Sampling Errors in Seasonal Forecasting

STEPHEN CUSACK AND ALBERTO ARRIBAS
Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 15 February 2008, in final form 2 September 2008)

ABSTRACT

The limited numbers of start dates and ensemble sizes in seasonal forecasts lead to sampling errors in predictions. Defining the magnitude of these sampling errors would be useful for end users as well as informing decisions on resource allocation to minimize total system error. A numerical experiment has been designed to measure them, and results indicate that sampling errors are substantial in state-of-the-art seasonal forecast systems. The standard solution of increasing sample sizes is of limited benefit in seasonal forecasting because of restrictions imposed by resource costs and nonstationary observations. Alternative options, based on the postprocessing of forecast and hindcast data, are presented in this paper. The spatial and temporal aggregations of data together with the appropriate use of theoretical distributions can reduce the effect of sampling errors on forecast quantities by an amount equivalent to increasing samples sizes by a factor of 4 of more, with insignificant losses of forecast information. These postprocessing techniques can be viewed as cost-effective methods of reducing the effects of sampling errors in seasonal forecast quantities.

1. Introduction

The future state of the atmosphere is influenced by chaotic internal dynamics (e.g., Lorenz 1969; Lau 1981; Hendon and Hartmann 1985; Branstator 1995) that can amplify the uncertainties in forecast system initialization and formulation to produce different possible seasonal climate states. Seasonal forecasting systems employ ensembles of simulations to sample these uncertainties, and the prediction of a meteorological quantity is most appropriately viewed as a probability distribution function (pdf). Typically, end users are presented with information derived from the forecast pdf together with an appropriate assessment of the performance of the forecasting system. Seasonal prediction systems use a number of historical reforecasts, usually referred to as hindcasts (e.g., Graham et al. 2005), to generate a set of independent forecasts for system assessment.

The forecasted meteorological quantities and associated skill assessments are based upon limited numbers of ensemble members and start dates, and sampling errors are an inevitable consequence of these finite data sizes. The effect of sampling errors on forecast information is required by end users. Furthermore, there is competition for resources to minimize errors from diverse sources such as forecast initialization, model error, and sampling errors. The optimum resource allocation to maximize the usefulness of forecasts requires the measurement of errors.

Kumar and Hoerling (2000, hereafter KH00) and Kumar et al. (2001, hereafter KBH01) designed a numerical experiment to investigate the effect of finite ensemble sizes on the assessment of system quality. Results from their experiment revealed that errors due to limited numbers of ensemble members produced a negative bias on commonly used measures of forecast quality. Section 2 contains a description of a numerical experiment designed to investigate the effect of finite ensemble sizes on the assessment of system quality. Results from their experiment revealed that errors due to limited numbers of ensemble members produced a negative bias on commonly used measures of forecast quality. Section 2 contains a description of a numerical experiment designed to extend the work of KH00 and KBH01 in three respects. First, the revised numerical experiment captures the effects of a finite number of historical states as well as limited ensemble sizes. The upper limit on the sampling of historical states in seasonal forecasting is governed by nonstationary observations and resource costs and is sufficiently small to have a significant impact upon assessments of system quality. Second, the impact of these sampling errors upon forecast variables are quantified in addition to their effects upon system assessment metrics since both quantities are supplied to end users. Third, the artificial datasets are based upon the characteristics of a real
seasonal hindcast set, permitting sampling errors to be estimated for different geographical locations at the spatial resolution of the forecast system. This third revision allows sampling errors to be quantified and displayed as global maps.

Estimates of the effects of the sampling errors on typical forecast information are presented in section 3. Our results indicate significant uncertainties in the value of a typical forecasted variable (i.e., tercile probabilities of near-surface air temperature). Furthermore, it is found that the estimated value of a standard assessment metric has significant uncertainty, in addition to the bias reported in KBH01. Section 4 contains discussions on the use of simple postprocessing methods to reduce the effects of sampling errors upon forecast information. The impacts of these methods are shown for a practical example. A development version of the Met Office Global Seasonal Forecast System, version 4 (GloSea4), is configured to have data samples 60% smaller than its predecessor, GloSea3, and the effects of applying the postprocessing methods to GloSea4 output are shown. Section 5 contains a summary of the main findings and some directions for future work.

2. Experimental setup

An experiment to measure the effect of sampling errors on forecast information in a robust and accurate fashion is now described. Traditionally, because of the relatively low skill of seasonal forecasts, the information issued to end users consists of a forecast quantity together with an estimate of the quality of such a prediction. This forecast information is extracted from an operational forecast consisting of $N_f$ members and a hindcast dataset comprising $S_h$ forecasts, or start dates, with $N_s$ ensemble members in each hindcast. Many sets of forecasts with associated hindcasts are required for a robust estimate of the effects of sampling errors in a single forecast and hindcast set. However, it is not feasible to generate many sets of forecast and hindcast data using an operational seasonal forecasting system because of computational resource limitations. Instead, a numerical experiment has been designed to mimic the results of an operational prediction system. The method is described in section 2a. Section 2b describes the method of quantifying the effects of sampling errors on estimates of a typical forecast quantity. Section 2c describes the effects of sampling errors on estimates of typical assessment metrics.

a. The creation of forecast and hindcast data

The forecast and hindcast data necessary for estimating the effects of sampling errors in forecast information are created using a method closely following that described in KH00. Briefly, many artificial forecast/hindcast sets are created by randomly sampling a theoretical pdf constructed from data obtained from GloSea3, the third version of the Met Office Global Seasonal Forecast System operational since 2005. GloSea3 is a state-of-the-art seasonal forecast system with similar configuration and performance to those in other operational centers. Results based on data from this system should be generally applicable to systems worldwide. Each hindcast set consists of $S_h$ start dates and $N_h$ ensemble members. Each hindcast set has an associated forecast comprising $N_f$ ensemble members.

The analysis in this paper focuses on near-surface air temperatures. A Gaussian statistical model is considered a good representation of the true underlying pdfs of this variable; therefore, both the operational forecast and the hindcast pdfs are fully specified by the mean ($\mu_f$) and standard deviation ($\sigma_f$). The values of the Gaussian parameters are estimated for each grid box in the seasonal climate model, permitting the analysis of the effect of sampling error on forecast information at this spatial scale. The values of these two parameters are derived from GloSea3.

The standard deviation of a quantity $x$ for the $s$th start date in the hindcast set, $\sigma_{f,s}$ ($s = 1, \ldots, N_h$) is estimated as follows:

$$
\sigma_{f,s} = \sqrt{\frac{\sum_{n=1}^{N_s} (x_{f,s,n} - \bar{x}_f)^2}{N_h - 1}}^{0.5},
$$

where $x_{f,s,n}$ is the value of the $n$th ensemble in the $s$th start date of the hindcast and $\bar{x}_f$ is the ensemble mean value of $x$ for the $s$th start date. Here $\sigma_f$ is specified to be the average of $\sigma_{f,s}$ over all start dates:

$$
\sigma_f = \frac{\sum_{s=1}^{S_h} \sigma_{f,s}}{S_h}.
$$

The mean value of a forecast, $\mu_f$, is a single value drawn from a Gaussian random number generator with a mean value of zero and a standard deviation $\sigma_f$ equal to the standard deviation of the ensemble mean forecast values over all appropriate start dates in the hindcast set:

$$
\sigma_s = \sqrt{\frac{\sum_{s=1}^{S_h} (x_s - \bar{x}_f)^2}{S_h - 1}}^{0.5}.
$$
where $\bar{x}$ is the mean value over all ensemble members and start dates.

An artificial forecast is created by drawing $N_f$ numbers from a Gaussian random number generator with a mean value of $\mu_f$ and a standard deviation of $\sigma_f$. A forecast for the $s$th start date in the hindcast set is created by drawing $N_h$ numbers from a Gaussian random number generator with a mean value of $\mu_f$ and a standard deviation of $\sigma_f$.

The assessment of hindcasts in operational seasonal forecasting centers involves the comparison of hindcasts with associated observations. This numerical experiment defines the observed value for the $s$th start date to be a random number drawn from a Gaussian distribution with parameters from the corresponding $s$th hindcast and so implicitly assumes a perfectly reliable probabilistic forecast. An assessment measure with sensitivity to the reliability of predictions will score better in the numerical experiment than in practice. Therefore, for the experiment to be of practical use, the effect of the sampling errors on the scores should have little dependence on the value of the score. Analysis of sampling errors in regions of the world with a wide variety of values of scores reveals that sampling error characteristics have little dependence on the score value.

In summary, the numerical experiment allows for the creation of an unlimited number of hindcast sets and associated operational forecasts. The results presented in later sections are based upon standard forecast quantities and assessments from each of 300 independent samples of operational forecast and hindcast sets. These results are compared to reference values to enable the effects of the sampling errors to be estimated. The reference values are estimated using the true underlying Gaussian parameters for the forecast and hindcast pdfs together with 1000 start dates in the hindcast set.

The forecast quantity requires the definition of the climatological pdf ($F_{\text{clim}}$) to define the location of the climatological terciles, as well as the forecast pdf ($F_f$).

In practice, the definition of $F_{\text{clim}}$ is based on a hindcast set consisting of $S_h \times N_h$ members, while $F_f$ is based on the sample size $N_f$. An estimate of the forecast quantity ($X_f$) is obtained for each of 300 forecast and hindcast sets generated by the numerical experiment. The values of the forecast quantity based upon the exact values of the Gaussian parameters and corresponding to very large samples ($X_f$) are calculated, and the sampling error effect is defined as the difference between $X_f$ and $X_t$. Because of the experimental design, the sampling error is expected to have no bias; therefore, only the spread of the sampling error needs to be considered. In future sections, the spread of the sampling error is measured by the 90% prediction interval (PI).

**c. The forecast assessment metrics**

The assessment of a forecasting system can be summarized using a variety of metrics (e.g., Murphy et al. 1989; Wilks 2005; Cusack and Arribas 2008). These general summary scores are designed to assess different aspects of forecast quality with a sensitivity specified by the use of a penalty function to punish differences from desired behavior. Two assessment metrics are used in this paper: the Brier score (BS; Brier 1950) and the Information Quantity score (IQ; Cusack and Arribas 2008).

The BS is a metric of the general performance of probabilistic forecasts and is used commonly in seasonal forecasting. The BS can be used to measure the probability of occurrence within terciles of the climatological pdf, thus, ensuring consistency between the score and the forecast quantity selected in section 2b. An unbiased definition of the terciles of the climatological pdf is achieved by including all ensemble members of the hindcast set except those from the start date being assessed. The BS is evaluated for each artificial hindcast set and can be converted to a skill score by the use of reference values of BS, to form the Brier skill score (BSS). A climatological forecast is chosen as the reference. BSS values of unity indicate perfect deterministic forecasts, while scores of zero are no better than the reference forecasts and negative scores indicate a forecast system performing worse than the reference.

Cusack and Arribas (2008) decomposed the usefulness of forecast information to end users into two independent components. The statistical consistency between the forecast and observations is measured by the full-pdf reliability ($R_{\text{pdf}}$), while the IQ measures the extra information contained in the forecast relative to a level of knowledge in the absence of such a forecast.
These scores are sensitive to the full details of all forecast pdfs in a hindcast set in contrast to the BS. Performance metrics that are sensitive to a greater amount of detail in the forecast pdf will be more sensitive to the effects of sampling errors. This associated increase in sensitivity to the effects of sampling errors can bring new information on the effects of sampling errors on performance metrics. The impact of sampling errors on the IQ score is presented in the following section. The results for $R_{pdf}$ are similar to those for IQ.

The estimation of the IQ score proceeds by sorting forecasts into $M$ homogeneous subsamples. Each subsample has an associated hindcast pdf representing all hindcasts in the subsample. The predictand is discretized into one of $N = 8$ bins, yielding hindcasts $F_{i,m}$, where $i = 1, \ldots, N$, $m = 1, \ldots, M$. The IQ is a positively oriented score, with a value of 0 indicating no new information and a value of 1 indicating the maximum possible amount of new information:

$$IQ = \frac{\sum_{m=1}^{M} Y_m}{2M},$$

where $Y_m$ is defined as

$$Y_m = \frac{\sum_{i=1}^{N} |F_{i,m} - C_i| \Delta_i}{\sum_{i=1}^{N} \Delta_i},$$

where $\Delta_i$ is the width of the bin (in the units of the predictand) and $C_i$ is the discretized climatological pdf.

### 3. The uncertainties in forecast information

The results presented in this section are based upon similar sample sizes to GloSea3, the operational seasonal forecast system at the Met Office from 2005. GloSea3 uses a hindcast set consisting of 15-member ensemble forecasts starting on the first of every month from 1987 to 2001, while the corresponding operational forecast comprises a 41-member ensemble. The following results are for the December–January–February (DJF) seasonal mean near-surface air temperatures for hindcasts and forecasts initialized on 1 November. The following results have little dependence on the forecast start month and lead times. The forecast quantity (near-surface air temperature) is compatible with the Gaussian theoretical distribution used in the numerical experiment.

#### a. The effects of sampling errors in the forecast quantity

The expected bias in the value of the forecasts quantity is zero; therefore, only the spread (defined as 90% of the PI) in the predicted value need be examined. The results shown in Fig. 1 provide a global summary of the uncertainty in this forecast quantity. Results in the Southern Hemisphere were similar to those of the Northern Hemisphere (not shown). These results indicate that many regions of the world require a predicted tercile probability differing by a wide margin from the climatological value of 0.33 for a forecast to be different from climatology at a statistically significant level. Also shown in Fig. 1 are the effects of increasing the sample size to values close to the maximum currently affordable at operational forecasting centers (estimated as a forecast ensemble of 100 members and a hindcast set comprising 40 ensemble members in each of 30 start dates). The uncertainty of the forecast quantity is reduced by this larger system but remains substantial. The 90% PI of a climatological forecast by this maximum-sized sample lies between about 0.23 and 0.43.

#### b. The effects of sampling errors on forecast assessment

The numerical experiment produces perfectly reliable forecasts by design, since observations and forecasts are based upon the same underlying pdfs. Perfect reliability ensures that BS values are greater than or equal to the BS values of the reference climatological forecast, or equivalently, that BSS values are greater than or equal to zero. Figure 2 provides a particular perspective on the effects of sampling errors on BSS values for lower tercile forecasts. Grid boxes with greater than 95% of all artificial hindcast sets having BSS values exceeding zero are shaded, indicating forecasts with sufficiently small sampling errors to be distinguished from the climatological reference forecast at the 5% significance level. The shaded regions in Fig. 2a highlight regions in which forecasts are detectably different from the reference forecast and potentially useful in the context of this numerical experiment. Note that the spatial extent of the shaded regions in Fig. 2a represent an upper bound on the detection of useful BSS values since the reliability of GloSea3 forecasts is expected to be less than that of the numerical experiment. Results for the middle and upper terciles are similar to the results shown in Fig. 2a.

Larger hindcasts sets have smaller sampling errors. Figure 2b is the corresponding plot to Fig. 2a for a hindcast set consisting of 30 start dates, and 40 ensemble members in each forecast. The number of grid points
significantly different from climatology at the 5% level is slightly more widespread with the larger hindcast set. However, the situation in the extratropics, and especially over land, shows little improvement.

The results displayed in Fig. 2 indicate that sampling errors have a very significant impact upon the assessment of a state-of-the-art seasonal forecasting system. The skill does not differ from climatology at the 5% significance level in most areas of the globe, yet the assumption of perfectly reliable forecasts in the experimental design ensures BSS values are greater than or equal to zero in the absence of sampling errors. Therefore, for most regions of the world, modern operational seasonal forecasting systems are unable to distinguish the skill of the system from a climatological forecast because of sampling issues. The expense in terms of computation, data storage, and human resources is not rewarded with robust detection of regional skill.

Figure 3 is similar to Fig. 1 and shows the 90% PI for IQ scores area averaged over the northern extratropics and the tropics for a GloSea3 hindcast configuration. The true value, defined using very large samples of forecast start dates together with perfect knowledge of the forecast and climatological pdf, is also shown in Fig. 3. It can be seen that the GloSea3 hindcast set has a bias that is comparable with the 90% PI. The sampling errors generate differences between the forecast and climatological pdfs that are spurious and lead to artificially high IQ scores. Furthermore, the 90% PI is similar between both geographic regions yet the expected true value is different, suggesting that the effects of sampling errors has little dependence on model errors for this particular metric.

4. Methods to reduce uncertainties in forecast information

As shown in previous sections, the numbers of ensemble members and hindcast start dates typically used in seasonal forecast systems produce sampling errors that have a significant impact on forecast information. Furthermore, increasing sample sizes to the maximum achievable with current computational resources does not significantly reduce the effect of these errors. Post-processing methods to reduce sampling errors are explored in this section.

a. Theoretical statistical distributions

Forecast and climatological pdfs for various meteorological quantities are known to be good fits to various theoretical distributions. For example, temperatures, and to a greater extent their temporal average, are
known to fit the Gaussian distribution. Precipitation data are a good fit to the gamma distribution, wind speed data have a shape corresponding to the Weibull distribution, and cloud amounts are well represented by the beta distribution. The use of appropriate theoretical distributions supplies new information thereby reducing uncertainties.

The reduction in uncertainty achieved by using theoretical distributions was estimated using the numerical experiment described in section 2. Each ensemble forecast of data values is used to estimate the parameters of a Gaussian distribution from which the issued forecast information is derived. The numerical experiment is the same as previously described in all other respects.

Figure 4 is analogous to Fig. 1 and displays the 90% PI for lower tercile forecasts of near-surface temperature. The fitting of statistical models to data samples reduces the effects of sampling errors by 10%–15%. Error magnitudes are expected to be proportional to the inverse of the square root of the sample size; therefore, such reductions in the random sampling error correspond to a 20%–40% increase in sample sizes. In this context, the use of statistical models cause a significant reduction in the effects of the sampling errors on forecast quantities.

The impact of applying the Gaussian model upon forecast assessment scores was examined in a similar fashion for both BSS values and IQ scores. It was found that the reduction in the spread of BSS values was small relative to the width of the 90% PI (not shown). On the contrary, the use of the Gaussian model leads to a significant reduction in the bias of estimates of the IQ score, as can be seen in Fig. 3 which shows the spread (90% PI) for IQ scores area averaged over the northern extratropics and the tropics. The true value, defined using very large samples of forecast start dates together with perfect knowledge of the forecast and climatological pdf, is also shown in Fig. 3 as an asterisk. The reason for this is that the IQ score was estimated by differencing the probability densities of the forecast and climatological pdfs with a discretization of eight bins spanning the full range of forecast values. The small sample sizes of both pdfs lead to large amounts of noise in the estimates of the probability density in each bin. This noise increases differences between the forecast and climatological pdfs, hence, inflating estimates of the IQ score. The use of a smoothly varying theoretical pdf improves the estimates of the IQ score.

b. Spatial aggregation of data

The comparison of the benefits of the reduction in the sampling error due to spatial aggregation with the accompanying costs due to the loss of information in the forecast signal forms the basis of the decision on whether to include spatial aggregation methods. The spatial scales of the signals at seasonal time scales are much bigger than the current operational forecasting system’s resolution (typically wavenumbers 10 and 100, respectively). Therefore, the spatial averaging of forecast information to a lower spatial resolution will lead to a little loss of signal and relatively small benefits from spatial aggregation would be sufficient to justify its use since the detrimental effects of loss of signal are small.
In the present study we analyze the result of spatial aggregation based upon data at the grid point together with data from its eight neighbors. Forecasts of seasonal mean anomalies have a larger spatial scale than the dimensions of a grid box. Therefore, anomalies at neighboring grid boxes are not independent and the spatial aggregation of data from \( X \) grid boxes will give an increase in the effective sample size by a factor less than \( X \). The method of generating artificial forecast data in the numerical experiment was altered to take account of the spatial dependence of forecast values and so obtain a more accurate and realistic estimate of the effect of spatial aggregation. This method is now described.

The spatial dependence of forecast values is accounted for by generating forecast values at the point \( j + 1 \) that are partly dependent upon the values at the central point \( j \). The method is an adaptation of a first-order autoregressive process [\( \text{AR}(1) \); e.g., Wilks 2005], with time being substituted by the spatial domain. The forecast value at \( j + 1 \) is the sum of two components, namely, the ensemble mean value and the anomaly from this mean, and both of these components are separately modeled as \( \text{AR}(1) \) processes. The degree of dependence between the values at point \( j \) and \( j + 1 \) is controlled by what is usually called the autoregressive parameter \( (\rho) \) between the two points. All values of \( \rho \) are based upon GloSea3 hindcasts. The ensemble mean value at point \( j + 1 \) uses a value of \( \rho \) based upon the correlation of the respective ensemble mean forecasts, whereas the anomaly component at the same point \( j + 1 \) uses a value of \( \rho \) based upon the correlation of the corresponding anomalies relative to the ensemble mean between points \( j \) and \( j + 1 \). To ensure the correct variance, the ensemble mean and anomaly components are both modified to produce the appropriate variance at \( j + 1 \) before combining to yield the forecast value. This procedure is repeated for all eight neighbors of the central grid point. When the point \( j + 1 \) is the central grid point, new data is generated for the halo of its eight adjacent neighbors, based on the above method.

The variable to be spatially aggregated requires consideration. The effects of sampling errors on the forecast quantity produce uncertainties in the measured value, with no expected bias. The reduction of the spread in values of the forecast quantity will be achieved by the spatial aggregation of this quantity. However, there are more factors to consider for forecast assessment metrics. KH00 and KBH01 show that sampling errors due to finite numbers of ensemble members produce a bias in the estimate of common assessment measures. Increasing the effective ensemble size by the spatial aggregation of the predictand addresses the bias component of the sampling error effects. The spread of estimated values of the performance metrics, caused by the finite number of start dates, can be reduced by the spatial aggregation of this quantity. A further consideration concerns consistency between the forecast quantity and the metric. If the forecast quantities are spatially aggregated before being issued to end users, then the estimate of the performance metrics should include such a spatial aggregation to ensure they reflect the skill of issued predictands. In the following, spatial aggregation has been applied to the assessment metric to illustrate its potential to reduce the spread of estimates. The practical choice depends upon the balance between ensemble members and start dates in a hindcast set and the resulting bias and spread of estimates of performance metrics.

Figure 4 shows the effects of sampling errors upon the lower tercile probability after a Gaussian distribution has been fitted and spatial aggregation over nine grid boxes applied. The magnitude of the uncertainty is reduced by about 40% (50% for middle tercile forecasts, not shown). This improvement is equivalent to the reductions achieved by increasing sample sizes in a hindcast set by a factor of 3 or 4.

Figure 5 shows a significant reduction in the spread of sampling errors of BSS values due to spatial averaging over nine grid boxes. Consistent with Fig. 3, the results suggest little dependence between sampling errors and forecast error for this metric: the width of the 90% PI in the extratropics is similar to the tropics despite the better BSS scores in the latter. The test that produced the results in Fig. 2 was rerun with spatial averaging and it was found that 55% of all points in the global domain have BSS values significantly different from zero at the 5% significance level, in marked contrast to the 10% obtained when no processing is applied.
Figure 3 shows the impact of spatial averaging on the sampling errors of IQ estimates. The magnitude of the uncertainty has been reduced by 35% by spatial averaging over 9 grid boxes, equivalent to an increase in the sample size by a factor of approximately 2.4.

Figures 3 and 5 show that the sampling errors for BSS and IQ have a bias relative to the values obtained with no sampling errors. The bias is much smaller than the spread in BSS values for unprocessed data. However, the estimates of the IQ score have a significant bias relative to the spread in the Northern Hemisphere and the spatial aggregation of both the forecast quantity and the metric may minimize the total error.

These results suggest that spatial aggregation permits more robust information to be extracted from a forecast system. The spatial aggregation method can reduce the uncertainty in the forecast quantities and assessment metrics by an amount equivalent to increasing the size of the hindcast set by a factor of 3 or 4. This benefit is gained at the cost of reducing the spatial resolution of forecasts. Forecast quantities have a spatial scale much larger than the grid box size; therefore, this cost is relatively small.

c. Temporal aggregation of data

Sample sizes can also be boosted by the aggregation of forecast information from different start dates. The decision on whether to apply temporal aggregation is based on the comparison of the reduction in error due to increasing the sample size and the degradation caused by merging data that differ for reasons other than sampling error. The forecasts of a seasonal-mean quantity could be anticipated to have different signals on monthly time scales and increases in sample sizes are offset by contaminating the forecast with data containing different signals. The new seasonal forecasting system being developed at the Met Office, GloSea4, will be run weekly and temporal aggregation of forecasts is an attractive option. The processes responsible for prediction skill on seasonal time scales are expected to exhibit small changes in behavior on short time scales in most regions of the world. In such cases, the benefits from the aggregation of performance metrics from different start dates will outweigh the cost of the reduced temporal resolution.

In the following, a method of estimating the potential reduction in uncertainties due to temporal aggregation of monthly BSS values is described, and results from this method are shown.

The BSS values in the months immediately before and after the start month being examined are estimated in the same fashion as before. The Gaussian parameters are estimated using GloSea3 data for the appropriate start month, and the numerical experiment generates many sets of hindcast data for each of the start months being studied. The estimate of the BSS value for a particular start month is defined as the average of the BSS values calculated for this start month and the months immediately before and after it. It is worth noting that there is a temporal correlation between the values of the forecast quantities in the three different start months in the same hindcast year, which is not accounted for by this method. The effect of this temporal correlation is to reduce the effective sampling of start dates; therefore, the following results represent an upper limit on the reduction of the effect of sampling errors on BSS values by the temporal aggregation of three different start months in a hindcast set.

Figure 5 presents the impact due to the temporal aggregation of BSS values from one month before and after the start date on the BSS uncertainties. The 90% PI is reduced in both regions because of temporal averaging. The assumption of independent forecasts for each start month ensures that the reduction in spread is equivalent to an increase in sample size by a factor of 3.

Finally, if all three postprocessing methods are considered, BSS uncertainties can be reduced considerably. Figure 6 is similar to Fig. 2a and shows the regions with BSS values significantly different from zero at the 5% significance level for the GloSea3 hindcast set processed using all three methods aimed at reducing uncertainties, namely, a fitted Gaussian theoretical distribution with spatial and temporal averaging.
members, and no postprocessing, and the second system is based on a development version of the new system, GloSea4, with a 15-yr hindcast set, six ensemble members, and postprocessing methods described earlier. The results in Fig. 7 indicate that the GloSea4-like system has smaller effects from sampling errors than GloSea3 despite having a smaller sample size.

5. Summary and future work

A numerical test was designed to measure the effects of sampling errors on seasonal forecasts. Results from the test reveal that sample sizes in state-of-the-art operational seasonal forecasting systems produce significant uncertainty in typical forecast information issued to end users, with modern systems unable to distinguish the skill of the system from a climatological forecast in most regions of the world. The standard solution of increasing sample sizes is of little benefit because of resource constraints and the limitations imposed by a nonstationary observing system.

The impact of standard postprocessing methods on sampling errors was assessed using the numerical experiment. Employing appropriate statistical distributions with parameters specified by the data sample can reduce the effect of sampling errors on forecast quantities by an amount equivalent to an increase in sample sizes by 20%–40%. The computational cost of this procedure can be considered negligible, while its cost in terms of reduced accuracy is small since common forecast quantities have distributions that agree well with theoretical distributions. The fitting of theoretical distributions to limited data samples yields little improvement to the BSS performance metric, although it can significantly reduce the bias of the IQ score.

Spatial aggregation leads to large reductions in the adverse effects of sampling errors on forecast quantities and assessment metrics. Spatial aggregation reduces the spatial resolution of the forecast information with little adverse effects because seasonal anomalies have spatial scales of approximately zonal wavenumber 10 and forecast models can typically resolve wavenumber 100. The benefits of spatial aggregation over 3 by 3 grid-box regions were found to be equivalent to an increase in the sample sizes by a factor of about 3.

The temporal aggregation of forecasts from three consecutive start months centered on the start month of interest was investigated. The cost of this method in terms of reduced accuracy depends upon the temporal variability of the information being aggregated. The annual cycle of BSS values are likely to have sufficiently small amplitudes since the processes responsible for skill have longer than monthly time scales. The temporal aggregation of BSS values reduced the effects of sampling errors by an amount equivalent to increasing sample sizes by a factor of 3.

In summary, these postprocessing methods result in a reduction of sampling errors by a factor of about 4 for forecasts of near-surface air temperatures and by a factor of about 10 for associated assessment metrics. More robust forecast information can be produced in a GloSea4-like system than the current operational system. This is noteworthy since GloSea4 has a smaller hindcast set than GloSea3 in order to reduce computation and data storage requirements and allow a faster development cycle.

Some areas of future work are the following: to apply the methods to more extreme quantiles, to assess the use of digital filters for the removal of high-frequency spatial variability rather than spatial averaging, and to extend the numerical experiment presented here to analyze resource allocation between the numbers of start dates and ensemble members in seasonal forecasting operational systems.

Acknowledgments. This work was supported by the Joint Defra and MoD Programme (Defra) GA01101 (MoD) CBC/2B/0417_Annex C5. The authors wish to thank the anonymous reviewers for their improvements to the final manuscript.

REFERENCES


