Predictive skill of climate indices compared to mean quantities in seasonal forecasts

Jonas Bhend,* Irina Mahlstein and Mark A. Liniger
Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland

*Correspondence to: J. Bhend, Federal Office of Meteorology and Climatology MeteoSwiss, Operation Center 1, PO Box 257, 8058 Zurich-Airport, Switzerland. E-mail: jonas.bhend@meteoswiss.ch

Seasonal forecasting models are increasingly being used to forecast application-relevant aspects of upcoming climatic conditions, often summarised by climate indices. Little is known, however, on how the predictive skill of such forecasts of climate indices relates to the predictive skill in forecasting seasonal mean conditions. Here we analyse forecasts of two generalised indices derived from daily minimum and maximum temperature: counts of events such as the number of frost days and accumulated threshold exceedances such as degree days. We find that the predictive skill of forecasts of these two types of indices is generally lower than the skill of seasonal mean daily minimum and maximum temperature. By use of a toy model we demonstrate that this reduction in skill is more pronounced for skillful forecasts and climate indices defined relative to a threshold at the tail of the distribution. Based on the toy model results we conclude that there is no indication of additional predictability in forecasts of these indices in excess of what is expected due to the predictability of the seasonal mean. To further support this hypothesis, we show that the skill in predicting climate indices can be statistically modelled successfully using the toy model and the skill in the seasonal mean.

Key Words: seasonal forecasting; climate indices; predictive skill; degree days; event frequency

Received 5 January 2016; Revised 2 August 2016; Accepted 22 August 2016; Published online in Wiley Online Library 20 October 2016

1. Introduction

Climate indices have become widely used in climate research studies (e.g. Stöller et al., 2013a, 2013b). Such indices can help to identify and quantify variability and changes in particular aspects of the climate system (Klein Tank et al., 2002). Additionally, climate indices can offer the possibility to describe the climate in a more approachable way than using the moments of the distribution such as the mean and the standard deviation. For example, the number of dry days per time period is more informative to most people than the summer mean precipitation. Hence, indices describe the climate in a way that users can more readily relate to. Some indices can directly be linked to user-related quantities, such as heating degree days to heating energy demand. These characteristics make the use of indices ideal for seasonal forecasting. If the user interest can be characterised by a climate index and the index can be skillfully forecasted, such a forecast conveys more relevant information to the user than a forecast of the underlying variable. Plus, there is no need to use a complex impact model. Processing daily data of seasonal forecasts on the other hand is computationally more expensive and more delicate bias removal techniques are needed (e.g. Mahlstein et al., 2015). These factors result in higher computational needs and a later delivery of operational forecasts. It remains unclear, to what extent the potential added value is worth this additional effort.

Hamilton et al. (2012) and Pepler et al. (2015) have recently shown that predictability of the frequency of very warm and cold days mainly stems from predictability in seasonal mean temperatures. In general, it remains to be determined how the skill of the seasonal forecast of a climate index compares to the skill of the underlying variable. Here, we propose a general framework to address this question using a toy model that allows us to perform controlled experiments. We distinguish between two different, archetypal types of indices: aggregation of threshold departures, such as degree days, and counts of threshold exceedances, such as the number of frost days. We compare the skill of these two types of indices to the skill of the underlying variable using two different verification metrics which address different aspects of forecast quality (section 3).

Toy models have been successfully used to study seasonal forecasts. Kharin and Zwiers (2003), for example, compare alternative ways to convert ensemble information to forecast probabilities and statistical methods for forecast improvement with a toy model. Weigel et al. (2008, 2009), on the other hand, study the effect of multi-model combination and recalibration on seasonal forecast skill. The above-mentioned studies use the toy model to study monthly or seasonal mean values, in particular seasonal forecast skill. The above-mentioned studies use the toy model to study monthly or seasonal mean values, in particular seasonal forecast skill. Here we expand on the existing literature by introducing a toy model for daily time series mimicking seasonal forecasts issued as daily values. The daily time series of pseudo seasonal forecasts with prescribed predictability can...
then be used to derive indices and compare forecasts of climate indices with the forecasts of the underlying variable in an idealized setting.

In the second part of this article (section 4), we test the conclusions drawn from the toy model by analysing real seasonal forecasts of climate indices using retrospective forecasts (referred to as hindcasts from now on) from an operational seasonal forecasting system. In general, little information is available on applications of climate indices in seasonal forecasting. Brands (2013) illustrates predictive skill in seasonal forecasts of boreal winter heating degree days. Hamilton et al. (2012) and Pepler et al. (2015) study seasonal forecasts of extremes and relate the skill in extremes to the skill in seasonal mean values. We add to the above by also analysing accumulated threshold exceedences and reflecting the results with the presented toy model.

We propose a general framework on how the toy model can be used to relate the skill of index forecasts to the skill of forecasts of seasonal averages of the underlying variable. In particular, we investigate if we can predict the skill of climate indices forecasts through the skill of seasonal means and the local day-to-day variability. Such a relationship would allow forecast providers and users to estimate the expected skill in forecasting a particular climate index without having to process daily forecast time series. Alternatively, we can use the relationship from the toy model to study if there is indication of enhanced or reduced predictability in indices compared to what is expected due to the predictability of the seasonal mean quantity.

2. Data and methodology

2.1. Forecasts and verifying observations

The seasonal hindcast data stem from the European Centre for Medium-range Weather Forecast’s (ECMWF) System 4 model. This is a fully coupled atmosphere–ocean forecasting system that provides operational seasonal predictions as well as hindcasts (Molteni et al., 2011). We use forecasts of daily minimum temperature for the boreal winter season (DJF) initialised on 1 November, and forecasts of daily maximum temperature for the summer (JJA) initialised on 1 May. For both initializations, simulations with 51 ensemble members for the following 7 months are available for all hindcasts. We analyse forecasts for the summers from 1981 to 2014 and the winters from 1982 to 2015 (i.e. from forecasts initialised in November 1981-2014). The forecast data have been regridded to a common 2 × 2 degree grid using bilinear interpolation previous to further analysis. For the skill analysis, the forecasts are compared to the corresponding daily time series from the ERA-Interim reanalysis (Dee et al., 2011) interpolated to the same grid.

2.2. Toy model

To analyse general characteristics of skill in forecasts of climate indices, we make use of the toy model TM2, which simulates the behaviour of single model ensemble seasonal forecasts as introduced by Weigel and Bowler (2009). In this toy model, the forecast skill (correlation) of the seasonal forecast can be specified using the parameter \( \alpha \). We expand the toy model to include daily variability based on the assumption that predictive skill is limited to seasonal mean quantities, and higher-order moments of the distribution of daily values cannot be predicted. Being able to control the forecast skill of the seasonal mean forecast (using \( \alpha \)) allows us to study the relationship between skill in the seasonal mean of the underlying variable and skill in the index. The results from the toy model are then compared to the predictive skill of forecasts from an operational forecasting system (ECMWF System 4) to determine whether the above assumption is valid.

Extending the notation of Weigel and Bowler (2009), we denote the forecast ensemble member \( i \) for lead time (in days) \( j \) as \( f_{ij} \) and the corresponding daily observations as \( x_i \). The season length is \( n \) days, i.e. \( 1 \leq j \leq n \). Daily observations \( x_i \) are represented as a combination of the seasonal mean predictable signal \( \mu_x \), a seasonal mean chaotic, non-predictable component \( \varepsilon \), and independent and identically distributed (iid) daily variability \( \xi_j \sim N(0, \gamma^2) \), where \( N(\mu, \sigma^2) \) denotes a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). The parameter \( \gamma \) denotes the standard deviation of the daily variability in the observations. This can be used to set the fraction of daily to seasonal variability.

Similarly, each daily forecast ensemble member \( f_{ij} \) is a combination of the seasonal forecast \( f_i \) and its daily variability that may differ in variance from the observed daily variability by the factor \( \delta \). The seasonal forecast \( f_i \) is a combination of the seasonal mean predictable signal \( \mu_x \) (where the term \( \mu_x \) allows us to model forecasts with negative correlation), a noise term \( \varepsilon_\beta \), that is identical for all ensemble members and that characterizes the overconfidence of the model, and the ensemble spread \( \varepsilon \). Parameter \( \beta \) is used to specify the reliability of the seasonal forecasts. Neglecting the effect of daily variability, the seasonal mean forecasts are reliable when \( \beta = 0 \) and the forecasts are overconfident when \( \beta^2 > 0 \).

The daily forecasts and observation time series are defined as shown in Eqs (1) and (2) respectively.

\[
\begin{align*}
  f_{ij} &= f_i + \varepsilon_i = \frac{\alpha}{|\alpha|} \mu_x + \varepsilon_\beta + \varepsilon_i + \varepsilon_j, \\
  x_i &= \mu_x + \varepsilon_x + \xi_j,
\end{align*}
\]

In this toy model, there is no predictability at the daily scale as daily variability between simulations and observations is uncorrelated (\( \varepsilon_i \) and \( \xi_j \) in Eqs (1) and (2) respectively). This reflects our understanding that there is only very little to no predictability in daily variability beyond a few days in the real world (Lorenz, 1963; Kalnay et al., 1998).

This toy model is designed such that the seasonally averaged values of forecasts and observations have the same parent distribution \( N(0,1) \) irrespective of the \( \alpha \) and \( \beta \) parameters (given the daily variability of forecasts \( \varepsilon_i \) and observations \( \varepsilon_j \) is of the same magnitude). Thus, the forecasts do not have a bias in terms of seasonal mean and climatological variability.

In the current study, we are interested in the behaviour of various skill metrics of the derived climate indices compared to the skill in seasonal mean quantities (e.g. \( f \) and \( x \), respectively). We study well calibrated or reliable forecast ensembles. In the presented toy model, reliable forecasts can be produced by setting \( \beta = 0 \) and \( \delta = 1 \). These well-calibrated forecasts then have seasonal and daily variability of the same magnitude as the observations, exhibit seasonal mean forecast skill as defined by \( \alpha \) (in the limit of infinite ensembles), and are reliable in the sense that the seasonal and daily mean squared error equals the intra-ensemble variance.

2.3. Definition of climate indices

To study the behaviour of climate indices in the idealised case of the toy model and in the case of operational seasonal forecasts, we define the categories of archetypal indices: counts and accumulated threshold exceedences relative to percentiles of the distribution of daily values. The former corresponds to indices such as the number of frost days; the latter corresponds to indices such as degree days.
Correspondingly, counts of threshold exceedances where \( F_{\delta} \) reanalysis daily percentiles are further smoothed using local centres about the day of interest. The resulting forecast and reanalysis, we estimate the daily percentiles from the 11 days from the daily values of all 51 ensemble members. For the percentile of daily maximum temperature in summer and counts clustering at the centre of the corresponding season due also avoids lead time dependencies and sampling issues through percentiles of forecasts and reanalysis respectively. This approach proposed in Zhang (1981–2010 instead of 1961–1990) and an extended moving below the 10th percentile), we use a different reference period TN10p (percentage of days with daily minimum temperature 1999). In contrast to ETCCDI’s TX90p (percentage of days with on Climate Change Detection and Indices (ETCCDI: Karl et al. 2000), we compute exceedances of the 75th, 90th, 95th and 98th for the summers (JJA) from 1981 to 2010 and daily minimum are estimated based on daily maximum temperature time series to the seasonal cycle. The forecast and observed daily percentiles counts clustering at the centre of the corresponding season due to the seasonal cycle. The forecast and observed daily percentiles are estimated based on daily maximum temperature time series for the summers (JJJA) from 1981 to 2010 and daily minimum temperature time series for the winters (DJF) from 1982 to 2011. We compute exceedances of the 75th, 90th, 95th and 98th percentile of daily maximum temperature in summer and counts of days below and accumulated departures from the 25th, 10th, 5th and 2nd percentile of daily minimum temperature in winter.

For the forecasts, we estimate the daily percentiles directly from the daily values of all 51 ensemble members. For the reanalysis, we estimate the daily percentiles from the 11 days centred about the day of interest. The resulting forecast and reanalysis daily percentiles are further smoothed using local linear regression (LOESS: Cleveland and Devlin, 1988) to reduce sampling variability as proposed in Mahlstein et al. (2015). Based on the daily percentiles, counts of threshold exceedances and accumulated threshold departures are computed analogously to Eqs (3) and (4).

These two types cover a wide range of user-oriented indices with applications in sectors like agriculture, glaciology, infrastructure or energy. The count indices used here are similar in spirit to some of the widely used indices proposed by the Expert Team on Climate Change Detection and Indices (ETCCDI: Karl et al., 1999). In contrast to ETCCDI’s TX90p (percentage of days with daily maximum temperature exceeding the 90th percentile) and TN10p (percentage of days with daily minimum temperature below the 10th percentile), we use a different reference period (1981–2010 instead of 1961–1990) and an extended moving window (11 instead of 5 days) to robustly estimate the more extreme percentiles. We do not include a smoothing procedure in Zhang et al. (2005), but we further smooth the daily percentiles in addition to using a moving window to estimate the percentiles as described above.

The accumulated threshold departures (DD) are closely related to degree days (e.g. growing degree days). In contrast to degree days defined relative to a fixed threshold, the definition of accumulated threshold departures relative to daily percentiles of the climatological distribution renders these indices insensitive to biases in the forecast climatology and errors in the representation of the seasonal cycle. Additionally, it allows for more general, more comparable conclusions, as the threshold is at the same location of the probability density function at all grid points in a particular analysis.

2.4. Verification metrics

We assess the association of the forecasts using the correlation of the verifying observations with the ensemble mean forecast. In contrast to previous studies (Eade et al., 2012; Hamilton et al., 2012; Pepler et al., 2015), we use the Pearson correlation coefficient despite the fact that for count indices, the Spearman rank correlation coefficient is further appropriate and for the correlation coefficient is chosen for comparison with the correlation of the seasonal mean quantity. The systematic effects on correlation when computed on bounded indices illustrate that our analysis is suited even for probabilistic index forecasts of rare events, for which appropriate verification metrics are often not readily available.

The accuracy of the forecasts is determined using the ranked probability score (RPS: Epstein, 1969; Murphy, 1969) for tercile category forecasts. The RPS is defined as

\[
RPS = \frac{1}{k} \sum_{i=1}^{k} \left( \sum_{j=1}^{i} p_j - \sum_{j=1}^{i} o_j \right)^2,
\]

with \( k \) the number of categories and \( p_j \) and \( o_j \) the forecast and observed probabilities for category \( j \) respectively. The RPS is a negatively oriented score ranging from 0 to 1 with \( RPS = 0 \) denoting a perfect forecast. Instead of analysing the RPS directly, we analyse the ranked probability skill score (RPS = 1 − (RPS\text{cal}/RPS\text{ref})), that is the relative accuracy of the forecast (measured by RPS\text{cal}) compared to a reference forecast (RPS\text{ref}). The RPS ranges from minus infinity to one. Positive RPS indicate that the forecast is more accurate than the reference forecast. Our reference forecast is a constant forecast with probability of one-third for each tercile category, the so-called climatological forecast.

The RPS is computed for tercile category forecasts with forecast and observed terciles estimated from the distribution of forecast and observed values respectively. Hence the RPS is largely independent of systematic errors of the forecasting system. Therefore, no bias correction of the forecasts is necessary before assessing predictive skill.

2.5. Prediction and significance assessment based on the toy model

We use the toy model results to statistically model the relationship between verification metrics for seasonal means and the corresponding indices. A brute force approach that involves producing a sufficiently large ensemble of toy model results to be sure to repeatedly capture all configurations of the operational forecast is deemed to be impractical. Instead we resort to setting up a local linear regression model (Cleveland and Devlin, 1988) to predict the estimated RPS or correlation for each index given the corresponding metric of the seasonal mean and the variance ratio of daily to seasonal variability (represented by parameter \( \gamma^2 \) in the toy model).

We set up local linear regression models for each type of index (CC and DD) and for each percentile threshold separately. The local linear models are fitted to Fisher-transformed correlation and to Fisher-transformed RPS with the transformation applied to positive RPS only. The local linear model takes into account the linear dependence on the transformed metric of the mean, the logarithmically transformed daily to seasonal variance ratio, and the interaction between the two terms such that for the correlation of the CC index for a given percentile, the local regression equation is the following:

\[
\text{atanh} (\text{corr}_{\text{CC}}) = \alpha + \beta \text{atanh} (\text{corr}_{\text{mean}}) + \gamma \log \frac{\sigma^2_{\text{daily}}}{\sigma^2_{\text{ seasonal}}} + \delta \text{atanh} (\text{corr}_{\text{mean}}) \log \frac{\sigma^2_{\text{daily}}}{\sigma^2_{\text{ seasonal}}} + \epsilon.
\]
In Eq. (6), \( \text{corr}_{CC} \) and \( \text{corr}_{mean} \) denote the correlation of the count index at a given threshold and the correlation of the seasonal mean respectively. \( \sigma_{daily}/\sigma_{season} \) is the ratio variance of daily variability about the seasonal mean to the interannual variability in seasonal means. The parameters \( a, b, \gamma \) and \( \delta \) are estimated during the fitting of the model. The residual variance \( \varepsilon \) is itself linearly dependent on the transformed metric of the mean (atanh(\( \text{corr}_{mean} \)) in the above example), the log variance ratio and the interaction between the two. The linear model for the residual variance is fitted to the log-transformed squared residuals \( \varepsilon \) after an initial fit with the local linear model. The final linear model used for prediction and comparison to the operational forecast is fitted using the inverse of the variance of \( \varepsilon \) as weights to take into account scaling of the residual variance.

To compare our expectations about the relationship of verification metrics of the index to verification metrics of the mean, we predict the correlation and RPSS of the index based on the daily-to-seasonal variance fraction and the correlation and RPSS of the operational system. To assess the significance of the difference between the expected skill derived from the fit to the toy model and the actual skill of the index for the operational forecast, we note that the difference (standardised with the standard error of the prediction) follows a \( t \)-distribution with the degrees of freedom determined by the window width of the local linear fit.

We account for multiple testing by controlling the False Discovery Rate as proposed by Benjamini and Hochberg (1995) and we use the tighter control on the False Discovery Rate as introduced in section 4 of Ventura et al. (2004).

3. Results

3.1. Toy model

In order to mimic the configuration of the operational System 4, we run the toy model simulating 34 forecasts of 90 days with 51 ensemble members and compute the seasonal mean and seasonal CC and DD indices from the resulting daily synthetic observed and forecast series. We repeat this procedure 500 times for each combination of the nominal correlation (\( a \)) and the ratio of daily variability to interannual variability in seasonal means (\( \gamma \)) and for different percentile thresholds.

In Figure 1, we show box and whisker plots of the distribution of correlation and RPSS values computed on the seasonal averages and CC and DD indices.

For positive \( a \) values we find the following: compared with the correlation for the seasonal mean quantity (dark grey boxes in Figure 1), the correlation for CC (grey boxes) and DD (white boxes) is reduced. This effect is strongest for indices computed with a threshold at the tail of the distribution of daily values (e.g. the 98th percentiles), skillful forecasts (\( a = 0.98 \)), and pronounced daily unpredictable variability (e.g. \( \gamma = 10 \), not shown). A slight decrease in correlation can even be found for forecasts with marginal skill (\( a = 0.2 \)), indices computed with a threshold at the 75th percentile and daily variability comparable to the variability in seasonal means (\( \gamma = 1 \) as shown in Figure 1). In general, the reduction of correlation is slightly more pronounced for the accumulated departures from the threshold (DD) compared to counts of exceedances of the threshold (CC). Similar differences are also found for the ranked probability skill score (RPSS), but the differences between RPSS for the indices and RPSS for the mean quantity are less pronounced.

For forecasts with negative skill (\( a < 0 \)), the correlation for the CC and DD indices are less negative than for the mean quantity. This effect is more pronounced for thresholds at the tail of the distribution, especially the 98th percentile. Furthermore, the count index (CC) shows consistently more negative correlation than the exceedance of threshold index (DD). While a comparable effect may be expected also for the RPSS, pronounced differences between the RPSS of the mean quantity and the RPSS of the indices are only found for the combination of strongly negative skill and high thresholds (Figure 1(f)).

When using the more appropriate Spearman rank correlation coefficient to assess correlation of count index forecasts from the toy model we find a slightly different pattern (not shown). In this case, the relationship of mean correlation and correlation of the index forecast is symmetric about zero without the levelling off of correlations for forecasts with negative correlation in the seasonal mean shown in Figure 1. Please note that the overall conclusions drawn are independent of the choice of correlation coefficient.

3.2. Operational forecasts

With the knowledge gained in section 3.1 we now analyse climate indices in a real-world seasonal forecasting system. In direct comparison with the counts (CC) and accumulated exceedances of thresholds (DD) presented in the previous section, we analyse counts and accumulated exceedances of daily percentiles of the climatological distribution based on daily minimum and maximum temperature forecasts.

In Figure 2 we show the distribution of correlation and RPSS for seasonal mean daily minimum and maximum temperature forecasts (dark grey boxes), counts of threshold exceedances (grey) and accumulated threshold exceedances (white) for the different percentile thresholds. Shown are skill metrics for land grid points in the Northern Hemisphere only. In contrast to Figure 1, results in Figure 2 are not stratified according to nominal correlation (parameter \( a \) in the toy model) as this is not known for the operational system.

Corroborating the findings from the toy model presented in section 3, we find that the correlation of forecasts of climate indices is at most as high as the correlation in the seasonal mean daily minimum and maximum temperature forecasts (upper panels in Figure 2). The reduction in correlation compared with the seasonal mean forecast is strongest for indices defined with respect to a threshold at the very tail of the distribution (e.g. defined with respect to the 98th and 2nd percentile respectively). Also, we find that the reduction in correlation is slightly more pronounced for exceedances of daily maximum temperatures in summer compared to temperatures falling below low daily minimum temperatures in winter.

As with the toy model, the effect on skill as measured by the RPSS is less pronounced, and consequently the RPSS for the seasonal mean and indices are rather similar (lower panels in Figure 2). A considerable reduction in RPSS can be found for indices defined relative to the 98th (2nd) percentile. Also, virtually no difference in RPSS is found when comparing counts and accumulated exceedances (grey and white boxes in Figure 2).

3.3. Using the toy model to predict skill in indices from skill in seasonal means

We compare the correlation and RPSS of the operational forecasting system presented in Figure 2 with the expected relationship between skill in seasonal means and indices based on the toy model as presented in section 3.1. First, we predict the correlation and RPSS of the forecasts of indices from the verification metric of the seasonal mean based on the toy model. In Figure 3 we show one example of such a prediction. For forecasts of summer (JJA) mean daily maximum temperature, we find significant positive correlations for most parts of the Northern Hemisphere except for areas in northern North America, northern Europe and Siberia (Figure 3(a)). Based on the fit of the local linear regression model to the toy model results, we predict the correlation of the DD90 index (Figure 3(b)). Predicted correlation of DD90 is slightly reduced, but follows closely the pattern of seasonal mean correlation. The actual correlation of DD90 derived from the operational forecasting system is shown in Figure 3(c) for comparison. The general spatial pattern of correlation of DD90...
These range from approximately $-5$ to $+5$ with the bulk of the residuals clustering around zero. For correlation, the median of the residuals for both count and accumulated threshold exceedances is slightly lower than zero, indicating that there is a tendency for the predictability (as measured by the correlation) of forecasts of indices in the operational system to be slightly worse than modelled by the predictability of the seasonal mean. For the RPSS, we find a similar tendency in winter except for the index forecast in the very tail. Residuals for the exceedances of the second percentile of daily minimum temperature in winter (rightmost group of boxes in Figure 4(d)) show a clear pattern of positive residuals for the count index and negative residuals for the accumulated threshold exceedance. Similarly, RPSS of exceedances of daily maximum temperature percentiles in summer (Figure 4(c)) is slightly enhanced with regard to expected RPSS based on the toy model for count indices but not for the accumulated threshold exceedances.

4. Discussion

We first start with a short discussion of the systematic effects found when verifying forecasts of indices from the toy model. In the following, we interpret the comparison of predictive skill of forecasts of indices and seasonal mean quantities of the operational forecasting system and discuss advantages and limitations of the proposed approach to predict skill of index forecasts of indices from the toy model for seasonal averages; the distribution of correlation and RPSS reflects the influence of sampling (500 random draws of 34 forecasts with 51 ensemble members). Boxes extend from the lower to the upper quartile with the median denoted by the solid vertical line. The whiskers extend to the extremes of the distribution or at most 1.5 times the interquartile range (box height) from the median. Further away than 1.5 times the interquartile range are marked with dots.

Figure 1. (a, b, c) Correlation and (d, e, f) RPSS of seasonal mean quantities (dark grey boxes) compared with CC (grey) and DD (white) indices computed for thresholds at the (a, d) 75th, (b, e) 95th and (c, f) 98th percentile, for index forecasts derived from cases with intraseasonal daily variability comparable to interannual variability in seasonal means (i.e. $\gamma = 1$). The parameter $\alpha$ denotes the nominal correlation of the toy model for seasonal averages; the distribution of correlation and RPSS reflects the influence of sampling (500 random draws of 34 forecasts with 51 ensemble members). Boxes extend from the lower to the upper quartile with the median denoted by the solid vertical line. The whiskers extend to the extremes of the distribution or at most 1.5 times the interquartile range (box height) from the median. Further away than 1.5 times the interquartile range are marked with dots.

Published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.
forecasts and to assess the significance of differences with the verification of the operational forecasts.

4.1. Understanding the toy model results

To get a better understanding of the effect on the correlation of accumulated threshold exceedances (DD) we use a simplified toy model. For simplicity, we compute DD on seasonal mean values with a deterministic forecast instead of computing DD on daily values with an ensemble forecast. That is, for a given correlation ρ, the observations and simplified seasonal forecasts are given by $x \sim N(0, 1)$ and $y = \alpha x + \varepsilon$ with $\varepsilon \sim N(0, 1 - \alpha^2)$. An example of 1000 such forecast observation pairs for different values of $\alpha$ is shown in Figure 5(a)–(c).

In analogy to computing accumulated threshold exceedances with daily forecasts, we illustrate the effect of introducing a threshold by setting all forecast and observation values smaller than the threshold (grey points in Figure 5(a) and (b)) to the threshold value (black points in Figure 5(a) and (b)). In cases of positive skill (Figure 5(a) and (c)), forecasts for observations that exceed the threshold often exceed the threshold themselves and most of the remaining pairs coincide at the threshold value. While the forecast observation pairs still follow the nominal correlation, actual correlation is reduced due to the lack of information about the relationship between forecasts and observations for pairs falling below the threshold. In cases of negative skill, the effect of computing indices is much more pronounced due to the fact that forecasts tend to fall below the threshold when observations do not (and vice versa, see Figure 5(b)).

The findings above can be summarised in one figure showing the dependence of correlation on threshold exceedance (Figure 5(c)). Reduction of correlation is moderate for positive skill with an increasing reduction with increasing threshold. For forecasts with negative correlation, the reduction in magnitude of the correlation is more pronounced, even for low thresholds. The Spearman rank correlation coefficient is less affected by this effect due to the fact that ties (multiple occurrences of zero) are taken into account. Please note that the results with the simplified toy model presented here are not directly comparable to the results presented in Figure 1. Due to the additional unpredictable daily variability in the toy model used in Figure 1, the correlation of forecasts of indices will be less strongly altered relative to the skill in the mean forecast than shown in Figure 5(c).

We conclude that within a well calibrated model environment, the correlation of the forecast of the mean quantity will be larger than that of the index if the model is skilful (i.e. $\alpha > 0$). The derivation of indices goes along with additional sampling uncertainty resulting in additional unpredictable noise. This additional noise reduces the correlation on average and increases the spread of estimates. Due to the assumption of no predictability in daily variability, this reduction in correlation is larger when the ratio of daily variability about the seasonal mean to the variability of seasonal means is larger. Please note that this reduction applies on average; individual draws with the toy model may well show correlations (or RPSS) for the index forecasts that exceed the correlations (RPSS) of the seasonal mean. For skilful forecasts ($\alpha = 0.9$) for example, 25% of the DD forecasts exhibit correlations that exceed the correlation of the seasonal mean when daily and seasonal variability are comparable ($\gamma = 1$), but this percentage quickly drops to zero as daily variability increases. This illustrates the effect of sampling variability which is also reflected in the size of the prediction intervals of the local linear regression introduced in section 2.5 (not shown).

Figure 2. Distribution of (a, b) correlation and (c, d) RPSS, for indices computed on (a, c) daily maximum temperature in summer (JJA) and (b, d) daily minimum temperature in winter (DJF). The indices are the seasonal mean (dark grey boxes) and counts (grey boxes) and accumulated exceedances (white boxes) of daily maximum and minimum temperatures defined with respect to the 75th, 90th and 95th daily percentiles of the climatological distribution of the ERA-Interim reanalysis and the forecasts. The distribution includes land grid boxes of the Northern Hemisphere only (Figure 3).
For negative correlation, the additional noise improves the forecasts. Hamilton et al. (2012) reported similar results when analysing extreme daily events and concluded that the forecast of daily extremes can only be improved by improving the forecast of the seasonal mean. However, if the model correlation is worse than climatology (negative correlation), the indices show better results, or in more accurate words, less negative correlation than the seasonal mean. For practical purposes, however, the results for forecasts with negative correlation are not relevant as such forecasts would be disregarded or improved by post-processing.

Please note that the toy model assumes no autocorrelation of the time series within each season. For the application to counts and accumulated threshold exceedances analysed here, this is irrelevant. To analyse indices describing the length of a period such as number of consecutive dry days or length of heat waves, the toy model is not directly applicable. Also, we assume Gaussian distributions for both the seasonal means and the daily variability about the seasonal mean. While this is a defendable assumption at least for temperature, mixtures of other distributions would be more appropriate to study other variables such as precipitation. We note that the toy model is general enough that both alternative distributions and additional features such as autocorrelation could be easily implemented. We do not study alternative distributions here, as such analyses are beyond the scope of this article.

4.2. Skill in operational forecasts

We find that the correlation and RPSS of forecasts of the temperature indices analysed is generally lower than the correlation and RPSS of seasonal mean daily minimum and maximum temperature. This reduction is mainly due to systematic effects related to the computation of the indices. The data truncation applied when computing the index decreases the signal-to-noise ratio of a potentially predictable signal by increasing the noise (see also discussion of toy model results in the previous section). This effect is well-known and has been previously highlighted to be responsible for differences in predictive skill (Hamilton et al., 2012). The use of a toy model allows us to quantify such systematic effects for different ensemble verification metrics as well as for forecasts with varying hindcast length and ensemble size.

Based on the empirical model fit to the toy model results, we predict the skill of the index forecast given the skill in the seasonal mean. The comparison of the predicted and actual skill values is further used to determine whether this empirical model can be used for prediction and if there is indication of enhanced (or reduced) skill in forecasts of indices in excess to what is expected due to predictability of the seasonal mean. To quantitatively compare the correlation and RPSS of the operational system with the predicted correlation and RPSS based on the toy model results, we compute p-values of the standardised residuals from the prediction shown in Figure 4. The fractions of land grid boxes in the Northern Hemisphere with residuals significantly (at the 5% level) different from zero are shown in Table 1. For most indices in both summer and winter, the fraction of significant grid boxes is close to 5%. This is the expected fraction of erroneous rejections when testing at the 5% level if the null hypothesis is in fact true.

As the above is a classical case of multiple testing, we also compute the fraction of significant departures from the model fit when taking into account multiple testing (see also section 2.5). After controlling the False Discovery Rate, only a small fraction of grid boxes with significant departures remain (numbers in brackets in Table 1). Field significant differences mainly occur for the RPSS of accumulated exceedances (DD). The majority of these are found in combination with strongly negative RPSS. Thus, we suspect that the occurrence of field significant departures is rather an artefact of the set-up of the local linear model than a real effect. Also, the departures that are field significant are patchy in space and no coherent patterns can be identified (not shown).

We find significant differences only in a small fraction of the northern hemispheric land grid points studied here. This indicates that the empirical model provides a good approximation of the
verification results of index forecasts with the operational system and the empirical model may therefore be used to predict the expected verification results for forecasts of indices. Also, we conclude that there is no strong indication that predictive skill in forecasts of counts and accumulated threshold exceedances is reduced or enhanced compared with predictive skill in seasonal means beyond the systematic effects discussed above. These conclusions are likely dependent on the assumptions of the toy model; in particular, we assume that daily variability about the seasonal mean is Gaussian and that the forecasts are well calibrated and reliable.

The assumption of Gaussian daily variability only affects the accumulated exceedances. The count index is by construction independent of the underlying distribution of daily values as counts are computed relative to the daily percentiles of the forecasts and verifying observations respectively. Thereby, the probability of exceeding the threshold is constant as opposed to when using fixed or seasonal thresholds for which the annual cycle becomes relevant. In contrast, the distribution of daily values above (or below) the threshold matters for accumulated exceedances as these departures are summed up. Consequently, the distributional assumptions and departures thereof may affect our interpretation of the significance of the results for the DD index. Whether a more heavy-tailed distribution for example leads to reduced or enhanced verification metrics is not obvious. This dependency may be studied using the toy model, but for the present study, we think this is of minor importance.

Using daily percentiles as the threshold against which to compute indices has further advantages for the comparison with seasonal mean forecasts. The constant occurrence probability eliminates the effect of degrading skill with increasing forecast lead time that would on average further reduce the skill of forecasts of events that are more likely to occur in the latter part of the season (Hamilton et al., 2012).

By computing the indices relative to the forecast and observed daily percentiles and by using two verification metrics that are independent of mean biases, we seek to maximise calibration of the forecasts. The major caveat when comparing the operational system with the toy model is that we assume that seasonal mean forecasts are reliable (in a statistical sense). Seasonal forecasts have often been found to be overconfident (Weigel et al., 2008, 2009; Weisheimer and Palmer, 2014). While estimates of the reliability of seasonal mean forecasts such as the spread to error ratio are readily available, we found that these are subject to considerable sampling variability when applied at the grid-point scale (see also Siegert et al. (2016) for more in-depth discussion on uncertainty in verification results). We refrain from factoring in the lack of reliability of seasonal mean forecasts in our assessment, but note that lack of reliability in seasonal means is expected to affect the seasonal mean forecast and the index forecast similarly (i.e. by leading to ensemble forecasts that are too narrow and not centred about the true response).

Finally, the toy model itself could have been chosen differently. Weigel and Bowler (2009) discuss variants of the toy model chosen here. Hamilton et al. (2012) have used an alternative set of synthetic forecasts to assess predictability in forecasts of climate indices. Their synthetic forecasts are constructed from the operational forecasts and the verifying observations directly and are thus arguably less prone to violations of the underlying assumptions than our Gaussian toy model. Their approach is limited in that the results might depend on the model used and that ensemble forecasts cannot be easily generated and therefore is not applicable to probabilistic ensemble forecast verification metrics. We see our approach as a valuable alternative and we note that the results with both approaches are largely consistent.

To conclude the discussion, we present how the above approach to assess the significance of differences between verification metrics of the seasonal mean and forecasts of indices could...
be used in practice. We present maps of correlation of daily maximum temperature forecasts in summer and daily minimum temperature forecasts in winter (Figure 6(a) and (b)). Along with the correlation of the seasonal mean quantity, we present the difference in correlation for the count and accumulated threshold exceedance index for a specific percentile threshold with the correlation in the seasonal mean (Figure 6(c)–(f)). Regions are stippled where correlations of the index forecasts significantly differ from the correlation we would expect due to predictability of the seasonal mean only. Stippling in Figure 6 indicates local significance. None of the grid points is globally or field significant after controlling the False Discovery Rate (see also Table 1).

Correlation of seasonal mean daily maximum temperature in summer is generally higher in lower latitudes compared to polar regions with the exception of Greenland where correlations exceed 0.6 (Figure 6(a)). In winter, seasonal mean daily minimum temperature shows the same general pattern with somewhat reduced correlations compared to summer (Figure 6(b)). Correlations of forecasts of the \( CC_{90} \) and \( DD_{90} \) indices in summer and the \( CC_{10} \) and \( DD_{10} \) indices in winter are generally slightly lower than those for seasonal means (predominance of blue colours in Figure 6(c)–(f)). Correlation of the index forecast exceeds correlation of the seasonal mean where the seasonal mean correlation is close to zero (e.g. southern Sahara and northwestern Russia in summer). One notable exception of this pattern is found in northwestern North America in winter, where index forecasts show more correlation than the seasonal mean despite significant correlations in the seasonal mean (Figure 6(d) and (f)). As none of the differences shown in Figure 6 are field significant, the deviations noted above could still be due to chance. Thus, the above example illustrates and supports the findings presented earlier.

5. Conclusions

The use of climate indices represents a direct way to characterize aspects of the climatological distribution that are related to specific applications. In this study we analyse how the skill of seasonal forecasts of climate indices compares to the skill in forecasts of the meteorological quantity from which these indices are derived. Using a toy model, we show that in situations with positive skill, the underlying mean quantity exhibits better skill than the index, and cumulative indices are slightly less skilful than count indices. This effect is strongest for skilful forecasts and for indices based on rare events. In cases of negative correlation, the findings are reversed in that the skill of the indices is less negative than the skill of the seasonal mean. The reason for this effect is illustrated with the aid of a simplified toy model. Due to the geometry of the problem, negative correlations cannot be sustained for indices with thresholds in the tail of the distribution since the introduced sampling uncertainty results in more random and thus fewer bad forecasts. In terms of skill and correlation, the operational forecasting system exhibits similar behaviour to the toy model. Skill and correlation of forecasts of indices are generally lower than those of forecasts of seasonal mean quantities. Also, this reduction is stronger for indices of rare events and more pronounced for accumulated threshold exceedances compared to counts of events. We further find that the differences between the verification of forecasts of indices from the operational forecasting system and the prediction of expected verification metrics based on the toy model results are not significantly different. We conclude from the lack of significance that: (i) we can use the toy model to predict the expected skill of forecasts without access to the daily forecast data and (ii) there is no evidence for systematically enhanced (or reduced) predictability of indices in addition to what is expected...
### Table 1. Percentage of land grid points with significant (at the 5% level) deviations in their scores from the expected score due to seasonal predictability only.

| Score | Index | 75th percentile | 90th percentile | 95th percentile | 98th percentile |
|-------|-------|-----------------|-----------------|-----------------|-----------------|
|       |       |                 |                 |                 |                 |
|       | Correlation in JJA |                 |                 |                 |                 |
|       | CC    | z > 0 4.4% (0.1%) | 4.8% (0%)       | 4.3% (0%)       | 5.1% (0%)       |
|       |       | z < 0 2.8% (0.1%) | 3% (0%)         | 2.9% (0%)       | 3.6% (0%)       |
|       | DD    | z > 0 1.6% (0%)  | 1.8% (0%)       | 1.4% (0%)       | 1.5% (0%)       |
|       |       | z < 0 5.4% (0.1%) | 5.1% (0%)       | 5.4% (0%)       | 6.4% (0%)       |
|       |       | z > 0 3.5% (0%)  | 3.7% (0%)       | 4% (0%)         | 5.3% (0%)       |
|       |       | z < 0 2% (0.1%)  | 1.5% (0%)       | 1.4% (0%)       | 1.1% (0%)       |
|       | RPSS in JJA |                 |                 |                 |                 |
|       | CC    | z > 0 5.3% (0.1%) | 5.2% (0.1%)     | 1.5% (0%)       | 1.4% (0.1%)     |
|       |       | z < 0 3.1% (0.1%) | 2.9% (0%)       | 1% (0%)         | 1.2% (0%)       |
|       | DD    | z > 0 2.2% (0%)  | 2.3% (0%)       | 0.5% (0%)       | 0.2% (0.1%)     |
|       |       | z < 0 9.1% (1.4%) | 6.9% (0.4%)     | 2.9% (0%)       | 3.4% (0.2%)     |
|       |       | z > 0 3.5% (0.1%) | 3.7% (0%)       | 4% (0%)         | 5.3% (0%)       |
|       |       | z < 0 2% (0.1%)  | 1.5% (0%)       | 1.4% (0%)       | 1.1% (0%)       |
|       |       | z > 0 5.3% (0.1%) | 5.2% (0.1%)     | 1.5% (0%)       | 1.4% (0.1%)     |
|       |       | z < 0 3.1% (0.1%) | 2.9% (0%)       | 1% (0%)         | 1.2% (0%)       |
|       |       | z > 0 2.2% (0%)  | 2.3% (0%)       | 0.5% (0%)       | 0.2% (0.1%)     |
|       |       | z < 0 9.1% (1.4%) | 6.9% (0.4%)     | 2.9% (0%)       | 3.4% (0.2%)     |
|       |       | z > 0 3.5% (0.1%) | 3.7% (0%)       | 4% (0%)         | 5.3% (0%)       |
|       |       | z < 0 2% (0.1%)  | 1.5% (0%)       | 1.4% (0%)       | 1.1% (0%)       |
|       | Correlation in DJF |                 |                 |                 |                 |
|       | CC    | z > 0 3.2% (0%)  | 5.3% (0%)       | 4.8% (0%)       | 6.1% (0%)       |
|       |       | z < 0 2.1% (0%)  | 3.2% (0%)       | 2.7% (0%)       | 3.3% (0%)       |
|       | DD    | z > 0 1.2% (0%)  | 2.1% (0%)       | 2.2% (0%)       | 2.9% (0%)       |
|       |       | z < 0 6.2% (0%)  | 6.4% (0%)       | 6.1% (0%)       | 5.3% (0%)       |
|       |       | z > 0 2.8% (0%)  | 3% (0%)         | 3.1% (0%)       | 3.2% (0%)       |
|       |       | z < 0 3.4% (0%)  | 3.4% (0%)       | 3% (0%)         | 2.1% (0%)       |
|       | RPSS in DJF |                 |                 |                 |                 |
|       | CC    | z > 0 5.2% (0.2%) | 5.3% (0.2%)     | 2% (0.2%)       | 1.4% (0.1%)     |
|       |       | z < 0 1.7% (0%)  | 1.6% (0%)       | 0.6% (0%)       | 0.5% (0%)       |
|       | DD    | z > 0 3.6% (0.2%) | 3.7% (0.2%)     | 1.4% (0.2%)     | 1% (0.1%)       |
|       |       | z < 0 8.6% (1.2%) | 7.5% (0.7%)     | 3.9% (0.3%)     | 5.9% (0.8%)     |
|       |       | z > 0 2.7% (0.1%) | 1.9% (0%)       | 0.3% (0%)       | 0.7% (0%)       |
|       |       | z < 0 5.9% (1.1%) | 5.6% (0.7%)     | 3.6% (0.3%)     | 5.1% (0.8%)     |

The percentages in brackets denote significance after accounting for multiple testing by controlling the False Discovery Rate. The fraction of positive and negative deviations is shown in the respective second and third rows for each index.

![Figure 6](image.png)

**Figure 6.** Correlation of seasonal mean daily (a) maximum temperature in JJA (1981–2014) and (b) minimum temperature in DJF (1982–2015) and difference in correlation of counts (c) above the 90th percentile and (d) below the 10th percentile, and (e, f) of accumulated threshold exceedances respectively. Blue areas in (c–f) indicate that the correlation of the index forecast is lower than the correlation of the seasonal mean quantity. Stippling in panels (a) and (b) indicates correlations significantly (at the 5% level) larger than zero. Stippling in panels (c–f) indicates correlations for the index that are significantly different from what is expected due to the correlation in the seasonal mean and predictability of the seasonal mean only. None of the differences in (c–f) are field significant after controlling the False Discovery Rate (see text and Table 1).

from the seasonal mean. Prediction of forecast skill based on the toy model results may be interesting from an application perspective as downloading and processing daily forecast data to verify forecasts of indices is computationally challenging. It should also be noted that this approach can be transferred to other model systems easily as the identified relationship does not depend on the forecasting model.

The toy model results, on the other hand, allow us to formulate a data-driven null hypothesis for significance testing. Thanks to its generality, this approach lends itself naturally to study a range of related questions, such as the dependence of predictive skill on errors in various aspects of the forecast (e.g. systematic and conditional bias, misrepresentation of daily variability, etc.). The toy model may also be expanded by including autocorrelated noise to study indices of sequences of events such as cold spell duration or heat-wave indices.

We conclude that forecasts of indices may be more relevant to users, yet at the same time such forecasts tend to be less skillful than forecasts of seasonal mean quantities. Our analyses suggest, however, that forecasts of counts and accumulated exceedances...
of thresholds can be issued without major loss in skill as long as these indices characterise events that are not very rare.

Acknowledgements

The research leading to these results has received funding from the European Union’s Seventh Framework Programme [FP7/2007–2013] under Grant Agreement no 308291. The analysis was carried out using the R packages easyVerification (MeteoSwiss, 2016) and SpecsVerification (Siegent, 2015). The authors wish to thank the Swiss National Supercomputing Centre (CSCS) for providing the technical infrastructure. The authors would also like to thank two anonymous referees for helpful comments on the manuscript.

References

Benjamin Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300, doi:10.2307/2346101.

Brands S. 2013. Skillful seasonal predictions of boreal winter accumulated heating degree-days and relevance for the weather derivative market. J. Appl. Meteorol. Climatol. 52: 1297–1302, doi:10.1175/JAMC-D-12-0303.1.

Cleveland WS, Devlin SJ. 1988. Locally weighted regression: An approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83: 596–610, doi:10.1080/01621458.1988.10478639.

Dee DP, Uppala SM, Simmons AI, Berrisford P, Poli P, Kobayashi S, Andrae M, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Holm EV, Isaksen L, Källberg P, Köhler M, Matricardi M, McNally A, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Thépaut J-N, Vitart F. 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Q. R. J. Meteorol. Soc. 137: 553–597, doi:10.1002/qj.828.

Eade R, Hamilton E, Smith DM, Graham RJ, Scaife AA. 2012. Forecasting the number of extreme daily events out to a decade ahead. J. Geophys. Res. 117: D21110, doi:10.1029/2012JD018015.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. J. Appl. Meteorol. 8: 985–987, doi:10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Hamilton E, Eade R, Graham RJ, Scaife AA, Smith DM, Maidens A, MacLachlan C. 2012. Forecasting the number of extreme daily events on seasonal timescales. J. Geophys. Res. 117: D03114, doi:10.1029/2011JD016541.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Grose J, Kanakidou M, Kistler R,<context省略>