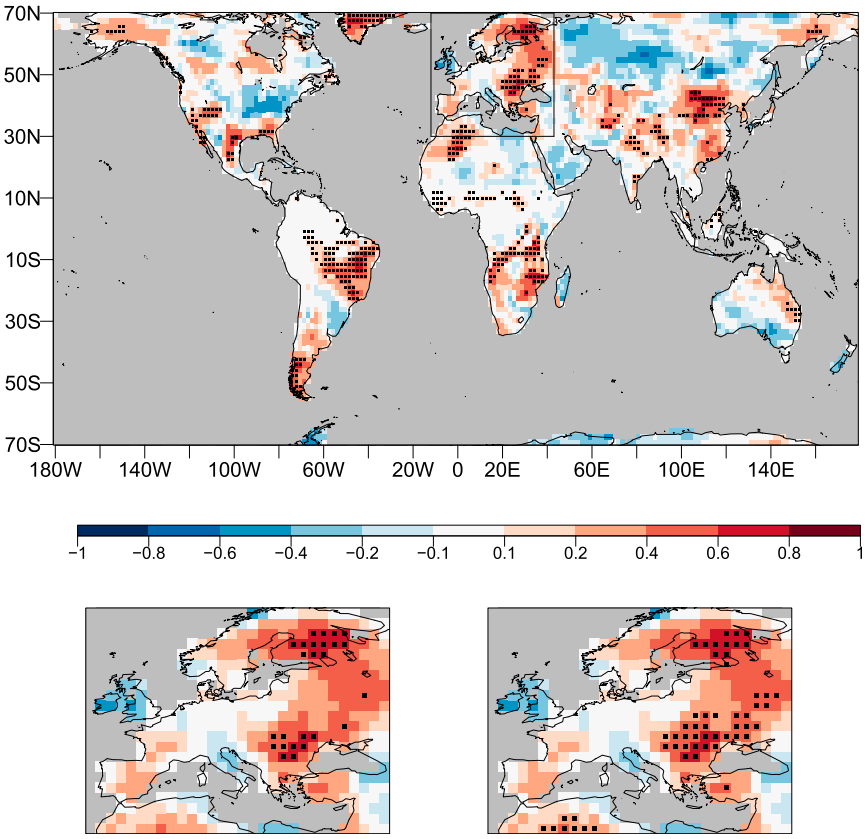

FIG. 3. (top) Map of correlation difference $r_{by} - r_{ay}$. Dots indicate differences that are significant at the 5% significance level (one-sided test based on the test statistic T_2). (bottom) Correlation differences in the boxed region, marking correlation differences that are deemed significantly larger than zero using (left) the test statistic T_1 and (right) the test statistic T_2.

rejecting H_0 might be different from the nominal value. Analyzing the type-I error rate will thus be useful to learn about the reliability of different statistical tests.

In the second question, a rejection of H_0 is clearly desired, because rejection amounts to detecting a genuine improvement in forecast quality. Statistical power (i.e., the chance of correctly rejecting a false H_0) depends on the details of the statistical test, on the actual difference between ρ_{by} and ρ_{ay} (effect size), and on the sample size n (Cohen 1992). Obviously, statistical tests with high power are desirable. An estimate of the power of the different tests will be useful because power characterizes our ability to detect improvements in forecast quality.

To analyze the power of a test, we have to know how often a test rejects the null hypothesis, given that one forecast is, in fact, more skillful than the other. For actual climate data, such as the data analyzed in section 4, one never knows exactly whether one forecast system has more skill than another. If one knew, there would be no need for statistical testing. To analyze power and type-I error rates, we thus have to use simulation studies, where we can control whether H_0 is true or false. For all analyses of the present section, we simulate forecasts of system A $\{a_1, \ldots, a_n\}$ and of system B $\{b_1, \ldots, b_n\}$, as well as their common verifying observations $\{y_1, \ldots, y_n\}$, by sampling from a trivariate normal distribution with expectation vector $\boldsymbol{\mu} = (0, 0, 0)^{\mathrm{T}}$ and covariance:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho_{ab} & \rho_{ay} \\ \rho_{ab} & 1 & \rho_{by} \\ \rho_{ay} & \rho_{by} & 1 \end{pmatrix}. \tag{10}$$

Such data can be interpreted as representing a climate index that was normalized to mean zero and unit variance. The off-diagonal elements of $\boldsymbol{\Sigma}$ indicate the correlations between forecasts and the observation. If we simulate data using a covariance matrix which has $\rho_{ay} = \rho_{by}$, both forecasts are equally skillful at predicting the observations, and the null hypothesis of zero correlation difference is, therefore, true. If we set $\rho_{ay} < \rho_{by}$ in

the covariance matrix, forecast B is more skillful at predicting the observations than forecast A, and the null hypothesis of zero correlation difference is, therefore, false. Note that ρ_{ay}, ρ_{by}, and ρ_{ab} must be chosen such that $\mathbf{\Sigma}$ is positive semidefinite, which is satisfied if all three ρs are in [–1, 1], and if $|\mathbf{\Sigma}| = 1 - \rho_{ay}^2 - \rho_{by}^2 - \rho_{ab}^2 + 2\rho_{ay}\rho_{by}\rho_{ab}$ is nonnegative.

To calculate power and type-I error rate of a given test, we use the following protocol:

1) Fix values for ρ_{ay}, ρ_{by}, and ρ_{ab}, as well as the sample size n.
2) Draw n triplets $\{a_t, b_t, y_t\}$, $t = 1, \ldots, n$, from the corresponding trivariate normal distribution and interpret these data a hindcast dataset of size n of two competing forecast systems A and B for the same observation.
3) Perform the given hypothesis test of the null hypothesis H_0: $\rho_{ay} = \rho_{by}$.
4) Record whether or not the test rejects H_0.
5) Repeat steps 2–4 a large number of times, each time with a different realization of artificial hindcast data.
6) Calculate the fraction of rejected null hypotheses.

If $\rho_{ay} = \rho_{by}$ the null hypothesis is true, and the fraction of rejected null hypotheses is, therefore, an estimate of the type-I error rate of the given test. If $\rho_{by} \neq \rho_{ay}$, the null hypothesis is false, and the fraction of rejected null hypotheses is, therefore, an estimate of the power of the given test.

We have analyzed type-I error rates of the hypothesis tests and confidence intervals presented in section 3. We simulate artificial hindcast datasets of sample size $n = 20$, similar to the data analyzed in section 4. We use $\rho_{ay} = \rho_{by} = 0.4$, which we consider reasonable values achievable by state-of-the-art climate forecast systems. Even though ρ_{ay} and ρ_{by} are equal, sample correlations r_{ay} and r_{by} calculated from a finite sample of size n are generally different from the population values, and different from each other. For a number of values of $\rho_{ab} \in [0, 0.99]$ we calculate 10^5 artificial hindcast datasets. We use hypothesis tests based on the test statistics T_1 and T_2, as well as the confidence interval calculated according to Eq. (9). Our statistical tests reject the null hypothesis if the p value of the two-sided test is smaller than 0.05, and if the central 95% confidence interval does not overlap the value zero. Since we chose $\rho_{ay} = \rho_{by}$, H_0 is true and the empirical type-I error rate should be equal to 5%.

Figure 4 shows that empirical type-I error rates are not always equal to the nominal value of 5%. Type-I error rates of the different tests are shown as a function of the between-forecast-correlation ρ_{ab} of the simulated hindcasts. If the forecasts are not or only weakly correlated ($\rho_{ab} < 0.1$) all tests behave as expected: the null

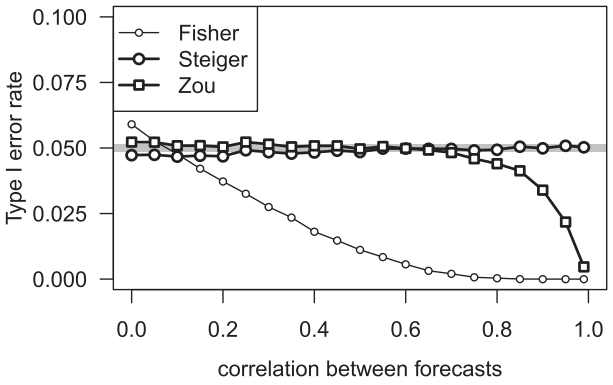

FIG. 4. Empirical type-I error rates of tests of correlation difference, based on simulated hindcast data with $\rho_{ay} = \rho_{by} = 0.4$ and $n = 20$. Hypothesis test based on the Fisher transformation (small circles; test statistic T_1), which ignores correlation between forecasts. Hypothesis test based on Steiger (1980) (big circles; test statistic T_2), which accounts for correlation between forecasts. (squares) Confidence intervals based on Zou (2007) (squares), which account for correlation between forecasts. The nominal type-I error rate used for the tests of 5% is indicated by the gray line.

hypothesis is rejected about 5% of the time, and the confidence intervals fail to cover the value of zero about 5% of the time. The tests behave differently if the forecasts are moderately or strongly correlated. The hypothesis test based on the test statistic T_1, which does not account for correlation between forecasts, has type-I error rates much smaller than 5%. The test is too conservative (i.e., it does not reject the null hypothesis often enough). DelSole and Tippett (2014) showed this analytically. By contrast, the test based on the test statistic T_2, which accounts for correlation between forecasts, rejects the null hypothesis 5% of the time, independent of the strength of the correlation between forecasts. The empirical and nominal type-I error rates agree (i.e., the test based on T_2 is reliable). The confidence intervals have correct coverage frequencies for all $\rho_{ab} < 0.8$. For strongly correlated forecasts, however, the confidence intervals become overdispersed; they cover the true value of zero correlation difference more often than indicated by their confidence coefficient of 95%. As a result, the empirical type-I error rate is smaller than the nominal 5%. However, the coverage frequencies of the confidence intervals at high values of ρ_{ab} improve for larger sample sizes n (not shown).

We also compare statistical power of one-sided tests for improvement based on the test statistics T_1 and T_2. We simulate hindcast datasets of size $n = 17$ under the assumption that the correlations ρ_{ay}, ρ_{by}, and ρ_{ab} are equal to the sample correlations r_{ay}, r_{by}, and r_{ab} in the four regions, as shown in Table 2. Using the correlation structure of each region, we perform 10^5 one-sided tests

TABLE 3. Power of statistical tests based on test statistic T_1 and T_2, assuming population correlations that are equal to the sample correlations of the four regions shown in Table 2.

Region	Power (T_1)	Power (T_2)
CEU	0.30	0.50
EAS	0.34	0.51
NEB	0.74	0.98
WAF	0.00	0.54

that reject H_0 if the p value is smaller than 0.05. Since $\rho_{by} > \rho_{ay}$ in each setting, statistical tests should reject H_0 as often as possible; each nonrejection constitutes a type-II error. Table 3 summarizes empirical rejection rates of tests based on the test statistics T_1 and T_2. In each setting the test based on T_2 achieves higher power than the often-used test that does not account for correlation between forecasts. The increase in power is substantial (between 16% and 54%). For the setting of region WAF, where the correlation difference is very small, but the correlation between forecasts is very high, we did not get a single rejection of H_0 when using the test based on T_1, compared to 54% correct rejections if we use T_2. Use of an appropriate statistical test has considerably improved our ability of detecting the (small) improvement in forecast quality.

In medical research, for example, it is common practice to demand that statistical tests at a significance level of 5% should achieve power of at least 80% (Cohen 1992). In three of the settings of Table 3, the power of the test based on T_2 is less than 80%. One way to increase power is to increase the sample size, because larger samples allow for more robust estimation of the correlation difference, which increases the chance of correctly detecting a genuine difference of the population correlations. In Fig. 5 the power in the four correlation settings is shown as a function of the sample size n. For the correlation structure of region NEB, power greater than 80% is already achieved at sample sizes of $n = 10$. The correlation structure of the other three regions requires sample sizes greater than $n = 40$ in order to detect the improvement with sufficient power.

The dependency between correlation structure and power is not straightforward, and it is worth analyzing this dependency further. Figure 6 shows how power of the test based on T_2 depends on ρ_{ay}, ρ_{by}, and ρ_{ab}. Three values of ρ_{ay} were considered (0.0, 0.3, and 0.6), and ρ_{by} was chosen greater or equal to ρ_{ay}. If $\rho_{ay} = \rho_{by}$, there is no improvement of forecast B over forecast A. The null hypothesis is therefore true. A test at significance level 5% should therefore reject the null hypothesis on average 5% of the time. This is confirmed by the plots in Fig. 6: the power curves meet at values of 0.05 at their

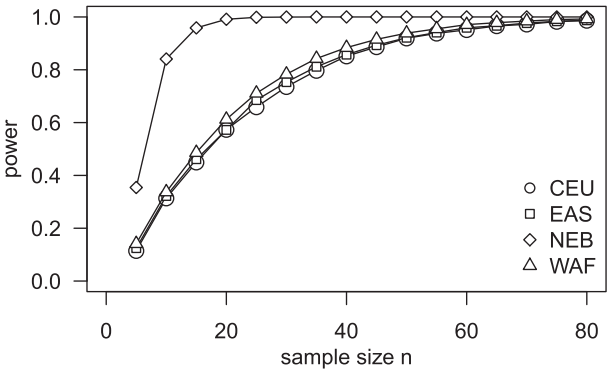

FIG. 5. Power as function of the hindcast sample size n, assuming population correlations that are equal to the sample correlations calculated for the four regions in Table 2.

leftmost points, where $\rho_{ay} = \rho_{by}$. If ρ_{by} is increased, the null hypothesis is false. The test rejects more often the bigger the difference is between ρ_{by} and ρ_{ay} (i.e., the bigger the improvement in correlation skill of forecast B compared to forecast A). If ρ_{by} approaches 1, the power converges to 1, independent of ρ_{ab} and ρ_{ay}. That is, a perfect forecast with correlation close to 1 can always be perfectly distinguished from an imperfect forecast. Furthermore, the more correlated the two forecasts are, the higher the power of detecting an improvement using the statistical test based on the test statistic T_2. Figure 6 shows that for each setting, small improvements in correlation of less than 0.2 cannot be detected with sufficient power, based on sample size of $n = 17$ and a 5% significance level. When ρ_{ay} is small, even an increase of correlation of 0.4 cannot be detected with sufficient power.

6. Improved predictions of climate indices by increasing model resolution

In this section we present an additional application of the statistical methodology of this paper. A standard approach to evaluate the ability of forecast systems at predicting regional climate variability is to check their skill to forecast the main modes of climate variability, such as El Niño–Southern Oscillation (ENSO; Trenberth 1997) or the North Atlantic Oscillation (NAO; Hurrell 1995). In this context, a rigorous application of statistical tests is essential to compare different forecast systems. The size of the sample of predictions that is used to compare the skill of different forecast systems should be chosen depending on the initial skill of the forecast system. Additionally, the high similarities between two versions of the same forecast system have to be taken into account when evaluating skill improvement. As an illustration, the EC-Earth model

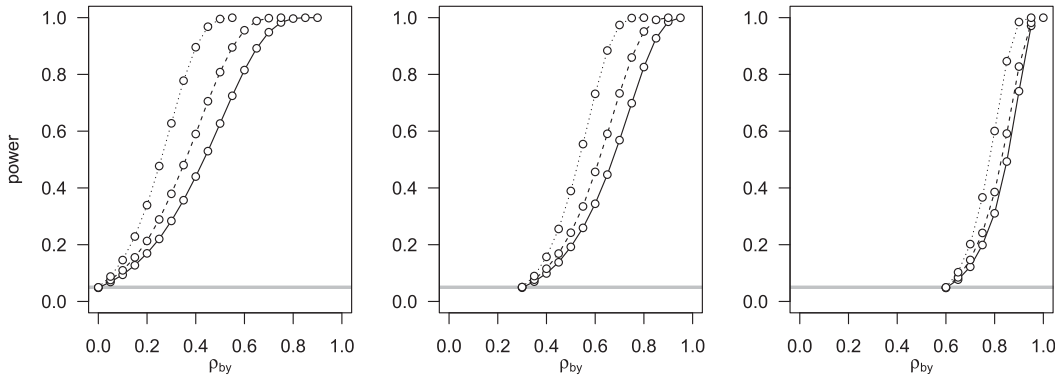

FIG. 6. Power as a function of the correlation skill improvement: (left) $\rho_{ay} = 0$, (middle) $\rho_{ay} = 0.3$, and (right) $\rho_{ay} = 0.6$. The line styles indicate the correlation between the forecasts: $\rho_{ab} = 0.4$ (solid), $\rho_{ab} = 0.6$ (dashed), and $\rho_{ab} = 0.8$ (dotted). The sample size is $n = 17$. The gray lines indicate the nominal type-I error rate of 0.05 used for all tests.

initialized in May with a resolution of ≈80 km in the atmosphere (T255) and ≈100 km in the ocean (1°) shows a relatively good skill to forecast ENSO for the months June–July–August (JJA): the correlation with the sea surface temperature observational dataset from Merchant et al. (2014) is $r_{ay} = 0.78$ over the period 1993–2009 ($n = 17$). Running the same forecast with a higher resolution, reaching ≈40 km in the atmosphere (T511) and ≈25 km in the ocean (0.25°) leads to a correlation of $r_{by} = 0.85$. The increase in correlation due to the higher resolution is relatively small: $r_{by} - r_{ay} = 0.07$. The increase is statistically significant at the 5% level when using the statistical test for differences between overlapping correlations; the p value based on the test statistics T_2 is 0.019. The same correlation difference of 0.07 is not statistically significant at the 5% level when the (high) correlation between forecasts ($r_{ab} = 0.971$) is neglected; the p value based on T_1 is 0.287.

Seasonal forecast of NAO is more challenging than for ENSO. Seasonal forecast systems typically obtain correlation skill between 0 and 0.3 at predicting the winter NAO on seasonal time scales (Shi et al. 2015). At commonly used sample sizes of around 20, these values are not statistically significant at the 5% significance level. Considering the low skill and the additional fact that two forecast systems with low skill are typically not highly correlated with each other, a large sample of predictions has to be used to detect any increase of the correlation from one model version to the next. We have estimated that samples with a size of $n = 120$ should be used to detect a correlation skill increase from 0.1 to 0.4 for winter NAO predictions with a power of 0.8. In addition, if we consider a sample of 17 winter NAO predictions with a correlation skill close to 0, a second set of predictions with the same size could be differentiated from the first one with power of 0.8 only if it had a correlation skill of about 0.7, which would be an exceptional increase.

7. Conclusions

A commonly used statistical test for detecting improvement in correlation skill was shown to be too conservative and underpowered, because it assumes that the two competing forecasts are uncorrelated with one another. Using an appropriate test that correctly accounts for the (high) correlation between forecasts improves the power of detecting genuine increases in forecast skill. We therefore strongly recommend using the test by Steiger (1980) based on the test statistics T_2 for comparative studies of correlation skill. The method by Zou (2007) for construction of confidence intervals for correlation differences is generally reliable, but strongly correlated forecasts and small sample sizes can lead to overdispersed confidence intervals.

The importance of power analysis has been pointed out in the climate literature by Jolliffe (2007) and Wilks (2010). Power analysis is common practice in designing medical studies, in order to determine the necessary sample size to detect a hypothesized effect of a given treatment. To our knowledge, power is not currently considered when designing hindcast experiments for comparing climate forecast systems. But with insufficient sample sizes, it is unlikely to detect significant differences in forecast skill, which limits the usefulness of the computationally expensive hindcast simulation. Clearly, analyzing differences in forecast skill is not the only purpose why hindcast datasets are simulated; different applications include diagnosing model errors and calculating bias corrections. But these applications are subject to

statistical uncertainties as well. There is always a chance of falsely diagnosing a model error or failing to diagnose an existing forecast bias due to insufficient sample size. Given the computational resources required to run hindcast experiments with state-of-the-art global climate forecast systems, statistical power should be taken more seriously if significance testing and confidence intervals are used to diagnose improvements. The present study demonstrates a simple simulation-based framework for investigating statistical power, and could be exploited for better design of hindcast experiments. Using the framework for power analysis presented here, more general settings can be analyzed. The present study only focused on differences in correlation skill of univariate data. In actual hindcast datasets, data are high dimensional, spatially and temporally correlated, and possibly nonnormal. These settings should be considered in future studies.

Comparative verification studies are often performed between forecast systems that are not very different from each other. It can be hypothesized that any improvement of one forecast over another is necessarily small (i.e., ρ_{by} is generally close to ρ_{ay}). We showed that the power of detecting small improvements in correlation skill tends to be low. Statistical tests rarely reject the null hypothesis of zero skill difference, even though there might be a difference. Therefore, uncertainty always remains about which forecast is the ''better'' one to be used for operational forecasting. The lack of power at picking the ''best'' forecast motivates a multimodel approach, where a multitude of available forecast systems are run in parallel, and a consensus forecast is calculated from all candidate forecasts.

As in the present study, forecasts are often calculated by averaging over a finite number of ensemble forecasts to average out internal model variability, and thus obtain a better estimate of the predictable signal of the model. Depending on the ensemble size, and the signal-to-noise ratio of the ensemble forecasts, there might be an inherent upper bound on the achievable correlation skill. Such an upper bound limits the possible magnitudes of improvement that, in turn, limits the power of detecting any improvements of ensemble forecasts.

Furthermore, power might be different for different evaluation criteria than correlation skill. But for different evaluation criteria, the notion of which forecast is better changes—we might find that forecast B has higher correlation than forecast A, but a worse ROC statistic or Brier score [for definitions, see Jolliffe and Stephenson (2012)] than forecast A. Given that different criteria yield different definitions of ''improvement,'' we do not generally recommend a comparison of statistical power between different evaluation criteria.

We have shown in the appendix that correlation is closely related to the mean squared error (MSE), so instead of analyzing differences in correlation one might analyze difference in the MSE of the recalibrated forecasts. The MSE has the benefit that it is a scoring rule; that is, it assigns an individual value to each pair (a_t, y_t) of forecast and observation, which is not the case for the correlation coefficient. If scoring rules are used for forecast evaluation, the statistical test of Diebold and Mariano (1995) can be used. This test is based on loss differentials and therefore takes into account correlation between forecasts. The test also includes a correction for serially correlated data.

This paper presented appropriate statistical tests for analyzing skill improvements, and power analysis as a method to evaluate such tests. The proposed tests were used to analyze seasonal hindcast datasets as practical examples, but can be applied to short-term weather forecasting and climate projections as well. It was shown that realistic land surface representation leads to significantly higher correlation skill in temperature forecasts. It was further shown that increased atmosphere and ocean resolution leads to significantly improved correlation of ENSO forecasts. For NAO predictions, for which most current systems have low skill, it was shown that very large hindcast datasets would be required to detect small increases in skill with sufficiently high power.

Acknowledgments. The authors acknowledge support by the European Union Program FP7/2007-13 under Grant Agreement 3038378 (SPECS). The work of O. Bellprat was funded by ESA under the Climate Change Initiative (CCI) Living Planet Fellowship VERITAS-CCI. Acknowledgment is made for the use of ECMWF's computing and archive facilities in this research, and the computer resources, technical expertise, and assistance provided by the Red Española de Supercomputación. The views expressed herein are those of the authors and do not necessarily reflect the views of their funding bodies or any of their subagencies. We wish to thank Timothy DelSole, two anonymous reviewers, and the editor for their comments that helped to improve the quality of the paper.

APPENDIX

Derivation: Correlation Squared is a Skill Score

Suppose a series of forecasts x_t for observations y_t $(t = 1, \ldots, n)$. The forecast x_t is recalibrated by linear regression on the observation to remove systematic

biases and scaling errors. The linearly recalibrated forecast $\hat{x}_t$ is given by

$$\hat{x}_t = \overline{y} + \frac{\mathrm{cov}(xy)}{\mathrm{var}(x)}(x_t - \overline{x}). \tag{A1}$$

The mean squared skill score (MSSS) of the recalibrated forecast with respect to the climatological forecast $\overline{y}$ is given by

$$\mathrm{MSSS} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{x}_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}, \tag{A2}$$

$$= 1 - \frac{\sum_{t=1}^{n}\left[y_t - \overline{y} - \frac{\mathrm{cov}(xy)}{\mathrm{var}(x)}(x_t - \overline{x})\right]^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}, \tag{A3}$$

$$= 1 - \frac{\mathrm{var}(y) + \frac{\mathrm{cov}(xy)^2}{\mathrm{var}(x)^2}\mathrm{var}(x) - 2\frac{\mathrm{cov}(xy)}{\mathrm{var}(x)}\mathrm{cov}(xy)}{\mathrm{var}(y)}, \tag{A4}$$

$$= \frac{\mathrm{cov}(xy)^2}{\mathrm{var}(x)\mathrm{var}(y)} = \mathrm{cor}(xy)^2. \tag{A5}$$

The MSSS of the linearly recalibrated forecast $\hat{x}_t$ is thus equal to the squared correlation between the forecasts x_t and observations y_t.

REFERENCES

Balsamo, G., and Coauthors, 2015: ERA-Interim/Land: A global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.*, **19**, 389–407, doi:10.5194/hess-19-389-2015.

Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär, 2013: Physical constraints for temperature biases in climate models. *Geophys. Res. Lett.*, **40**, 4042–4047, doi:10.1002/grl.50737.

Cohen, J., 1992: Statistical power analysis. *Curr. Dir. Psychol. Sci.*, **1**, 98–101, doi:10.1111/1467-8721.ep10768783.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:10.1002/qj.828.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, doi:10.1175/MWR-D-14-00045.1.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13** (3), 134–144, doi:10.1198/073500102753410444.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013a: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, doi:10.1002/wcc.217.

——, and Coauthors, 2013b: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, doi:10.1038/ncomms2704.

Du, H., F. Doblas-Reyes, J. Garca-Serrano, V. Guemas, Y. Soufflet, and B. Wouters, 2012: Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. *Climate Dyn.*, **39**, 2013–2023, doi:10.1007/s00382-011-1285-9.

Ferry, N., and Coauthors, 2012: GLORYS2V1 global ocean reanalysis of the altimetric era (1993–2009) at meso scale. *Mercator Ocean Quart. Newsl.*, **44**, 28–39. [Available online at http://www.mercator-ocean.fr/wp-content/uploads/2015/05/Mercator-Ocean-newsletter-2012_44.pdf.]

Guemas, V., F. J. Doblas-Reyes, K. Mogensen, S. Keeley, and Y. Tang, 2014: Ensemble of sea ice initial conditions for interannual climate predictions. *Climate Dyn.*, **43**, 2813–2829, doi:10.1007/s00382-014-2095-7.

Hazeleger, W., and Coauthors, 2012: EC-Earth V2.2: Description and validation of a new seamless earth system prediction model. *Climate Dyn.*, **39**, 2611–2629, doi:10.1007/s00382-011-1228-5.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, **269**, 676–679, doi:10.1126/science.269.5224.676.

IPCC, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, 582 pp. [Available online at http://ipcc-wg2.gov/SREX/report/full-report/.]

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637–650, doi:10.1175/WAF989.1.

——, and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp.

Keenlyside, N. S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84–88, doi:10.1038/nature06921.

Koster, R. D., and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, doi:10.1126/science.1100217.

Livezey, R. E., and W. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111**, 46–59, doi:10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2.

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). *Geosci. Data J.*, **1**, 179–191, doi:10.1002/gdj3.20.

Murphy, A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–582, doi:10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Pepler, A. S., L. B. Díaz, C. Prodhomme, F. J. Doblas-Reyes, and A. Kumar, 2015: The ability of a multi-model seasonal forecasting ensemble to forecast the frequency of warm, cold and wet extremes. *Wea. Climate Extremes*, **9**, 68–77, doi:10.1016/j.wace.2015.06.005.

Rodgers, J. L., and W. A. Nicewander, 1988: Thirteen ways to look at the correlation coefficient. *Amer. Stat.*, **42**, 59–66, doi:10.2307/2685263.

Shi, W., N. Schaller, D. MacLeod, T. N. Palmer, and A. Weisheimer, 2015: Impact of hindcast length on estimates of seasonal climate predictability. *Geophys. Res. Lett.*, **42**, 1554–1559, doi:10.1002/2014GL062829.

Siegert, S., D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas, 2016: A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *J. Climate*, **29**, 995–1012, doi:10.1175/JCLI-D-15-0196.1.

Steiger, J. H., 1980: Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, **87**, 245–251, doi:10.1037/0033-2909.87.2.245.

Trenberth, K. E., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771–2777, doi:10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2.

Von Storch, H., and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press, 496 pp.

Wilks, D., 2010: Sampling distributions of the Brier score and Brier skill score under serial dependence. *Quart. J. Roy. Meteor. Soc.*, **136**, 2109–2118, doi:10.1002/qj.709.

——, 2011: *Statistical Methods in the Atmospheric Sciences*. Vol. 100, Academic Press, 704 pp.

Williams, E. J., 1959: The comparison of regression variables. *J. Roy. Stat. Soc. B*, **21** (2), 396–399. [Available online at http://www.jstor.org/stable/2983809.]

Zhang, J., L. Wu, and W. Dong, 2011: Land-atmosphere coupling and summer climate variability over East Asia. *J. Geophys. Res.*, **116**, D05117, doi:10.1029/2010JD014714.

Zou, G. Y., 2007: Toward using confidence intervals to compare correlations. *Psychol. Methods*, **12**, 399–413, doi:10.1037/1082-989X.12.4.399.