Reliability of decadal predictions

S. Corti,[1,2] A. Weisheimer,[1,3] T. N. Palmer,[1,3] F. J. Doblas-Reyes,[4,5] and L. Magnusson[1]

Received 31 July 2012; revised 6 October 2012; accepted 8 October 2012; published 15 November 2012.

[1] The reliability of multi-year predictions of climate is assessed using probabilistic Attributes Diagrams for near-surface air temperature and sea surface temperature, based on 54 member ensembles of initialised decadal hindcasts using the ECMWF coupled model. It is shown that the reliability from the ensemble system is good over global land areas, Europe and Africa and for the North Atlantic, Indian Ocean and, to a lesser extent, North Pacific basins for lead times up to 6–9 years. North Atlantic SSTs are reliably predicted even when the climate trend is removed, consistent with the known predictability for this region. By contrast, reliability in the Indian Ocean, where external forcing accounts for most of the variability, deteriorates severely after detrending. More conventional measures of forecast quality, such as the anomaly correlation coefficient (ACC) of the ensemble mean, are also considered, showing that the ensemble has significant skill in predicting multi-annual temperature averages.
Citation: Corti, S., A. Weisheimer, T. N. Palmer, F. J. Doblas-Reyes, and L. Magnusson (2012), Reliability of decadal predictions, *Geophys. Res. Lett.*, *39*, L21712, doi:10.1029/2012GL053354.

1. Introduction

[2] There is a considerable upsurge of interest in the decadal prediction of climate, considered as a combined initial and boundary-value problem [*Goddard et al.*, 2012]. Indeed, for the first time, the potential for such prediction is being assessed by the Intergovernmental Panel on Climate Change (in its Fifth Assessment Report). Such predictions must be probabilistic, arising from inevitable uncertainties, firstly in knowledge of the initial state, secondly in the computational representation of the underlying equations of motion, and thirdly in the so-called "forcing" terms, which include not only greenhouse gas concentrations, but also volcanic and other aerosols [*Hawkins and Sutton*, 2009].

[3] If such predictions could be shown to be reliable when predicting non-climatological probabilities, there can be little doubt about their utility across a range of application sectors. However, has the science of decadal forecasting advanced to the stage that potential users could indeed rely on specific multi-year predictions?

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK.
[2]Istituto di Scienze dell'Atmosfera e del Clima, Consiglio Nazionale delle Ricerche, Bologna, Italy.
[3]National Centre for Atmospheric Science, Department of Physics, Atmospheric, Oceanic and Planetary Physics, Oxford University, Oxford, UK.
[4]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.
[5]Institut Català de Ciències del Clima, Barcelona, Spain.

Corresponding author: S. Corti, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK. (susanna.corti@ecmwf.int)

0094-8276/12/2012GL053354

[4] The word "reliable" has a specific technical meaning in probability forecasting, a meaning that can allow potential users to assess whether decadal forecasts might have value. Suppose a decadal forecast probability of some event E – say that that temperature lies above the long-term climatological median value - is equal to 0.7. For a reliable forecast system, one could assert that E would actually occur on 70% of occasions where E was forecast with a probability of 0.7.

[5] But are decadal forecasts reliable in this way? We address this question using Attributes Diagrams [*Hsu and Murphy*, 1986; *Mason*, 2004] (which provide a graphical display of information, not only for forecast probabilities of 0.7, but for the entire range of forecast probabilities). These Attributes Diagrams are commonly used to assess the reliability of medium range and seasonal forecasts [*Palmer et al.*, 2008], but have not, as yet, been used to assess the reliability of decadal prediction systems.

[6] The decadal prediction system under study here is based on four different versions of the ECMWF coupled model. We assess the reliability and the prediction skill of such a system in simulating the variability of surface and near-surface temperature over multi-year time scales for six selected geographical regions.

2. Experimental Design and Methodology

2.1. Experimental Setup

[7] Ten sets of decadal climate hindcasts were carried out with the ECMWF coupled system to represent the key uncertainties which give rise to forecast error in near-term climate predictions, such as uncertainties in the initial conditions, in model formulation and in future (and past) radiative forcing. Uncertainty in initial conditions is taken into account by perturbing randomly the ocean initial conditions and using four re-analysis data sets with two different ocean initialisation techniques such as full initialisation and anomaly initialisation (L. Magnusson et al., Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors, submitted to *Climate Dynamics*, 2012). To represent model uncertainty we use four different versions of the ECMWF coupled model and apply stochastic perturbations to the physical tendencies in the atmospheric component of the coupled system [*Palmer et al.*, 2009]. Finally, we consider the uncertainty arising from our limited knowledge of (some components of) radiative forcing by including (or not) tropospheric and/or stratospheric aerosols, associated with the effects of anthropogenic pollution and volcanic eruptions.

[8] Each experiment includes at least 3 and as many as 7 ensemble members generated by slightly different initial conditions. The simulations are 10 years long and were started on the 1st of November, once every five years over the period 1960 to 2005 [*Taylor et al.*, 2012]. The total 54 ensemble members from ten hindcast experiments

provide the values for the ECMWF "Multi-Model" Ensemble (EC_MME). A summary of each experimental configuration is given in the auxiliary material (see Text S1 and Table S1).[1]

2.2. Data and Computation of the Anomalies

[9] Near-surface air temperature and sea surface temperature data from ERA Interim [*Dee et al.*, 2011] and, prior to 1979, from ERA40 [*Uppala et al.*, 2005] are used to evaluate the hindcasts. The reference climatology is taken to be the average of the entire period from 1960 to 2010. To take into account the model systematic error, forecast anomalies for each experiment are calculated as: $X'_{j\tau} = X_{j\tau} - \bar{X}_{k\tau}$ where j is the starting year ($j = 1, n$), τ is the forecast month ($\tau = 1{,}120$) and $\bar{X}_{k\tau} = \frac{1}{n-1}\sum_{j\neq k}^{n-1} X_{j\tau}$ is the forecast average estimated in cross-validation mode (i.e., estimated removing the model climate for the specific forecast period, see *Doblas-Reyes et al.* [2011] for more details).

[10] Observation anomalies can be estimated either using the standard definition $O'_{j\tau} = O_{j\tau} - \bar{O}$ or basing the average on the period for which both observations and hindcasts are available, i.e., as: $O'_{j\tau} = O_{j\tau} - \bar{O}_{\tau}$. We applied the first definition, however the results do not change significantly if a *per-pairs* selection of the years is considered [*García-Serrano and Doblas-Reyes*, 2012].

[11] Skill in these decadal hindcasts comes mainly from two sources: radiative forcing and the predictive component of natural climate variability. The main component of radiative forcing consists in the rising trend of well-mixed greenhouse gases [*Intergovernmental Panel on Climate Change*, 2007]. The low-frequency climate variability can be (in principle) captured by the initialisation of the ocean [*Keenlyside et al.*, 2008; *Meehl et al.*, 2009; *Pohlmann et al.*, 2009]. A comparison between initialised and non-initialised hindcasts is the best way to assess the relative importance of initial conditions with respect to forcing. However, when, as here, companion non-initialised experiments are not available, linear detrending of anomalies, the technique used here, is a good approximation to filter out the effect of greenhouse gas warming [*van Oldenborgh et al.*, 2012; *Guemas et al.*, 2012a].

3. Forecast Quality Assessment

3.1. Measures of Skill

[12] The methods for evaluating ensemble prediction skill include deterministic and probabilistic measures. Deterministic measures of skill compare the ensemble-mean prediction anomaly against corresponding observations. A conventional metric used to measure the skill of ensemble forecasts is the anomaly correlation (AC) of the ensemble mean. Here we score the average anomaly for the lead times of 2–5 year and 6–9 year of surface temperature and near-surface air temperature.

[13] A shortcoming of deterministic measures of skill is that information about prediction uncertainties is not available. A different set of skill measures based on metrics for categorical probabilistic predictions can be used for this purpose. In this study we use the Brier skill score (BSS) [*Wilks*, 1995; *Mason*, 2004], which measures the improvement of the forecast relative to a reference forecast (in our case the sample climatology). A positive value of BSS indicates a forecast that is better than climatology. The BSS can be decomposed as follows [*Murphy*, 1973]: $BSS = BSS_{rel} + BSS_{res} - 1$, where BSS_{rel} and BSS_{res} are the reliability and the resolution components of the score. The reliability measures how close the forecast probabilities are to the observed frequencies. The resolution measures how much the forecast probabilities differ from the climatological probability of the event. By definition when the forecast is always the climatological probability, the system is perfectly reliable ($BSS_{rel} = 1$), but has no resolution ($BSS_{res} = 0$) and thus no skill ($BSS = 0$). The relative role played by reliability and resolution in determining the Brier Skill Score will be illustrated in the next section.

3.2. Reliability and Attributes Diagrams

[14] Decadal climate hindcasts (and forecasts) belong to the category of probabilistic predictions. As such they are issued in ensemble mode and ultimately they have to be evaluated on the basis of whether they give an accurate estimation of the relative frequency of the predicted outcome [*Murphy*, 1993]. When this happens they are considered "reliable". Reliability is indeed a necessary pre-condition for issuing useful probabilistic predictions. A forecast system may have virtually no skill (compared for example to climatology) and still be useful, as long as the forecasts are statistically reliable, and hence can be trusted when predicting non-climatological probabilities.

[15] Here we assess the reliability and the skill of EC_MME decadal hindcasts for sea surface temperature and near surface air temperature over different oceanic and land regions. For these variables we consider the following binary event $E(x)$: anomaly above the median at a particular point x. (It would be more useful to define events using tercile thresholds, but the sample space of hindcasts is too small to allow this. Therefore using terciles would further increase the uncertainty of results. Some examples with terciles are given in the auxiliary material).

[16] Attributes Diagrams [*Hsu and Murphy*, 1986; *Palmer et al.*, 2008] are used to illustrate the EC_MME decadal hindcast reliability. They measure how closely the forecast probabilities of an event correspond to the actual chance of observing the event. They are based on a discrete binning of many forecast probabilities taken over a given geographical region $\langle x \rangle$. In section 4 below, Figures 2 and 3, we will illustrate such Attributes Diagrams for six selected regions.

[17] For perfect reliability (i.e., $BSS_{rel} = 1$) the forecast probability and the frequency of occurrence should be equal, and the plotted points should, within their uncertainty ranges, lie on the diagonal (solid black line in the figures). When the line joining the bullets (the reliability curve) has positive slope, it indicates that as the forecast probability of the event occurring increases, so too does the verified chance of observing the event. The forecasts therefore have some reliability. However, if the slope of the reliability curve is flatter than the diagonal, then the forecast system is overconfident. If the reliability curve is mainly horizontal, then the frequency of occurrence of the event does not depend on the forecast probabilities. In this situation a user might make some very poor decisions based on such uncalibrated probabilities. The black dashed line separates skilful and unskillful regions in the diagram: points with forecast probability smaller (larger) than the climatological frequency which fall

[1] Auxiliary materials are available in the HTML. doi:10.1029/2012GL053354.

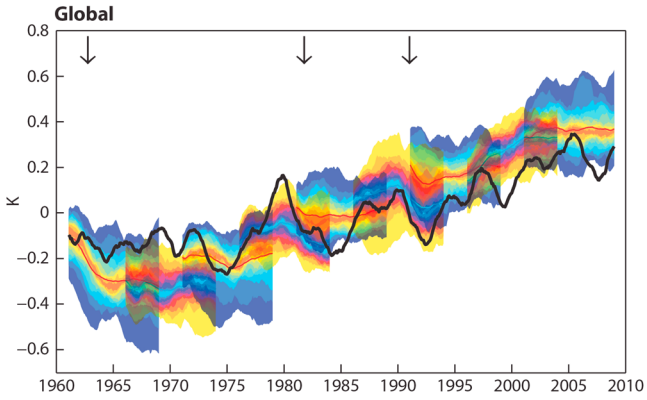

Figure 1. Time series of globally averaged near-surface air temperature anomaly over land smoothed with a 25-month running mean for reanalysis (black) and the ensemble distribution of EC_MME decadal hindcasts. The coloured shades represent the percentiles (from 5% to 95% every 5%) of the probability distribution associated with the spread of the 54 ensemble members composing the EC_MME. The colour scale for the shading alternates every starting date. Vertical black arrows on the top indicate the Agung, El Chichon and Pinatubo volcanic eruptions.

below (above) the dashed line contribute to positive BSS (i.e., $BSS_{res} > 1 - BSS_{rel}$); otherwise they contribute negatively to BSS. Bins associated with non-climatological forecast probabilities, particularly those with abscissa close to 0 and 1, define the "sharpness" of the forecast. If these bins are populated by the ensemble, then this is an indication of the ability of the system to produce a non-climatological forecast probability. However, the reliability of these bins centred at the tails of the climatological distribution depends on how close they are to the diagonal. An ideal forecast should have a high resolution (i.e., it should successfully distinguish cases in which the probability of an event is high from those in which the probability is low) whilst retaining reliability. In other words it should be both sharp and reliable.

4. Results

[18] Figure 1 shows the evolution of the low-pass filtered global air surface temperature anomaly over land from the observation and the ensemble mean. (Similar time series for air surface temperatures over Europe and Africa are shown in Figures S1 and S2, while sea surface temperatures over North Atlantic, North Pacific and Tropical Indian Ocean can be found in Figures S3–S5.) The shaded area represents the spread of the distribution associated with the 54 ensemble members. Overall, the ECMWF decadal hindcasts follow the evolution of the reanalysis curve, however it is worth noticing that the average amplitude of the ensemble spread is comparable to the inter-decadal variability. Most of the simulations underestimate the temperature anomaly in the early decades compared to the reanalysis, and tend to overestimate the warming from 1990 onwards. It appears that this is a common problem to all models from CMIP5 decadal hindcast experiment [*Kim et al.*, 2012].

[19] The resulting maps from the computation of the anomaly correlation coefficients for the lead times of 2–5 year and 6–9 year of surface air temperature (over land) and sea surface temperature (over sea) are shown in Figure S6. The EC_MME skill is high at the 95% confident level over large regions up to 6–9 years. North Atlantic, Indian Ocean and Subtropical Eastern Pacific are the oceanic areas with the highest scores. This result compares well with other model predictions [see, e.g., *van Oldenborgh et al.*, 2012; *Kim et al.*, 2012]. A more detailed discussion of the results from the anomaly correlation can be found in the auxiliary material (see Text S2 and Figure S6).

[20] Figures 2 and 3 show Attributes Diagrams for $E(x)_{T2m}$ for three selected regions over land (Global, Europe and Africa) and for $E(x)_{SST}$ for three ocean basins (North Pacific, North Atlantic and Tropical Indian Ocean) respectively. There are three diagrams for each region: at 2–5 and 6–9 lead times and at 2–5 year lead time for detrended data (lead time 6–9 year is shown in Figure S7). Overall, in all the geographical regions considered, with the exception of the North Pacific Ocean, the EC_MME predictions appear to be reliable when the climate trend is not filtered out. The Tropical Indian Ocean stands out as the region with the highest significant BSS (0.52) at both lead times. (BSS, BSS_{rel}, and BSS_{res} are indicated on the top left corner of each diagram). This particularly high score is related to the resolution of the forecast, i.e., high populations and high reliability of the non-climatological bins. Significant positive BSS are found for $E(x)_{T2m}$ over Europe and Africa, while the scores over Global land, even if positive, are not significant (at 95% of confidence estimated with bootstrap resampling procedure). All the BSSs except the one for the North Pacific Ocean, decrease after detrending. In the light of results shown in Figure S6, this is not unexpected: a non-negligible portion of the signal disappears when an estimate of the impact of the natural and anthropogenic climate forcing is filtered out. However, Attributes Diagrams convey information about the reliability of a forecast as well (and this information cannot be easily extracted from anomaly correlation maps). All the diagrams for detrended anomalies, with the (important) exception of the Tropical Indian Ocean, appear substantially reliable. The sample population is in general more concentrated on the climatological bins (i.e., BSS_{res} decreases with detrending), but the bullets lie close to the diagonal (i.e., BSS_{rel} values are relatively stable), indicating that the hindcasts are reliable.

[21] The results over the North Atlantic (Figures 3a–3c) deserve special attention since this is a region where the influence of ocean initial conditions in decadal predictions is detectable [*Pohlmann et al.*, 2009; *Branstator and Teng*, 2012]. The Attributes Diagrams for the North Atlantic show high reliability (the highest amongst the regions considered here) even when the climate trend is filtered out. The BSS scores are positive for all lead times. When detrended anomalies are considered (Figures 3c and S7b), the BSS decreases slightly for lead time 2–5 year (from 0.2 to 0.14) and more substantially for longer lead times (from 0.27 to 0.05). The modest decrease in skill for shorter lead times is in agreement with the recent results of *Branstator and Teng* [2012]. These authors found that the impact of initial conditions over the North Atlantic becomes secondary (with respect to the forcing) after 8 years. (Indeed in Figures S8a and S8b it is shown that BSS at 5–8 year lead time is still of 0.12 after detrending. At longer lead time some skill is found over the Tropical Atlantic, see Figures S9c and S9d).

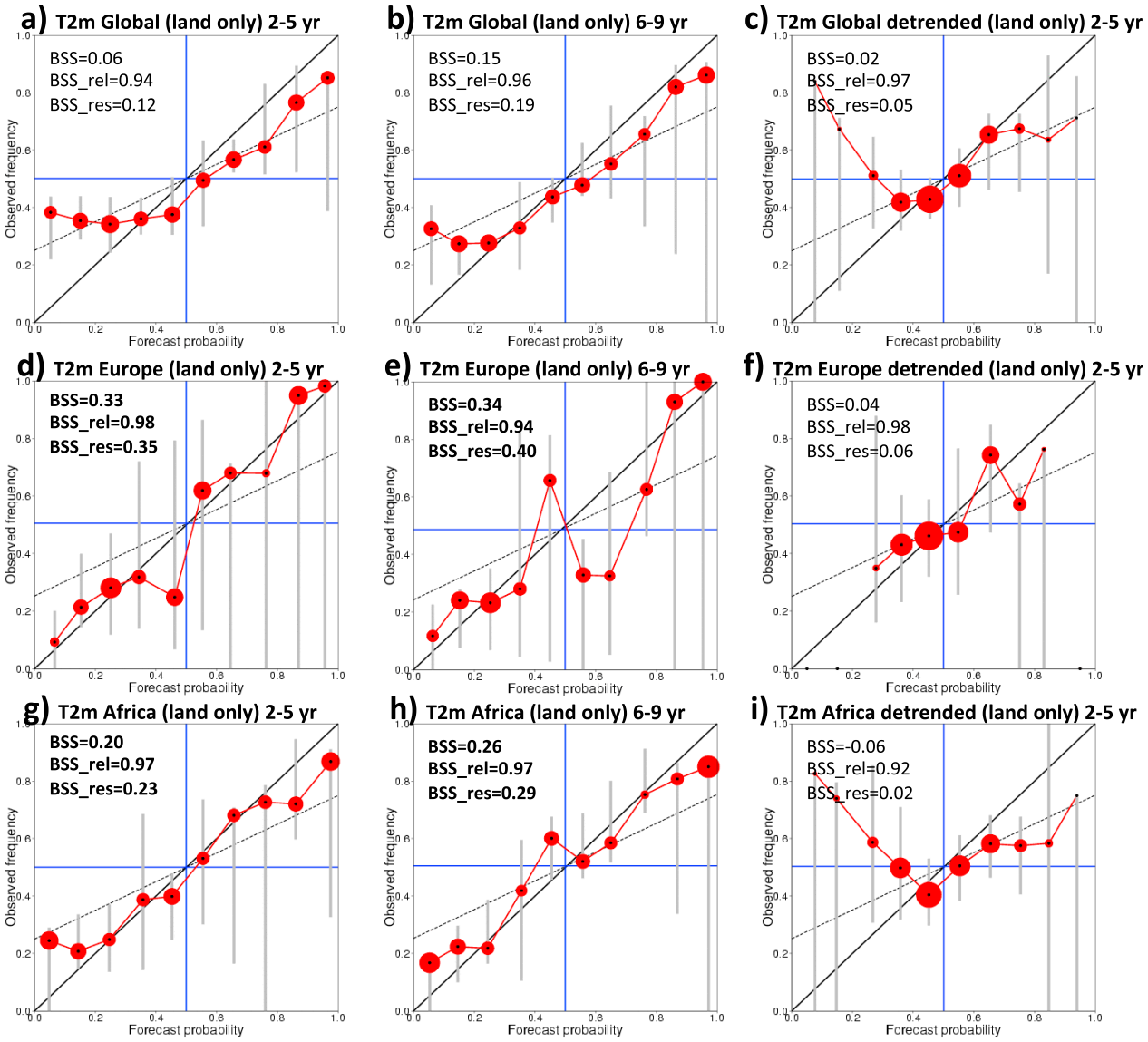

Figure 2. Attributes Diagrams for ECMWF multimodel decadal hindcasts for $E(x)_{T2m}$ above the median for selected land regions, lead times and de-trended. The size of the bullets represents the number of forecasts in each probability bin (sharpness). The blue horizontal and vertical lines indicate the climatological frequency of the event in observation and forecast, respectively. Grey vertical bars indicate the uncertainty in the observed frequency for each probability bin estimated at 95% level of confidence with a bootstrap resampling procedure based on 1000 samples. The Brier Skill Score and its components (reliability and resolution) are indicated on the top left corner of each diagram (in bold when BSS is not significantly negative at 95% level of confidence). The black dashed line separates skillful from unskillful regions (see text for further details).

[22] The results from the Attributes Diagrams are encouraging. However the uncertainty associated with them is non negligible. This is mainly due to the exiguous number of hindcasts. In fact, in some of the cases shown here (see for example Figures 2e and 3h), the uncertainty associated with each observed frequency (the grey bars in Figures 2, 3, S7–S9) can encompass all the range of possible frequencies. Because of this, it was decided to focus our analysis on the frequent events $E(x)_{SST}$ or $E(x)_{T2M}$ above the median. As an example in Figures S10 (Figure S11) the Attributes Diagrams for North Atlantic SSTs for the events $E(x)_{SST}$ > upper tercile ($E(x)_{SST}$ < lower tercile) for lead times 2–5 and 6–9 year and for non-detrended and detrended data are shown. The results are comparable with our findings for the "above the median" event (it is interesting to note that there is an asymmetry in the BSS between the two terciles: warm events are better predicted), however the uncertainty associated with the terciles is larger. To reduce this uncertainty it will be necessary to repeat this exercise with a larger sample of hindcasts.

5. Conclusions

[23] Currently there is considerable interest in Decadal Prediction ("the fascinating baby that all wish to talk about" [*Goddard et al.*, 2012]) and initialised decadal predictions

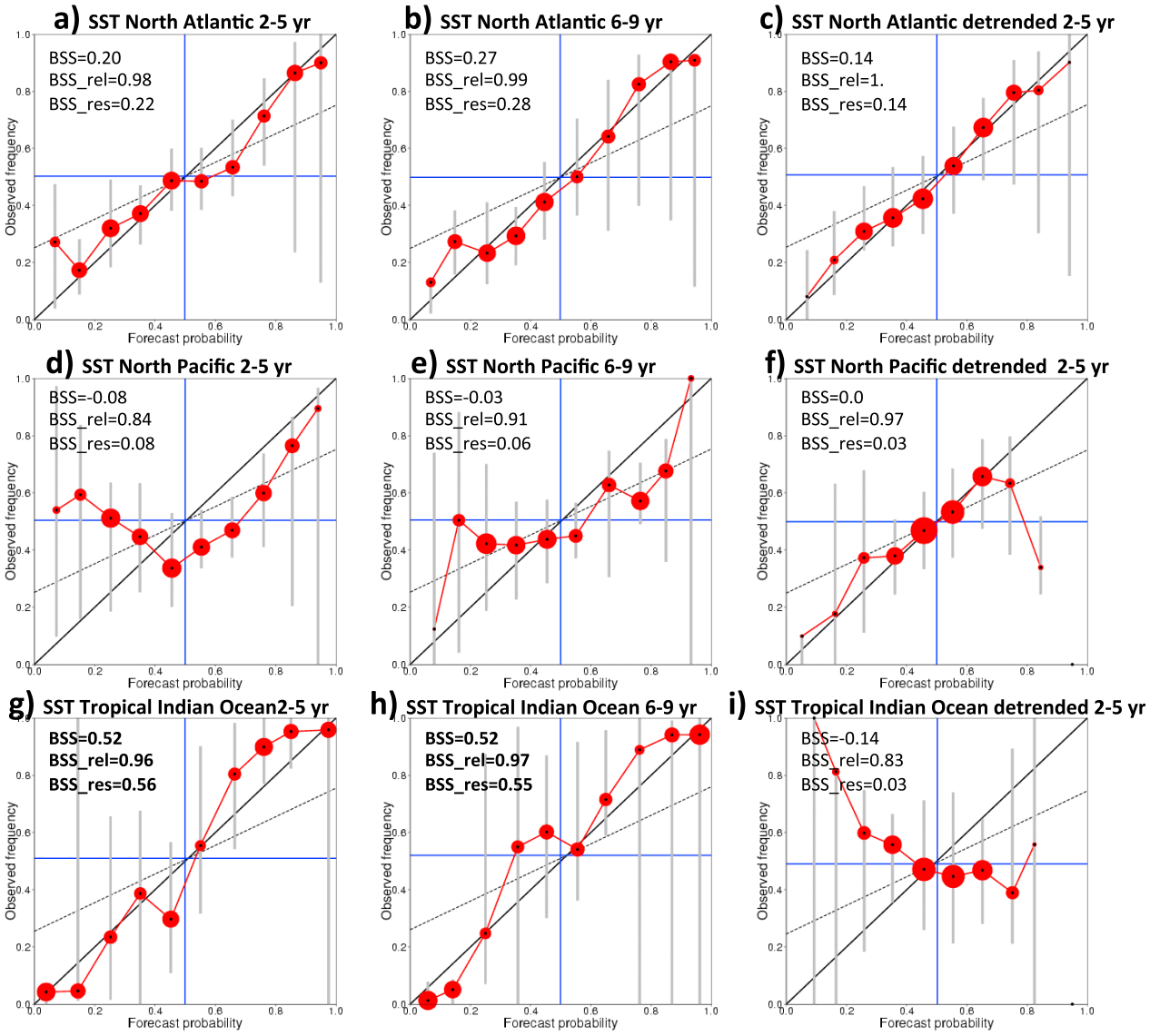

Figure 3. As Figure 2 for $E(x)_{SST}$ above the median for selected ocean basins indicated in the subpanel titles.

are being assessed in the forthcoming IPCC Fifth Assessment Report. Estimates of skill from the CMIP5 decadal hindcasts suggest modest but positive skill. However, are such hindcasts reliable in the probabilistic sense? It is argued here that the usefulness of decadal (indeed all) forecasts can be most readily assessed from Attributes Diagrams which illustrate graphically the reliability and sharpness of the probability forecasts. Above all, a potential user can assess from an Attributes Diagram whether forecasts, particularly where the forecast probabilities differ from climatology, are reliable.

[24] Here Attributes Diagrams for temperature have been shown for a "multi-model" ensemble based on the ECMWF coupled model system. Conventional skill scores based on Anomaly Correlation Coefficients for the ensemble mean are comparable with values from the CMIP5 ensemble mean hindcasts. Attributes Diagrams show that the reliability from the ECMWF ensemble system is remarkably good for most regions studied. Over large continental areas as Europe and Africa the prediction skill decreases when the climate trend is filtered out, but reliability is preserved. Over the North Atlantic the hindcasts are both sharp and reliable, even after detrending, confirming the importance of initialisation in that region. On the other hand, in the Indian Ocean, the region where the ratio of internally-generated to the externally-forced variability is the lowest, predictions are reliable only for the trend component. Decadal predictions are less reliable (and skilful) over the North Pacific where very strong natural variations are observed. These points are consistent with results found in a number of decadal prediction studies [e.g., *Kim et al.*, 2012; *Guemas et al.*, 2012b]. However, the results have to be tempered by the fact that the sample size is small and hence error bars in the Attributes Diagram are relatively large. We suggest that a larger sample of hindcasts should be generated by groups studying the decadal prediction problem, extending the protocol recommended by CMIP5 [*Taylor et al.*, 2012].

[25] It should be stressed that these results are for sea surface and near-surface temperature only and it is not

expected that decadal precipitation forecasts will be reliable in the same way as temperature is.

[26] **Acknowledgments.** This work was supported by the EU-funded projects THOR (FP7/2007–2013) under grant agreement 212643, COMBINE (FP7/2007–2013) under grant agreement 226520 and the MICINN-funded RUCSS (CGL2010-20657) project.

[27] The Editor thanks the two anonymous reviewers for their assistance in evaluating this paper.

References

Branstator, G., and H. Teng (2012), Potential impact of initialization on decadal predictions as assessed for CMIP5, *Geophys. Res. Lett.*, doi:10.1029/2012GL051974, in press.

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, *137*, 553–597, doi:10.1002/qj.828.

Doblas-Reyes, F. J., M. A. Balmaseda, A. Weisheimer, and T. N. Palmer (2011), Decadal climate prediction with the European Centre for Medium-Range Weather Forecasts coupled forecast system: Impact of ocean observations, *J. Geophys. Res.*, *116*, D19111, doi:10.1029/2010JD015394.

García-Serrano, J., and F. J. Doblas-Reyes (2012), On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast, *Clim. Dyn.*, *39*, 2025–2040, doi:10.1007/s00382-012-1413-1.

Goddard, L., J. W. Hurrell, B. P. Kirtman, J. Murphy, T. Stockdale, and C. Vera (2012), Two time scales for the price of one (almost), *Bull. Am. Meteorol. Soc.*, *93*, 621–629, doi:10.1175/BAMS-D-11-00220.1.

Guemas, V., S. Corti, J. Garcìa-Serrano, F. Doblas-Reyes, M. Balmaseda, and L. Magnusson (2012a), The Indian Ocean: The region of highest skill worldwide in decadal climate prediction, *J. Clim.*, doi:10.1175/JCLI-D-12-00049.1, in press.

Guemas, V., F. Doblas-Reyes, F. Lienert, Y. Soufflet, and H. Du (2012b), Identifying the causes of the poor decadal climate prediction skill over the North Pacific, *J. Geophys. Res.*, doi:10.1029/2012JD018004, in press.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*, 1095–1107, doi:10.1175/2009BAMS2607.1.

Hsu, W.-R., and A. H. Murphy (1986), The attributes diagram: A geometrical framework for assessing the quality of probability forecasts, *Int. J. Forecast.*, *2*, 285–293, doi:10.1016/0169-2070(86)90048-8.

Intergovernmental Panel on Climate Change (2007), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, Cambridge, U. K.

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

Kim, H.-M., P. J. Webster, and J. A. Curry (2012), Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, *Geophys. Res. Lett.*, *39*, L10701, doi:10.1029/2012GL051644.

Mason, S. J. (2004), On using "climatology" as a reference strategy in the Brier and ranked probability skill scores, *Mon. Weather Rev.*, *132*, 1891–1895, doi:10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2.

Meehl, G. A., et al. (2009), Decadal prediction, *Bull. Am. Meteorol. Soc.*, *90*(10), 1467–1485, doi:10.1175/2009BAMS2778.1.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, A. H. (1993), What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather Forecast.*, *8*, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell (2008), Toward seamless prediction: calibration of climate change projections using seasonal forecasts, *Bull. Am. Meteorol. Soc.*, *89*, 459–470, doi:10.1175/BAMS-89-4-459.

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, J. Shutts, G. M. Steinheimer, and A. Weisheimer (2009), Stochastic parametrization and model uncertainty, *Tech. Memo. 598*, Eur. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Pohlmann, H., J. H. Jungclaus, A. Kohl, D. Stammer, and J. Marotzke (2009), Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic, *J. Clim.*, *22*, 3926–3938, doi:10.1175/2009JCLI2535.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Uppala, S. M., et al. (2005), The ERA-40 reanalysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012, doi:10.1256/qj.04.176.

van Oldenborgh, G. J., F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger (2012), Decadal prediction skill in a multi-model ensemble, *Clim. Dyn.*, *38*, 1263–1280, doi:10.1007/s00382-012-1313-4.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, San Diego, Calif.