Vol. 55: 181–200, 2013
doi: 10.3354/cr01135

CLIMATE RESEARCH
Clim Res

Published online January 10

OPEN ACCESS

Methodological aspects of the validation of decadal predictions

Reidun Gangstø[1,2,*], Andreas P. Weigel[1], Mark A. Liniger[1], Christof Appenzeller[1]

[1]**Federal Office of Meteorology and Climatology MeteoSwiss, 8044 Zurich, Switzerland**

[2]***Present address:*** **The Norwegian Meteorological Institute, 0313 Oslo, Norway**

ABSTRACT: Validation techniques based on past performance are well-established in seasonal forecasting. So far there is no consensus on the degree to which these are applicable to decadal predictions. We contribute to this discussion by assessing the effects of drift-correction, cross-validation and de-trending. The study employs decadal hindcasts of 2 m temperature from the EU FP6 ENSEMBLES project database and a synthetic toy model. Decadal predictions can be subject to substantial lead-time dependent model drifts. The conventional drift-correction method has a considerable sampling uncertainty, amounting to up to 40% of the potentially predictable signal. Introducing a smooth drift curve allows this uncertainty to be reduced by about 30% for annual values. For drift-corrected decadal predictions the leave-one-out cross-validation procedure may lead to biased skill estimates for decadal prediction due to the small number of hindcasts available. We identify this effect and show that 'jackknifing' represents a suitable technique for estimating potential skill without bias and to estimate sampling uncertainty. Results indicate significant correlation skill on the order of 0.7 to 0.9 for predicting global annual mean temperature on all lead-times. If linear trends are removed prior to verification, skill is still clearly above 0 in the first year for global mean temperature. On a local scale, some specific regions exhibit skill additional to the trend even at longer lead times. With the limited dataset analyzed here, the strong sampling uncertainty still prohibits drawing a final conclusion, by means of verification, on whether or not decadal predictions have skill in predicting climate variability beyond the trend.

KEY WORDS: Decadal predictions · Cross validation · Bias/drift correction · Jackknife · Seasonal forecasts · Sampling uncertainty · Detrending

1. INTRODUCTION

There is a growing demand for future climate information on different time scales ranging from a season to a century ahead, as evident for example during the Third World Climate Conference (WCC-3) in Geneva in 2009. With existing climate modeling and prediction capabilities, some of these demands have been addressed in recent years. For instance, seasonal predictions, which rely heavily on the prediction of the El Niño Southern Oscillation (ENSO), have become a well-established technique. They are now routinely issued by several weather and climate services, and have many applications in climate risk management. On the other hand, anthropogenic climate change towards the end of the century can be estimated by global climate models under prescribed scenarios of greenhouse gas and aerosol emissions and corresponding changes in radiative forcing (IPCC 2007). However, on the time-scale of a few years to a decade, i.e. the time-scale between seasonal forecasting and multi-decadal climate change projections, there is still a substantial gap in prediction capability (Meehl et al. 2009). This time-scale has high relevance for stakeholders and decision makers in many sectors, such as infrastructure planning, water resource management, energy production, insurance and agriculture (Vera et al. 2010).

In recent years, substantial efforts have been made to start filling this gap. The existence of modes of nat-

*Email: reidung@met.no

ural oceanic variability on decadal time-scales, such as the Pacific Decadal Oscillation (PDO) and the Atlantic Multidecadal Oscillation (AMO), triggered the hope that some of this variability could be predicted and the impacts on the atmosphere be estimated by initializing coupled ocean-atmosphere global circulation models (AOGCMs) with ocean data, especially the upper-ocean heat content. Since then, a number of pioneering modeling studies have been carried out, some of them with a focus on global near-surface temperature (Smith et al. 2007, Doblas-Reyes et al. 2011, Fyfe et al. 2011, Matei et al. 2012), others with a more regional focus such as the North Pacific (Mochizuki et al. 2010) and the North Atlantic (Keenlyside et al. 2008, Pohlmann et al. 2009), and a few focusing on variables other than temperature, such as Atlantic hurricane activity (Smith et al. 2010). In the context of the EU funded FP6 ENSEMBLES Project (van der Linden & Mitchell 2009, Doblas-Reyes et al. 2010) and, very recently, the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al. 2012), large data sets of decadal reforecasts (or hindcasts) have been performed in a coordinated effort with multiple AOGCMs. This allows for a multi-model assessment of decadal predictability (van Oldenborgh et al. 2012, Kim et al. 2012), so that more insight can be gained into the consequences of uncertainties resulting from model error (e.g. Hagedorn et al. 2005, Doblas-Reyes et al. 2005, Weigel et al. 2008a). These dynamical modeling approaches have been complemented by studies with simpler modeling approaches, including statistical modeling (e.g. Collins et al. 2006, Lean & Rind 2009, Krueger & Von Storch 2011, Dunstone et al. 2011).

Overall, the studies carried out up to now indicate that some aspects of ocean variability is predictable by initialized AOGCMs, but decadal prediction skill over land may be limited to a few specific regions (Smith et al. 2010). The field of decadal predictions is still in its infancy, and there are still plenty of technical and scientific issues to be resolved (Murphy et al. 2010) before one can arrive at a final judgment on the true prediction skill and value of decadal hindcasts for decision making. These open questions refer in particular to the underlying processes of decadal predictability, and to the optimum technique for initializing AOGCMs. But a further open question is how optimal strategies for the validation and post-processing of decadal forecasts can be developed. This question was the focus of this study.

For instance, at the moment there is no consensus on what is the best strategy to validate the prediction skill of decadal forecasts. In weather and seasonal forecasting, it has become common practice to assess prediction skill by computing suitable scores over a sufficiently large set of past forecasts, or hindcasts, comparing these to corresponding verifying observations (e.g. Weigel et al. 2007, Weigel 2011). Unfortunately, it is not straightforward to apply such an approach to decadal prediction systems because the number of independent verification samples provided by present-day modeling is very small, i.e. lower than 10 values for multiannual averages. The reason is that the long lead-times involved require that hindcasts are computed from initial conditions in pre-satellite times to obtain reasonable sample sizes. This is particularly difficult since sub-surface ocean data are needed for initializing decadal forecasts (Smith et al. 2007); however historical sub-surface ocean data are scarce and may be subject to bias (Domingues et al. 2008, Ishii & Kimoto 2009). Another unresolved issue is the question of how to deal with unpredictable explosive volcanic eruptions in decadal hindcasts (Meehl et al. 2009). Including them in the hindcasts would lead to an overestimation of true prediction skill, while excluding them may lead to an underestimation. But even if the hindcasts could be interpreted as representative samples of true predictive capability, and could be validated by a formal verification, the small sample sizes involved remain a key problem that, under certain circumstances, can lead to significant biases in skill estimates, as we will show in more detail in this study.

Another specific feature of decadal forecasts, at least if initialized with a realistic state of the climate system, is the development of a substantial model drift, i.e. systematic errors that have a tendency to build up with lead-time and may reach magnitudes of up to 10°C in some regions (Latif et al. 2010). Such model drift may have multiple causes, including inconsistencies between the radiation budgets used for the ocean and atmospheric reanalyses, and model errors due to missing processes (Doblas-Reyes et al. 2011), as well as insufficient model resolution (Scaife et al. 2011). Different approaches to reduce model drift are currently being explored, such as improving model physics (Doblas-Reyes et al. 2011), constraining model parameters on the basis of observations (Zhang 2011), and using anomaly initialization, i.e. initializing model runs with observed anomalies rather than observed values (e.g. Keenlyside et al. 2008, Pohlmann et al. 2009, Matei et al. 2012). However, as long as the drift problem is not fully resolved, an empirical *a posteriori* bias-correction derived from the hindcasts is the obvious option to provide unbiased forecasts. Such an empirical drift-correction is

common practice in the context of seasonal forecasts (Stockdale 1997), and has also been recommended and applied for decadal predictions (International CLIVAR Project Office 2011, Yeager et al. 2012). In this paper, we discuss the applicability of this approach to decadal forecasts and propose a simple but effective improvement to this procedure.

While most studies indicate that models have at least some skill in predicting 2 m temperature with decadal forecasts, it is not yet clear how much of this skill can be attributed to the forced climate change signal, and how much effectively results from forecast initialization. The obvious way to assess the skill resulting from initialization is the parallel use of initialized and non-initialized model runs for a given setting of radiative forcing (e.g. Smith et al. 2007). However, such accompanying model runs are often not available (as is the case in the ENSEMBLES hindcasts dataset, for example). A simple alternative strategy is to remove both the observed and modeled trends from the data and verify the residuals (van Oldenborgh et al. 2012), a strategy that has also been applied in the context of seasonal forecasting (Liniger et al. 2007). However, distinguishing between natural and externally forced variations is not easy (Solomon et al. 2011). In particular, it is not yet known how the external forcing interacts with natural modes of variability, and a separation might introduce noise and obscure the original signal so that the results obtained were not robust, particularly at local scale. It is clear that external forcing is likely not equivalent to a linear trend. We believe that it is nevertheless justified to use the linear trend as a first 'guess' reference benchmark, firstly as it is more robust than other approaches (recall the small sample size), and secondly because it already has been widely used, particularly by the end-user community, and thus provides a basis for comparisons. For instance, for stakeholders and decision makers, it may be reasonable to carry out such a trend separation approach when statements need to be made on the added value of initialized decadal predictions with respect to alternative simple model strategies, such as a linear trend extrapolation consistent with climate change scenarios. Van Oldenborgh et al. (2012) stated (but did not explicitly show) that the exact definition of trend has no effect on the results. In this study we pick up on this discussion, assess the sensitivity of the results with respect to the choice of de-trending method, and discuss the results in the context of the sampling uncertainty resulting from small sample size.

In summary, this study provides some comments and suggestions on methodological issues related to drift-correction, verification and trend separation of decadal forecasts. The analyses are primarily based on the decadal hindcasts of 2 m temperature from the EU FP6 ENSEMBLES project database, and additionally on analysis of data generated with a synthetic toy model.

2. DATA AND METHODS

2.1. Hindcast data and observations

Our analyses are based on the evaluation of decadal hindcasts of 2 m temperature that have been computed within the EU FP6 ENSEMBLES project (Van der Linden & Mitchell 2009, Doblas-Reyes et al. 2010). The ENSEMBLES multi-model comprises the following 4 coupled atmosphere-ocean prediction systems: ARPEGE4.6-NEMO (institution: Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique [CERFACS]), IFS33r1-HOPE/E (European Centre for Medium-Range Weather Forecasts [ECMWF]), ECHAM5-MPI/OM1 (Leibniz Institute of Marine Sciences at the University of Kiel [IFM-GEOMAR]) and HadGEM2-OA (UK Met Office [UKMO]). Three of the 4 models (CERFACS, ECMWF and UKMO) were employed with full initialization, that is, the oceans were initialized by an ocean analysis close to the true observations; whereas the IFM-GEOMAR model was used with anomaly initialization (Keenlyside et al. 2008), that is, the model was initialized by determining the anomaly of the observed state and adding it to the model climatology. The role of initialization has not been assessed in this study, and the fact that 3 out of 4 models employed full initialization is not meant to suggest any preference for this method. This question is currently subject to intensive research by CMIP5 (Taylor et al. 2012) and the German coordinated research program 'MiKlip' (www.fona-miklip.de/en/), but is only of secondary relevance for our paper since we mainly focus on methodological issues in the evaluation procedure. The models evaluated in this study did not consider any aerosol effects related to volcanic eruptions, apart from the ECHAM5 runs where volcanic aerosol concentrations from eruptions occurring before the analysis date were relaxed to zero with a time scale of 1 yr. For more information on the configuration and technical background of the ENSEMBLES decadal hindcasts, the reader is referred to Doblas-Reyes et al. (2010).

With each of these 4 models, decadal hindcasts were integrated in intervals of 5 yr, starting with

initial conditions on 1 November 1960, for the period 1960–2005. Each model was run with 3 ensemble members started from different initial conditions. Due to the short-term memory of the atmosphere, the first month of the integrations consistently has much higher skill than any other following month (a feature that it also known from seasonal forecasting). The first month was therefore removed from the analyses in order to focus on the skill resulting from the memory of the non-atmospheric components of the system such as the ocean. In order words, forecasts of 'Year 1' consider lead-times of 2 to 13 mo (i.e. the period December–November), forecasts of 'Year 2' consider lead-times of 14 to 25 mo, and so forth. Since the 10th hindcast year only comprises 11 mo due to this definition, it was excluded from the analyses.

As an observational reference for calibration and verification, 2 m temperature fields from the ERA-40 reanalysis (Uppala et al. 2005) were applied for the time period December 1959 until November 1989, and fields from the ERA-Interim analysis (Dee et al. 2011) were applied thereafter. Due to differences in terms of methodology and data used, the combined series of ERA-40 and ERA-Interim reveals an inhomogeneity in 1989 where the 2 datasets are merged. This is depicted in Fig. 1, which shows the time-series of annual global mean 2 m temperature for both datasets. The overlap period between 1990 and 2001 (grey shading) reveals that ERA-40 is systematically shifted with respect to ERA-Interim by about 0.15°C. Before merging the 2 datasets, the ERA-40 data were therefore calibrated against the ERA-Interim data by subtracting the mean difference during the overlap period at each grid-point separately. The resulting combined dataset (dot-dashed line in Fig. 1) is henceforth simply referred to as ERA. Both hindcasts and verifying observations are interpolated on a grid with 1° × 1° resolution, covering the latitudes from 87° S to 87° N. When calculating global averages, the individual grid-points were weighted according to their size.

We are aware that, depending on the region considered, datasets other than the ERA reanalysis may be more appropriate as observational reference of near-surface temperature. For instance, both van Oldenborgh et al. (2012) and Doblas-Reyes et al. (2011) propose a specific combination of the (NCEP) GHCN/CAMS, (NCEP) ERSST V3b and GISTEMP datasets. However, since the focus of this study is predominantly on methodological aspects rather than on the actual interpretation of decadal prediction skill, the choice of observational reference is only of secondary importance in the present context.

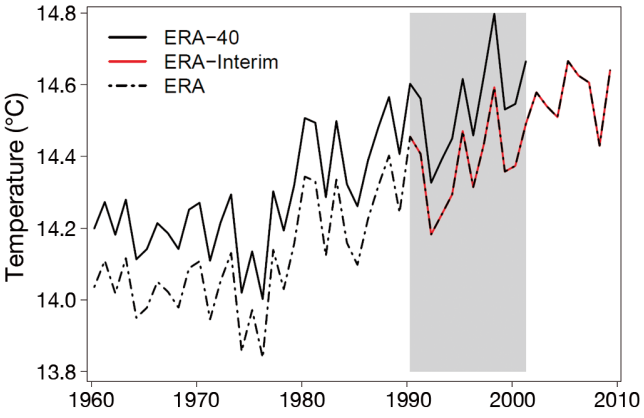

Fig. 1. Global annual mean temperature for the time period December 1959 to November 2010, based on ERA-40 (black solid line), ERA-Interim (red line), and the combined dataset that is here referred to as 'ERA' (black dot-dashed line). The grey shading indicates the overlap period that has been used to calibrate the ERA-40 data against ERA-Interim

2.2. Jackknife estimator

In this study, a statistical resampling technique called 'jackknife' is applied to obtain debiased estimates of the expectation and the variance of a parameter. In the following, a brief technical summary of this technique is provided. For a more detailed description and derivation of the jackknife estimator, the reader is referred to Efron & Gong (1983) and von Storch & Zwiers (1999).

Consider a set of n random samples $X_1, X_2,..., X_n$ that have been sampled from an unknown probability distribution F. Let $\theta(X)$ be a parameter of F, and let $\hat{\theta} = \hat{\theta}(X_1, X_2,..., X_n)$ be an estimator of θ as obtained from the n random samples. $\hat{\theta}$ could for example be the sample mean, the sample median, or the sample variance, while θ would then be the (usually unknown) population mean, population median, or population variance. $\hat{\theta}$ depends on the values of the n random samples. Consequently, $\hat{\theta}$ is subject to sampling variability and may vary between different sampling experiments. Moreover, $\hat{\theta}$ may be systematically biased with respect to $\theta(X)$, i.e. the expectation of $\hat{\theta}(X_1, X_2,..., X_n)$ may be different from $\theta(X)$.

Jackknifing is a relatively simple non-parametric approach that can be applied to estimate the systematic bias of an estimator $\hat{\theta}$, as well as its variance. The basic idea is that bias and variance of the estimator can be estimated by repeatedly recomputing the estimator leaving out a different sample value each time. More specifically: Let $\hat{\theta}_{(i)}$ be the estimator that is calculated using all n samples apart from X_i, and let $\hat{\theta}_{(\bullet)}$ be the average over all $\hat{\theta}_{(i)}$. That is,

$$\hat{\theta}_{(i)} = \hat{\theta}\left(X_1,...,X_{i-1},X_{i+1},...,X_n\right), \text{ and}$$

$$\hat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \quad (1)$$

Following Efron & Gong (1983), the jackknife estimate of the bias of $\hat{\theta}$, $B_{\hat{\theta}}$, is given by

$$B_{\hat{\theta}} = (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta}) \quad (2)$$

Subtracting $B_{\hat{\theta}}$ from $\hat{\theta}$ yields the bias corrected jackknife estimate $\hat{\theta}_J$:

$$\hat{\theta}_J = \hat{\theta} - (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta}) \quad (3)$$

The jackknife estimate of the variance of $\hat{\theta}$ is given by

$$\hat{\sigma}^2_{\hat{\theta}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)}\right)^2 \quad (4)$$

As an example, consider a 2-dimensional normal distribution $f(X,Y)$ with mean $(0,0)$, standard deviations $\sigma_X = \sigma_Y = 1$, and correlation coefficient $\mathrm{cor}(X,Y) = \rho$:

$$f(X,Y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}(X^2-2\rho XY+Y^2)} \quad (5)$$

Let there be n random samples from this distribution, (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n), and let $\hat{\rho}_n$ be the sample correlation between the vectors $(X_1,...,X_n)$ and $(Y_1,...,Y_n)$. The X-Y pairs could for example be interpreted as forecast-observation pairs generated with a toy climate model, $\hat{\rho}_n$ as an estimate of prediction skill on the basis of the n verification samples, and ρ as the true underlying model skill (e.g. Weigel et al. 2008a). Fig. 2 shows (as open circles) the expectation of $\hat{\rho}_n$ as a function of n for fixed $\rho = 0.5$. The values have been obtained by averaging over 100 000 resampling experiments from $f(X,Y)$. It can be seen that $\hat{\rho}_n$ is a slightly negatively biased estimator of ρ, with the magnitude of bias growing as sample size is decreased. The corresponding expected jackknife estimates are shown as filled triangles. They have been obtained by Eq. (3), setting $\hat{\theta} = \hat{\rho}_n$ and $\hat{\theta}_{(i)} = \hat{\rho}[(X_1,...X_{i-1},X_{i+1},...X_n),(Y_1,...Y_{i-1},Y_{i+1},...Y_n)]$. Along with the expected jackknife estimates, the plot shows their variability (± 1 SD), as diagnosed from the 100 000 experiments (grey shading), and as estimated from the jackknife variance estimator of Eq. (4) (error bars). The plot illustrates that (1) the jackknife mean estimates provide almost unbiased estimates of the underlying population correlation, and (2) that the jackknife variance estimator on average provides a reasonable approximation of the true uncertainty of the jackknife estimator. These findings are not sensitive to the choice of correlation coefficient (results not shown).

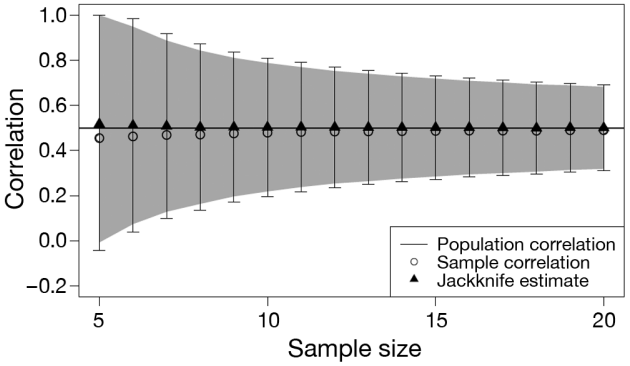

Fig. 2. (○) Sample correlation and (▲) jackknife estimates of correlation as a function of sample size for random samples from a 2-dimensional Gaussian distribution $f(X,Y)$ with mean $(0,0)$, standard deviations $\sigma_X = \sigma_Y = 1$, and population correlation $\rho(X,Y) = 0.5$. The values shown have been derived from 100 000 resampling experiments. The grey shading indicates the variability (± 1 σ) as diagnosed from the 100 000 jackknife estimates, while the error bars indicate the variability as estimated from the jackknife variance estimator (Eq. 4)

3. DRIFT-CORRECTION

3.1. Conventional drift-correction (CONV)

As outlined in the introduction, systematic model biases are unfortunately a typical feature of seasonal and decadal prediction systems. This is also evident in the ENSEMBLES decadal hindcasts. Fig. 3 shows the observed global mean temperature (as obtained from the ERA dataset) along with all decadal hindcast ensemble runs for each of the 4 models considered. The CERFACS-model runs in particular (Fig. 3a) are seen to be systematically too cold, with the magnitude of bias increasing from about –1°C to about –2°C at longer lead-times. The other 2 fully initialized models (ECMWF and UKMO) reveal a smaller, but still clearly visible drift behavior. Only the anomaly-initialized IFM-GEOMAR (Fig. 3d) runs seem to be unbiased. However this finding is not supported when the regional distribution of bias is analyzed as shown in Fig. 4. The 4 maps display the average difference between all 10-yr hindcast temperature means (only considering data for which corresponding observations are available) and the corresponding observed decadal means for each gridpoint and each model. It becomes apparent that all models reveal a pronounced regional variability of the magnitude of bias. In particular, the IFM-GEOMAR model (Fig. 4d), i.e. that model that has been seen to lack systematic bias when global averages are considered (Fig. 3d), reveals substantial biases, on the order of several degrees, in certain regions.

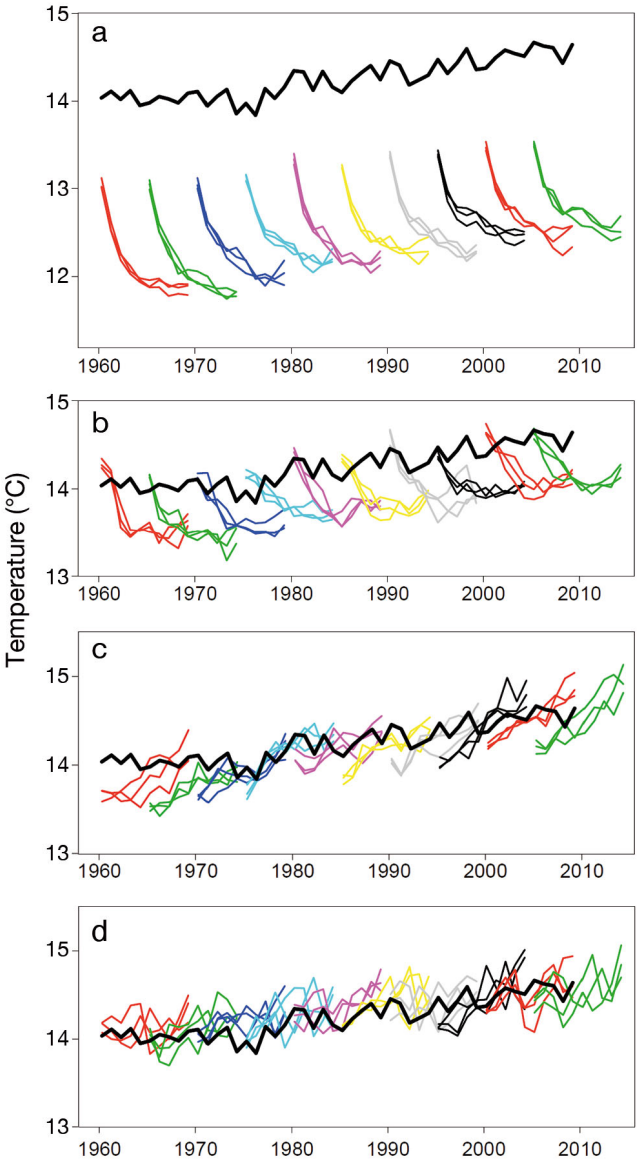

Fig. 3. Evolution of global annual mean near-surface (2 m) temperature as observed (black solid line) and simulations (colored lines) of 3 ensemble members in 10 hindcasts, for each of the 4 models considered: (a) CERFACS; (b) ECMWF; (c) UKMO; (d) IFM-GEOMAR

In monthly and seasonal forecasting, systematic biases can be considered either by issuing anomaly forecasts (i.e. by removing the mean of all reforecasts initialized at the same date for each year, and evaluated for the same lead-time), or by carrying out an explicit *a posteriori* bias correction on the basis of past forecast observation pairs. In the latter case, the following approach is conventionally applied (hereafter referred to as 'CONV'): Let $T^f(\tau)$ be a forecast of temperature for a specific grid-point or region with lead-time τ. The mean bias $\hat{d}_{\mathrm{CONV}}(\tau)$ for this prediction context is estimated by the average difference between all hindcasts $T^h(\tau)$ that have been initialized at the same time of the year and evaluated for the same lead-time τ, and the corresponding observations $T^o(\tau)$. A bias-corrected forecast $T^f_{\mathrm{cor}}(\tau)$ is then obtained by

$$T^f_{cor}(\tau) = T^f(\tau) - \hat{d}_{CONV}(\tau)$$

$$\text{with} \qquad \hat{d}_{CONV}(\tau) = \langle T^h(\tau) \rangle - \langle T^o(\tau) \rangle \tag{6}$$

This approach has also been proposed and applied for decadal predictions (International CLIVAR Project Office 2011, Doblas-Reyes et al. 2011, van Oldenborgh et al. 2012, Yeager et al. 2012). Fig. 5 shows the drift estimates (crosses) obtained with Eq. (6) for the 4 ENSEMBLES models as a function of lead-time, for global mean temperature (Fig. 5a), and for an arbitrarily selected grid-point (Fig. 5b). The plots reveal 3 aspects: (1) They confirm what has already been discussed above, namely that the magnitude of systematic model error at a specific grid-point can be substantially larger than for the global mean. (2) They confirm that model error indeed often has a tendency to increase with lead-time, consistent with the paradigm of a model drifting towards its own climate mean state. This tendency is particularly evident for the CERFACS and ECMWF models. (3) The error growth with lead-time does not follow a smooth line but reveals pronounced year-to-year variability, particularly at grid-point level. The model drift is often interpreted as a kind of relaxation process that brings a model initialized with 'real' observations gradually back towards its model climatology (e.g. Doblas-Reyes et al. 2011), and one would therefore expect the error evolution to be a smooth process (as shown with solid lines in Fig. 5; see Section 3.2 for more detail). The variability of the bias estimates found can be primarily explained by sampling uncertainty that is induced by the small number of hindcasts available. This bias variability may reach substantial values. For instance, for annual global mean temperature, the variance of the bias estimator, i.e. the uncertainty of the bias correction procedure alone, is already on the order of 40% of the year-to-year-variability of the predictand, i.e. the observed global mean temperature. Fig. 5b illustrates how this variability may vary from lead-time to lead-time. While the arbitrarily chosen point reflects a typical behavior, the behavior varies from point to point, with a variability that is sometimes smaller, sometimes larger than the point shown. In an attempt to reduce the impacts of this bias variability, we pro-

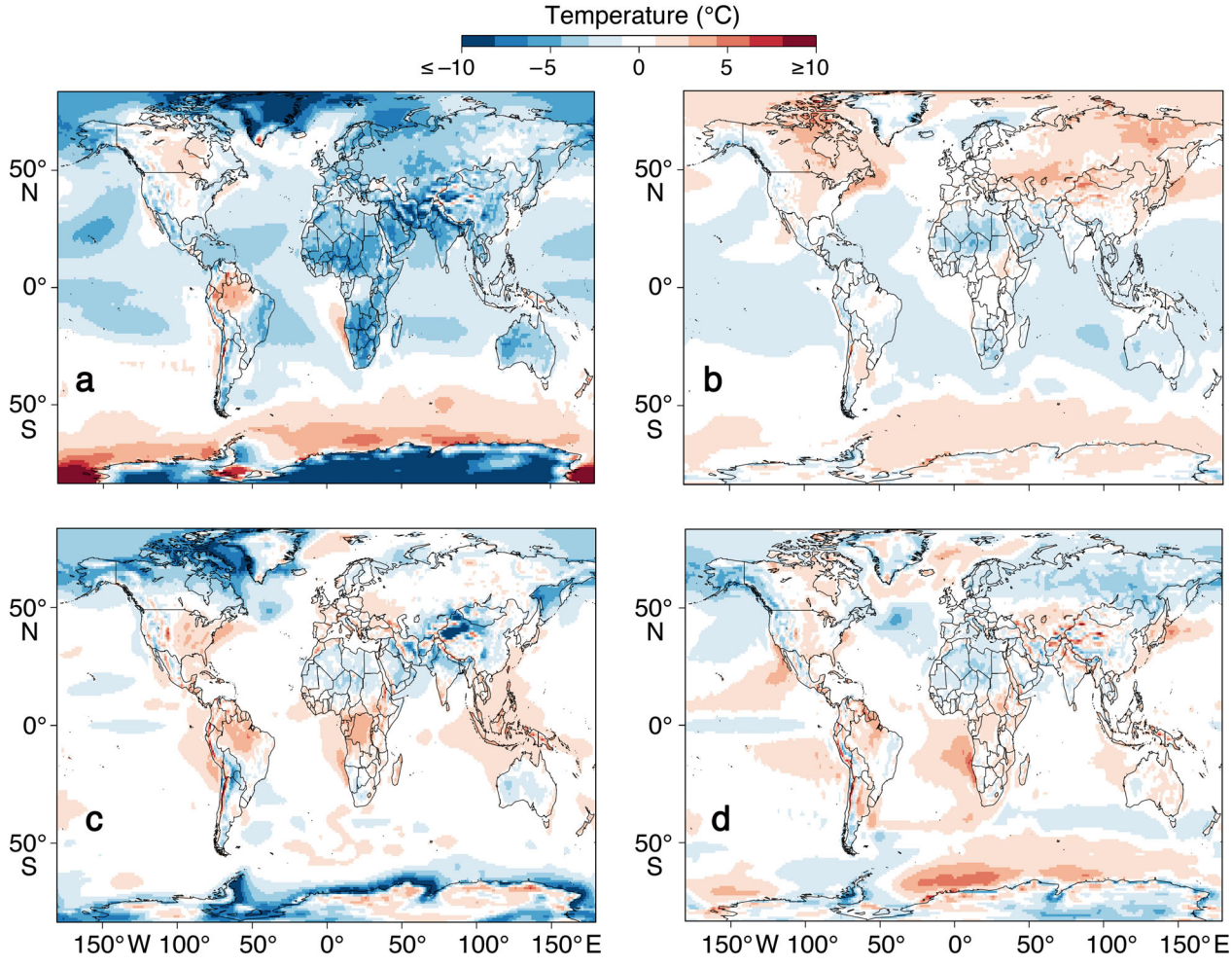

Fig. 4. Mean bias (ensemble mean minus ERA observation) for 2 m temperature (°C) for (a) CERFACS, (b) ECMWF, (c) UKMO, and (d) IFM-GEOMAR

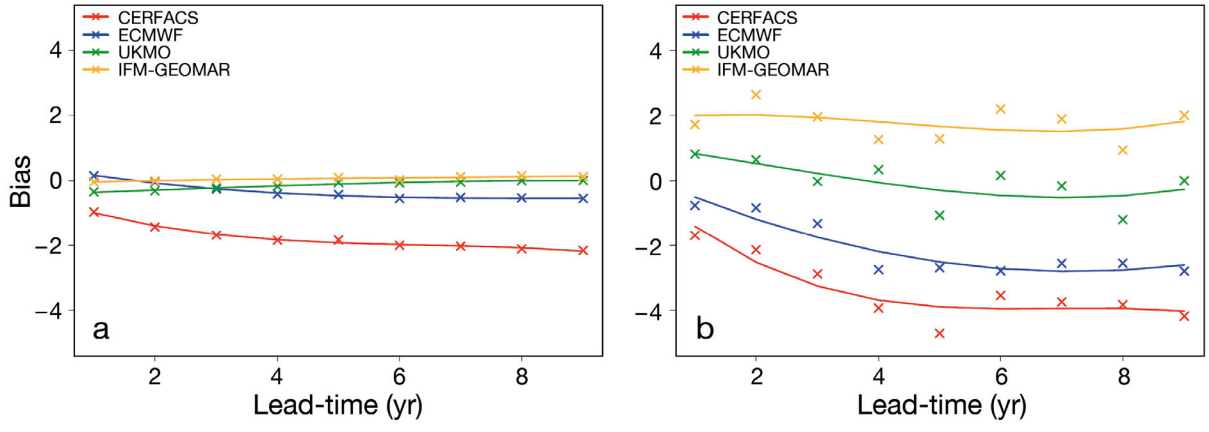

Fig. 5. Evolution of model drift defined as annual mean bias (ensemble mean minus ERA observation) with lead-time in the 4 models considered (CERFACS, ECMWF, UKMO and IFM-GEOMAR) for (a) global mean 2 m temperature (°C) and (b) 2 m temperature (°C) at an arbitrarily selected grid-point. The crosses represent the model drift obtained by the conventional drift-correction method (CONV), while the solid lines represent the model drift estimated by fitting a relaxation curve (FIT). See Sections 3.1 and 3.2 for details of the calculation of mean biases ($\hat{d}_{CONV}$) and ($\hat{d}_{FIT}$) by the two methods

pose an alternative drift-correction method which is designed to reduce sampling uncertainty by fitting a suitable 'decay function' to the drift-estimates $\hat{d}_{\mathrm{CONV}}(\tau)$ obtained by Eq. (6).

3.2. Drift-correction by fitting a relaxation curve

The method proposed here builds upon the interpretation of model drift as a smooth relaxation of a model state initialized with 'real' observations towards its mean state, i.e. towards its own model climate which may be different from the mean of the observed climate (Doblas-Reyes et al. 2011). This relaxation becomes manifest as a continuous growth of systematic error, where the error change rate is largest in the first years after initialization, and levels off as longer lead-times are considered (see Fig. 5). While in physics such relaxation processes are often modeled by an exponential decay function with offset (i.e. $a_0 + a_1\exp(-\tau/a_2)$, with a_0, a_1, and a_2 being the parameters to be estimated from the fit) we have chosen to use a third-order polynomial here instead. This is mainly for 2 reasons: (1) The fit of an exponential with offset is relatively difficult and unreliable if sample sizes are small. (2) We do not want to rule out the option that the evolution of drift may be non-monotonous and reveal a local maximum or minimum at a specific lead-time, as different climate processes and respective model components might have an influence on different time scales. Non-monotonous developments have also been observed with seasonal forecasts (Balmaseda & Anderson 2009). In comparison to higher or lower order polynomials, a third order polynomial has proved to be a good compromise between minimizing the number of parameters to be estimated from 9 to 10 data values while still having some flexibility in the shape of the curve to be fit to the data. Thus, we propose to model the lead-time dependent bias as

$$\hat{d}_{FIT}(\tau) = a_0 + a_1\tau + a_2\tau^2 + a_3\tau^3 \tag{7}$$

with a_0, a_1, a_2, and a_3, being determined from a least-squares fit of Eq. (7) to the bias-estimates $\hat{d}_{\mathrm{CONV}}(\tau)$ of Eq. (6). In Fig. 5, the drift curve estimates obtained with Eq. (7) have been included in the plot as solid lines.

The curve fitting approach of estimating systematic bias (Eq. 7) will henceforth be referred to as FIT. We now discuss the benefits of FIT compared to the conventional method (CONV) (Eq. 6).

3.3. Benefits of FIT compared to CONV

In weather or seasonal forecasting, the added value of a new model or a new post-processing scheme is, whenever possible, judged by verification. Due to the small number of hindcasts available and due to related methodological issues, such an approach is not applicable to decadal hindcasts (see the next section for a more detailed discussion). However, as an alternative strategy one can estimate the direct effects of the 2 methods (CONV and FIT) on the expectation and variance of the drift estimates obtained, and thus indirectly also judge the effects on the assessment of the prediction skill. This is done below (this section). We demonstrate that, compared to CONV, FIT yields—in a statistical sense—unbiased estimates of model drift, while reducing the sampling uncertainty. These 2 aspects of FIT imply a methodological improvement.

The analysis is carried out with the help of the jackknife estimator introduced in Section 2.2 by setting $\hat{\theta} = \hat{d}_{\mathrm{CONV}}$ or $\hat{\theta} = \hat{d}_{\mathrm{FIT}}$ in Eqs. (2) & (4). More specifically, the existence of a systematic statistical bias that might potentially be inherent to FIT was tested by calculating $B_{\hat{d}_{\mathrm{FIT}}}$ with Eq. (2), i.e. the jackknife bias estimator of $\hat{d}_{\mathrm{FIT}}$ at a given lead-time τ:

$$B_{\hat{d},FIT} = (n-1)(\hat{d}_{FIT(\bullet)} - \hat{d}_{FIT}) \tag{8}$$

with $\hat{d}_{FIT(\bullet)} = \frac{1}{n}\sum_i \hat{d}_{FIT(i)}$, and with $\hat{d}_{\mathrm{FIT}(i)}$ being the drift-corrector obtained with FIT using all hindcasts, apart from the ith one. $B_{\hat{d},\mathrm{FIT}}$ was calculated for all models, all grid-points and all lead-times and was never observed to be larger than 10^{-10} (not shown). Hence we can conclude that FIT provides unbiased estimators of model drift.

As a second step, the jackknife procedure was used to estimate the sampling uncertainty of drift estimates obtained with both methods by calculating the variance of $\hat{d}_{\mathrm{CONV}}$ and $\hat{d}_{\mathrm{FIT}}$ with Eq. (4):

$$\hat{\sigma}^2_{\hat{d},CONV} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{d}_{CONV,(i)} - \hat{d}_{CONV,(\bullet)}\right)^2$$
$$\text{and} \qquad \hat{\sigma}^2_{\hat{d},FIT} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{d}_{FIT,(i)} - \hat{d}_{FIT,(\bullet)}\right)^2 \tag{9}$$

Fig. (6) shows the variance estimates obtained for the CONV drift corrector plotted against the variance estimates obtained for the FIT drift corrector for each model and each lead-time, for global average temperature (Fig. 6a), and temperature at each grid-point (Fig. 6b). It can be seen that the sampling uncertainty is clearly reduced, on average by about 30%. For the case of global mean temperature, this reduction cor-

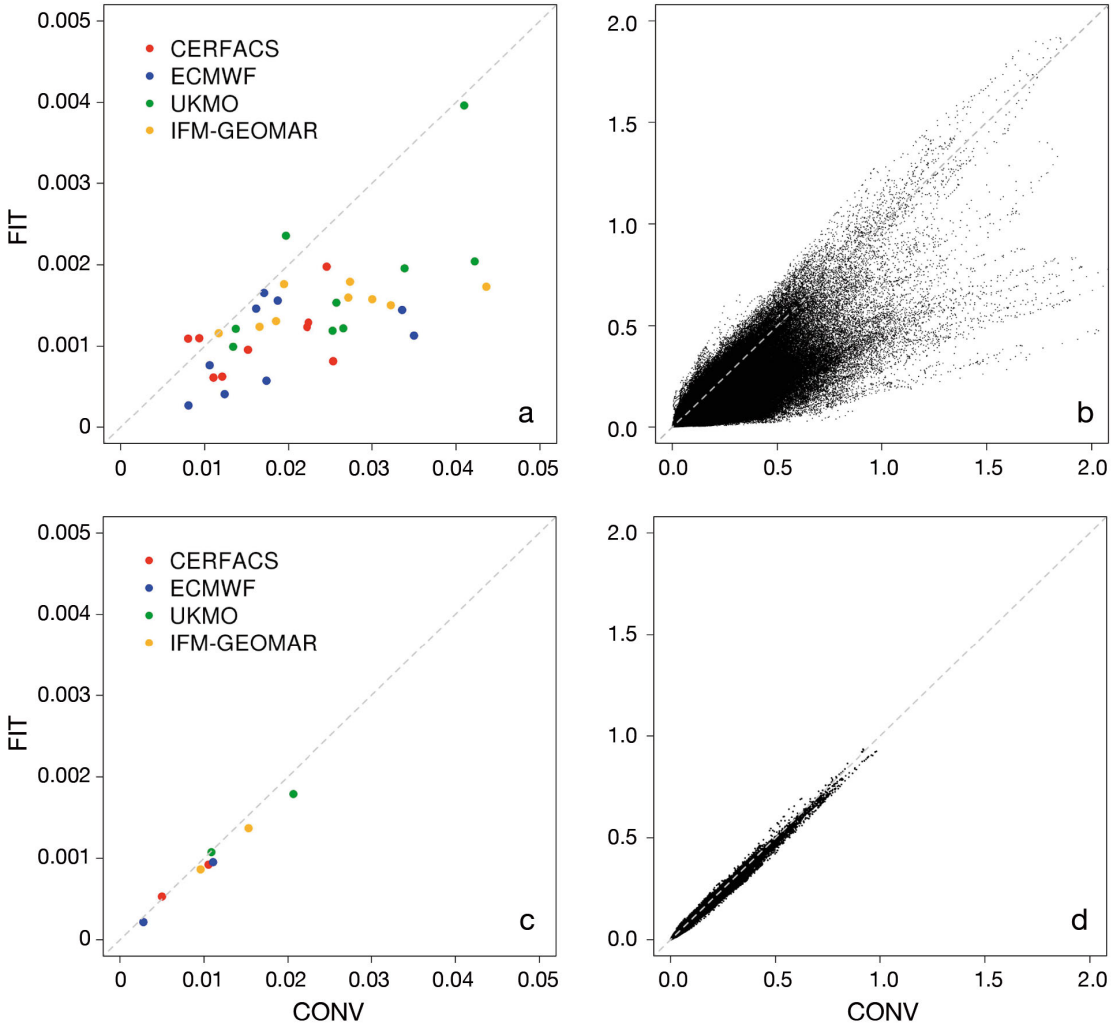

Fig. 6. Jackknife estimates of the variance of the temperature drift estimates (°C) obtained with FIT, plotted against the corresponding estimates obtained with CONV (see Fig. 5 legend for explanation of FIT and CONV), for annual averages (a) of global mean temperature and (b) of temperature at each grid-point, and (c,d) for corresponding averages over the two 4 yr lead-time periods (Years 2–5 and 6–9)

responds to about 15% of the variability of the observations, i.e. of the variability of the potentially predictable signal. We think that this reduction in sampling uncertainty is large enough to justify applying a polynomial fit, such as FIT, to obtain more robust estimates of model drift. However, we stress that this only holds if individual years or seasons are to be validated. Often, decadal prediction skill is assessed for multi-year averages and the averaging procedure implies smoothing of the bias-variance as well. For instance, van Oldenborgh et al. (2012) calculated skill scores for the averages of forecast Years 2 to 5 and 6 to 9. In this case, much of the bias variability found in the CONV drift estimates from forecast year to forecast year is indeed averaged out, thus reducing the superiority of drift estimates obtained with FIT. This is evident in Fig. 6c,d, where the sampling uncertainties (for global averaged temperature, and for the individual grid-points, respectively) of the 2 drift-correction methods applied to the averages of Years 2 to 5 and 6 to 9 are plotted against each other. Here, FIT reduces sampling uncertainty by only about 5 to 10%.

4. VERIFICATION OF SMALL SAMPLES

The quality of monthly and seasonal forecasts is usually assessed by computing skill scores over a sufficiently large set of past forecasts and hindcasts. However, it is not straightforward to apply such an approach to decadal prediction systems because (1)

the sample size is very small and (2) decadal hindcasts may not be representative of true decadal prediction skill because of the low reliability of the dataset used for initialization and verification (see Section 1). Given these conceptual problems, Mason (2011) proposed alternative strategies for the assessment of decadal prediction systems, which do not yield direct estimates of prediction skill, but which may at least provide some basic information on the trustworthiness of the models. Such a strategy could be, for example, the evaluation of how well the model climatology matches the observed climatology, or of how well one ensemble member predicts another ensemble member (perfect model approach). However, Mason (2011, p. 219) also stresses that:

> It is still worth calculating verification scores with whatever data are available. While it may be impossible to demonstrate statistically significant skill, the extent to which the models improve their simulation of the observed large-scale climate variability as improved datasets are assimilated, for example, reinforces the belief that the models may be able to make useful predictions.

It is not the aim of this paper to expand on this discussion. In fact, some verification studies of decadal forecasts have already been carried out (e.g. van Oldenborgh et al. 2012, García-Serrano & Doblas-Reyes 2012). However, we believe that it is worth highlighting a potential pitfall that may arise from the verification of small samples, as is commonly the case for decadal hindcasts. Note that the line of argumentation in Section 4.1 only holds for drift-corrected forecasts, not for the assessment of forecasts formulated as anomalies.

4.1. Cross-validation bias

The computation of skill scores relies on the comparison of a set of past forecasts or hindcasts with a corresponding set of verifying observations. Often model forecasts are modified and refined by statistical methods prior to being issued. Such a modification could, for example, be the correction of a systematic bias due to model drift as discussed above, or the removal of a linear trend as discussed in Section 5.2 (and in Liniger et al. 2007), but also more sophisticated statistical post-processing techniques such as a recalibration of ensemble spread (Weigel et al. 2009). While calibration and statistical-post-processing have become common practice in weather and seasonal forecasting, the verification of such post-processed forecasts remains a challenging issue. The key problem is that the post-processing parameters are usually determined by comparing past forecasts with the corresponding observations, i.e. they are based on the same dataset that is available for verification. However, if the verifying observations are not independent of the data used for model training, this may lead to skill estimates that are positively biased. For this reason, it has been recommended to carry out verifications in retroactive mode, meaning that for each target year to be verified only data prior to the target year are used as training data to compute the post-processing parameters (Mason & Baddour 2008). However such an approach is not feasible if sample sizes are small, at least if they are as small as is typical for decadal hindcasts. As an alternative and less data-intensive approach, a leave-one-out cross-validation technique is often applied (e.g. Wilks 2006), meaning that all years available, apart from the target year, are used as training data, including years after the target year. For instance, if a decadal hindcast initialized in 1975 is to be drift-corrected and then validated, cross-validation would imply that the drift-correction needs to be derived on the basis of all hindcasts apart from the one launched in 1975 (i.e. 1960, 1965, 1970, 1980, 1985, 1990, 1995 and 2000).

Here we argue that cross-validation bears the risk of yielding negatively biased skill estimates if applied to drift-corrected decadal hindcasts. The underlying problem is that cross-validation may lead to an implicit leakage of information from the training data to the verification sample (as discussed by e.g. von Storch & Zwiers 1999), which may in turn lead to a degeneracy in skill (as shown by Barnston & van den Dool 1993)—a problem which becomes larger the smaller the sample size is. In the following, we assess the magnitude and nature of this skill degeneracy in a prediction context that is typical for decadal hindcasts. More specifically, we assess how the correlation coefficient between a set of forecasts and corresponding observations is affected if the forecasts have been drift-corrected by CONV (for simplicity we do not consider FIT here) prior to verification, by applying a leave-one-out cross-validation approach. This will be done on the basis of synthetic forecast-observation pairs generated with a synthetic toy model, and in Section 5 also with hindcasts from the ENSEMBLES database.

Consider the following unrealistic but illustrative example. Consider a set of 4 (identical) forecasts $y_1 = y_2 = y_3 = y_4 = 12°C$, and 4 corresponding (alternating) observations $x_1 = 8°C$, $x_2 = 10°C$, $x_3 = 8°C$, and $x_4 = 10°C$. It is easy to see that the sample correlation between these 4 forecasts and the corresponding obser-

vations is zero. Now assume that each of the forecasts is bias-corrected using CONV in leave-one-out cross-validation. The bias estimator of, say, forecast y_1 is then given by $\frac{1}{3}\left[\left(y_2 - x_2\right)+\left(y_3 - x_3\right)+\left(y_4 - x_4\right)\right] = 2.67$, and the bias-corrected forecast becomes $y_{1,cor}$ = 9.33°C. Applying the same procedure on the other forecasts yields $y_{2,cor}$ = 8.67°C, $y_{3,cor}$ = 9.33°C, and $y_{4,cor}$ = 8.67°C. The new bias-corrected forecasts and the observations are now perfectly anti-correlated. As already mentioned above, the reason for this degeneracy of skill is leakage of information from the training data (which are used to calculate the bias corrector) to the forecast–observation pair to be verified. Indeed, if the difference between the omitted forecast and the corresponding observation is larger (smaller) than the average difference between the training forecasts and observations, the bias corrector becomes smaller (larger) than the average bias of all forecast-observation pairs.

This introductory example was designed to generate extreme behavior and to illustrate that cross-validation can potentially result in unexpected results. For a more systematic evaluation, we apply a stochastic Gaussian generator of forecast–observation pairs. This generator was designed such that it generates observations x and corresponding forecasts y fulfilling preset conditions with respect to forecast skill, variance and drift. These conditions are controlled by 3 free parameters, α, r, and d, with $\alpha \in [-1,1]$, $r > 0$, and d being any real number. For given α, r, and d, the following 2 steps are undertaken to generate a forecast-observation pair:

Step 1: An observation x is sampled according to:

$$x \sim N(0,1) \tag{10}$$

~N(μ,σ) thereby indicates a random number drawn from a normal distribution with mean μ and standard deviation σ.

Step 2: A corresponding forecast y is constructed by imposing a bias d and an independently sampled perturbation ε on the observation x:

$$y = r^{1/2}\left(\alpha x + \varepsilon\right) + d \tag{11}$$

with

$$\varepsilon \sim N(0, \sqrt{1-\alpha^2}) \tag{12}$$

A more in-depth discussion of toy models of this kind is provided in Weigel & Bowler (2009). α is the population correlation between the forecasts and the observations, r is the ratio between the variance of the forecasts and the variance of the observations, and d is a systematic bias term resembling the idea of model drift. Without loss of generality, d is set to 1 (the results described in the following are not sensitive to the choice of d). For fixed settings of α and r, the toy model is used to generate n forecast-observation pairs, resembling a set of n decadal hindcasts for a specific lead-time. The n synthetic forecasts $y_1,...,y_n$ are then drift-corrected with CONV in leave-one-out cross-validation, yielding a new set of corrected forecasts $y_{1,cor}$, ..., $y_{n,cor}$ with

$$y_{i,cor} = y_i - \frac{1}{n-1}\sum_{j \neq i}(y_j - x_j) \tag{13}$$

These bias-corrected forecasts y_{cor} are correlated with the observations x, yielding a sample correlation value $\hat{\rho}$. This procedure is repeated 10 000 times for a given setting of α and r, and the expected sample correlation $\langle\hat{\rho}(n,\alpha,r)\rangle$ is determined. Results of $\langle\hat{\rho}(n,\alpha,r)\rangle$ as obtained for different combinations of n, α and r are shown in Fig. 7. Fig. 7a shows $\langle\hat{\rho}(n,\alpha,r)\rangle$ as a function of n for α = 0 and r = 1, i.e. forecasts and observations are modeled to be independent from each other, and to have the same variance. This mimics the situation of bias-correcting and verifying one ensemble member without skill on the basis of n hindcasts. The estimates of $\langle\hat{\rho}(n,\alpha,r)\rangle$ are shown as open circles. Fig. 7b is the same as Fig. 7a, but with r = 1/12, mimicking a theoretical multi-model with 4 fully independent models, each with 3 fully independent ensemble members (thus representing an upper boundary for the effective number of independent ensemble members in our dataset). In Fig. 7c, $\langle\hat{\rho}(n,\alpha,r)\rangle$ is plotted as a function of α for n = 10 and r = 1/12. The results confirm that a cross-validated drift-correction leads to a significant negative bias of the estimates of correlation skill (the forecasts have zero skill by definition in this experiment). The magnitude of bias increases (1) if the sample size is reduced (i.e. if fewer hindcasts are available); (2) if the variance of the forecasts is reduced with respect to the variance of the observations (e.g. if several ensemble members are averaged); and (3) if the population correlation is low (i.e. if the forecasts have only little predictability). For a prediction context that is typical for decadal hindcasts (low skill and small sample sizes on the order of 10 or less), the negative bias in correlation skill would be on the order of 0.1 to 0.3, or even larger, depending on the variance ratio between observations and forecasts. This negative bias can almost entirely be attributed to the cross-validated drift-correction procedure, rather than to the bias of the classical sample correlation estimator as depicted in Fig. 2. This becomes evident in Fig. 7d, which is as Fig. 7c, but showing the sample correlation for toy model fore-

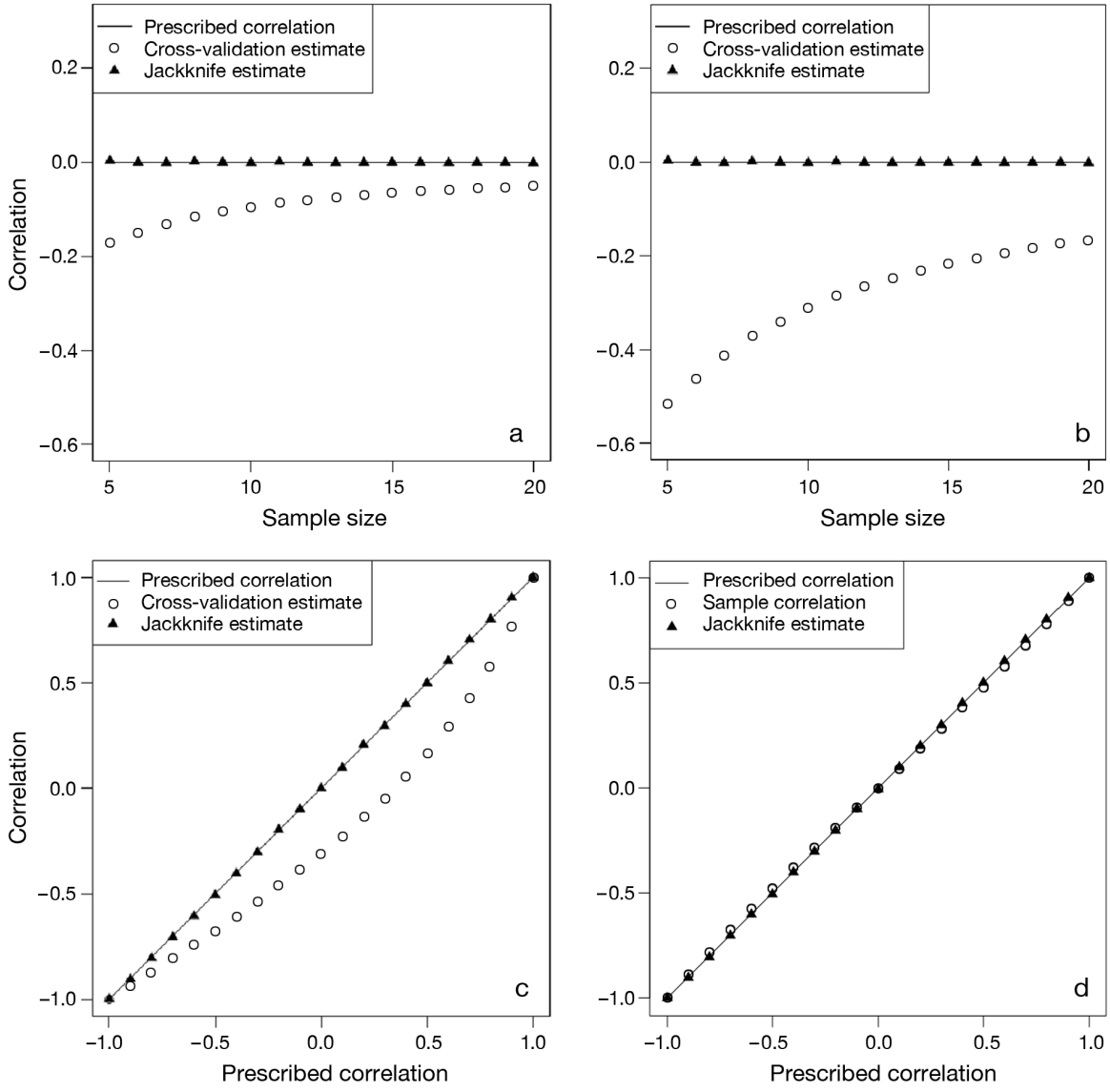

Fig. 7. Cross-validation and jackknife estimates for toy model experiments. For prescribed population correlation, prescribed ratio between forecast and observation variance, prescribed systematic forecast bias and prescribed sample size, the plots show the sample correlation to be expected after bias-correcting the forecasts with CONV in (○) leave-one-out cross-validation, and (▲) the corresponding jackknife estimates. Each value shown is based on 10 000 toy model experiments. (a) Expected correlation estimates as a function of sample size for population correlation 0 and for forecasts and observations having the same variance. (b) As (a), except with variance ratio 1:12. (c) Expected correlation as a function of prescribed population correlation for sample size 10 and variance ratio 1:12. (d) As (c), but for forecasts that lack a systematic bias and therefore do not require a drift-correction nor a cross-validation

casts that lack systematic bias (i.e. $d = 0$) and therefore do not need to undergo a bias correction procedure in cross-validation (resembling a prediction context where forecasts are interpreted as anomalies with respect to a well-defined model mean state). One can see that Fig. 7d lacks the large biases that have been apparent in Fig. 7c. Only closer inspection reveals that the magnitudes of the correlation estimates still slightly undercut those of the prescribed 'true' correlation, but to a much smaller degree than in the case of drift-corrected forecasts.

Of course, the example presented here is rather specific and theoretical. However, it illustrates that even seemingly small data manipulations such as a simple drift-correction can lead to severely biased skill estimates if validated in cross-validation, which in turn may lead to a misinterpretation of the results. The nature of this cross-validation bias may become

even more difficult and less transparent if more sophisticated statistical post-processing techniques are applied, such as de-trending (see Section 5.2) or recalibration, and it may become manifest in different ways for different forecasting techniques. For instance, if forecasts are formulated as anomalies rather than drift-corrected absolute values, with the anomalies being determined in leave-one-out cross-validation, then the leakage of information due to cross-validation does not affect correlation values, but does have an impact on other metrics such as the root-mean-squared error (not shown here).

As a consequence of this finding, we recommend that cross-validation approaches should be critically questioned in the context of the verification of drift-corrected decadal hindcasts. However, omitting cross-validation may lead to an overestimation of skill. Therefore, rather than attempting to validate the actual prediction skill of post-processed decadal forecasts on the basis of a very small hindcast data set, we think the focus should be more on an assessment of the potential prediction skill, i.e. the skill that could be expected were the statistical model parameters (such as systematic model bias) accurately known. Note that this definition of potential skill is not identical to the one used in e.g. perfect model experiments, where decadal potential predictability is defined as the ratio of variance on a decadal time scale to the total variance (e. g. Latif et al. 2006). Positive values of potential skill do not imply that the actual prediction skill of the forecasts is positive and significant as well, but they represent a necessary condition that needs to be satisfied for decadal forecasts to be useful at all. The jackknife method described in Section 2.2 provides a very efficient tool to directly obtain such estimates of potential prediction skill. We illustrate this at the example of the toy model simulations discussed above.

4.2. Estimating potential skill by jackknifing

Similarly to the way the jackknife estimator was used to estimate the expected model drift (Section 3.4), or to estimate the population correlation from samples of a 2-dimensional normal distribution (Section 2.2), it can be applied to estimate the underlying potential 'population skill' without bias from a set of forecast-observation pairs, whether the forecasts are post-processed or not. Applying Eq. (3) to the aforementioned toy model example, the jackknife estimator of the correlation skill after drift-correcting the forecasts is given by:

$$\hat{\theta}_J = \hat{\theta} - (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta}) \quad (14)$$

with

$$\hat{\theta} = \hat{\rho}\left[x,\ y - \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right) \right] \quad (15)$$

with $\hat{\theta}_{(\bullet)}$ being defined in an equivalent way as in Eq. (1).

The resulting jackknife estimates are shown in Fig. 7 as filled triangles. It can be seen that the jackknife estimates consistently represent good approximations of the underlying prescribed correlation skill. Also for the situation of forecasts that are not subject to systematic drift and therefore not undergoing a cross-validated drift-correction procedure (Fig. 7d), the jackknife estimators removes the small bias that remains from the sample correlation estimator (see also Fig. 2). In practice, this situation usually applies to anomaly forecasts with a well-defined model mean state.

Although cross-validation and jackknifing have similarities, in that both rely on the re-computation of skill values from subsets of the verification data, they also have very different functions. Cross-validation aims at inferring future forecasting performance from a finite sample, whereas the jackknife characterizes the underlying 2-dimensional distribution that connects the sampled observations with the forecasts.

5. APPLICATION TO ENSEMBLES DECADAL HINDCASTS

5.1. Cross-validation and jackknifing

The effects of cross-validation and jackknifing on correlation skill of a real decadal prediction system were examined for the 9 yr ENSEMBLES multi-model mean hindcasts of 2 m temperature. The correlation skill was calculated once for global mean temperature, and once for each individual grid-point. In both cases, the following procedure was applied: (1) The 3 initial condition ensemble members of each of the 4 models were averaged. (2) The resulting 4 sets of ensemble mean hindcasts (1 set for each model) were then drift-corrected individually, either by CONV or FIT. (3) The drift-corrected single model hindcasts were then averaged, yielding a set of drift-corrected multi-model mean hindcasts. (4) The correlation between these multi-model means and the corresponding observations was calculated for each lead-time. The entire procedure was done by leave-one-out cross-validation as well as with jackknifing. Fig. 8 shows, as a function of lead-time, the results ob-

tained for global mean temperature, and Fig. 9 shows the average of the correlation values obtained for each grid-point. The black lines are for a validation of annual means, while the grey lines are for a validation of 4-yr means (averages of forecast Years 2 to 5 and 6 to 9). In Figs. 8 & 9, the dotted lines indicate the skill values obtained in leave-one-out cross-validation with drift-correction method CONV, while the dashed lines are based on cross-validation with FIT. The solid lines are the jackknife estimates obtained with either FIT or CONV (they are identical, as discussed below, this section), and the grey shading indicates the corresponding jackknife estimates of sampling uncertainty at grid point level (shown as ±1.65 standard deviations, corresponding to 90% confidence intervals).

We start with a discussion of the evaluations of annual mean temperature (black lines). It can be seen that both for the global means and for the grid-point averages, the hindcasts that have been drift-corrected with FIT in cross-validation yield higher correlation values than those that have been drift-corrected with CONV, and that the jackknife estimates exceed the cross-validated ones for either method of drift-correction. Skill values of global mean temperature are on the order of 0.7 to 0.9.

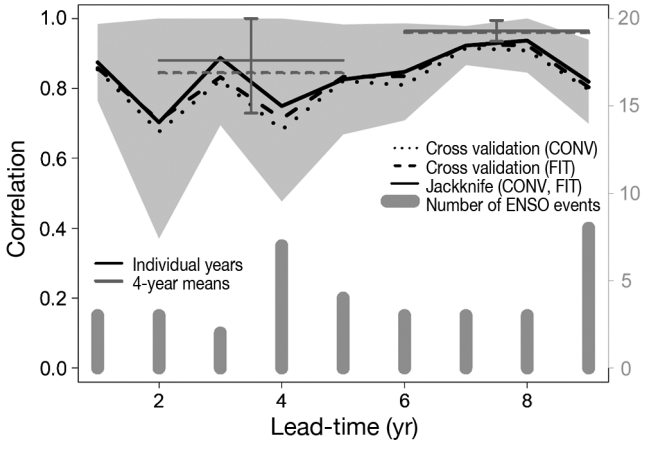

Fig. 8. Correlation of decadal predictions of global mean 2 m temperature (multi-model mean) with observations as a function of lead-time. The black lines show results for annual means, and the grey lines for 4 yr means (Years 2–5 and 6–9). The validation was carried out in cross-validation (dotted and dashed lines) and with jackknifing (solid lines). Dotted lines: drift-correction with CONV [for the 4 yr means, grey dashed lines partly obscure these lines], dashed lines: drift-correction with FIT, and solid lines: independent of drift-correction method. The jackknife estimates of sampling uncertainty (90% confidence intervals) are shown as grey shading and grey error bars, for annual and 4 yr means, respectively. The grey bars denote the number of ENSO events recorded for each lead-time

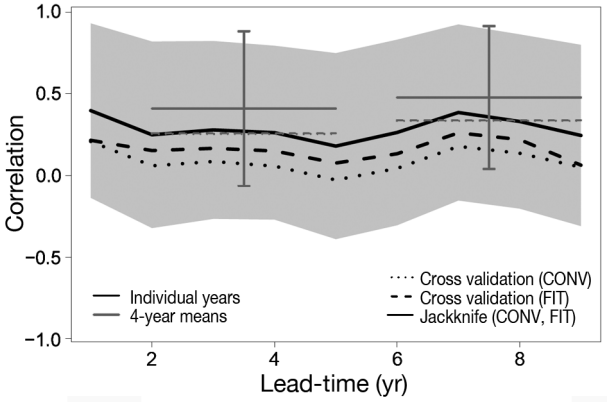

Fig. 9. Correlation of decadal predictions of 2 m temperature calculated at each grid-point (multi-model mean) and then averaged over the globe, with observations as a function of lead-time. See Fig. 8 legend for explanation of lines and shading. See Fig. 8 for no. of ENSO events

Van Oldenborgh et al. (2012) showed that the main source of these high skill values is the long-term trend, which is primarily forced by greenhouse gases and aerosols (see also Section 5.2 below). The average correlation at grid-point level is on the order of 0.2 to 0.4, and thus substantially lower. This difference is plausible, since the latter are sensitive to the models' ability to reproduce the spatial distribution of the predictable signal, while the former are not. While it would be interesting to assess regions to identify minimum spatial scales with predictive information, we leave this for future research. A viable strategy to find these scales could be the approach described by Masson & Knutti (2011) in the context of CMIP3 climate change projections.

The skill difference between the jackknife estimates and the CONV-corrected cross-validated skill values is much smaller for the global averages (on the order of 0.05 or less; Fig. 8) than for the grid-pointwise evaluation (on the order of 0.2; Fig. 9). This is consistent with the toy model simulations shown in Fig. 7c, which reveal that the cross-validation bias disappears as the correlation approaches higher values. In both Figs. 8 & 9, the FIT-corrected forecasts have a lower bias (i.e. higher skill) than the CONV-corrected ones. This is due to the reduced sampling uncertainty of the drift estimates as discussed above, and implies that less 'false' variability is imposed on the forecasts in the course of the cross-validation procedure. In contrast to the cross-validated skill values, the jackknife estimates are insensitive to the drift-correction method applied. This can be explained by the fact that in cross-validation each of the forecasts to be considered in the correlation is corrected by a different bias estimate derived from the remaining

forecasts, implying that there is sensitivity to the drift-correction method applied. In jackknifing, on the other hand, all forecasts considered in a correlation estimate, be it the correlation of the full set ($\hat{\theta}$ in Eq. 3) or the correlations of the reduced sets ($\hat{\theta}_{(i)}$ in Eq. 3), are bias-corrected with the same value, implying that there is no sensitivity to the drift-correction method, since the correlation operator is invariant to a constant being added to the forecasts.

If 4 yr averages are considered (grey lines), skill values are generally higher as compared to the annual means. This is a known effect, and can be explained by the fact that some of the unpredictable noise (such as ENSO events) is averaged out as the prediction intervals grow (e.g. Weigel et al. 2008b, García-Serrano & Doblas-Reyes 2012). Moreover, the cross-validation bias gets very small for correlations close to 1 (see forecast for Years 6 to 9 in Fig. 8). For these averages, the choice of drift-correction method has hardly any impact on the skill values obtained. This is consistent with the findings of Fig. 6c that revealed that the sampling uncertainty of the 2 methods converges as multi-annual averages are considered.

The variance estimates provided by jackknifing indicate that sampling uncertainty is comparatively large (grey shading and grey error bars in Figs. 8 & 9). For the global mean temperature forecasts, the 90% intervals shown may range from 0.4 to 1, depending on lead-time. For the grid-pointwise verification, the intervals represent the averaged sampling uncertainty at a single grid point and the ranges are even wider, ranging from −0.3 to about 0.8.

While it is safe to argue that, at least for the global averages, predictability is significantly positive for all lead-times, the skill variability found between individual lead-time years is not statistically significant. For instance, is the correlation skill at Years 2, 4 and 9 really lower than at the other years? García-Serrano & Doblas-Reyes (2012) found that dips in prediction skill often correspond to forecast years that contain a disproportionately high number of ENSO-years, a finding that has also been reproduced in the present study. It is consistent with the fact that current models are not able to predict ENSO on a time scale beyond ~2 yr (Luo et al. 2008). Fig. 8 shows, as grey bars, the number of ENSO events recorded for each simulated lead-time (ENSO events were diagnosed on the basis of the multivariate ENSO index of the 4 mo period November–February; Wolter & Timlin 2011). It can be seen that 2 of the 3 low-skill lead-time years indeed correspond to a high number of ENSO events (Years 4 and 9). However, since the uncertainty ranges are very large, the question as to whether or not this is a coincidence cannot be answered on the basis of such a small sample of verification data.

5.2. The effect of de-trending

Van Oldenborgh et al. (2012) showed that the main source of the positive skill seen in Figs. 8 & 9 is the long-term trend. Here we expand on this issue by applying the FIT drift-correction and jackknife skill estimator to de-trended hindcast data. The goal is on the one hand to assess the effect of de-trending on potential predictability, and on the other to judge the sensitivity of the results to the choice of de-trending method. Three different procedures to de-trend the data were implemented: (1) Observations and model data are de-trended with a linear trend as diagnosed from the observations over the period considered. (2) The observations are de-trended as in (1), but the model data are de-trended separately with a linear trend as diagnosed from the hindcasts (Liniger et al. 2007). (3) Observations and model data are de-trended by a regression of the observations to global annual mean CO_2-concentrations. That is, the trend is interpreted as the part of the signal that is proportional to the rising CO_2 concentrations (van Oldenborgh et al. 2012). In all cases, the model data were drift-corrected by FIT prior to determining and removing the model trend (see Section 3). The correlation skill of the de-trended multi-model mean forecasts was calculated for global mean temperature, and for temperature at each grid-point. Fig. 10 shows the results for global mean temperature, and (Fig. 11) shows the global average of the grid-point validation. As above, the evaluations of annual means are shown in black, and the 4 yr means are shown in grey, and the corresponding jackknife uncertainty estimates are displayed respectively as grey shading and grey error bars. The correlation of de-trended global temperature in Fig. 10 is on average substantially lower (average correlation 0.14) than in the corresponding case without de-trending in Fig. 8 (average correlation 0.84). In Year 1, the correlation is comparatively high and of the order of 0.65. This is likely due to persistence of sea surface temperature anomalies in combination with the evolution of ENSO. The development of skill with lead-time does not reveal a clear and consistent tendency, but fluctuates with large amplitude around zero, in the range between about −0.6 (in Year 4) and 0.8 (in Year 7). Similar differences are seen for the 4 yr averages, again with

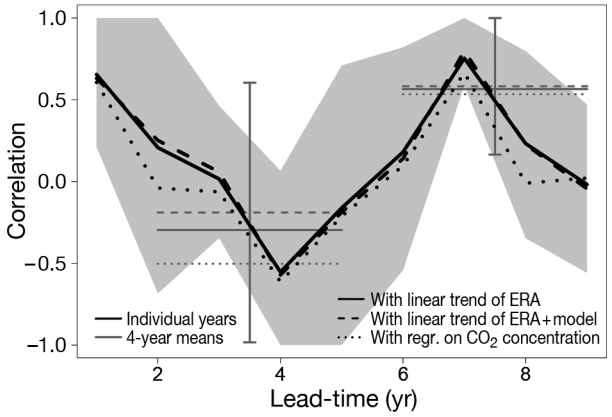

Fig. 10. Correlation of de-trended decadal predictions of global mean 2 m temperature (multi-model mean) with de-trended observations as a function of lead-time. Black lines: annual means; grey lines: 4 yr means (Years 2–5 and 6–9). The validation was carried out with jackknifing, and model-drift was corrected with FIT. The jackknife estimates of sampling uncertainty (90% confidence intervals) are shown as grey shading and grey error bars, for annual and 4 yr means, respectively. Three methods to remove the trend were applied as follows. Solid line: linear trend of observations subtracted both from observations and hindcasts. Dashed line: linear trend of observations subtracted from observations, and linear trend of hindcasts subtracted from hindcasts. Dotted line: regression of the observations to global annual mean CO_2 concentrations subtracted from both observations and hindcasts

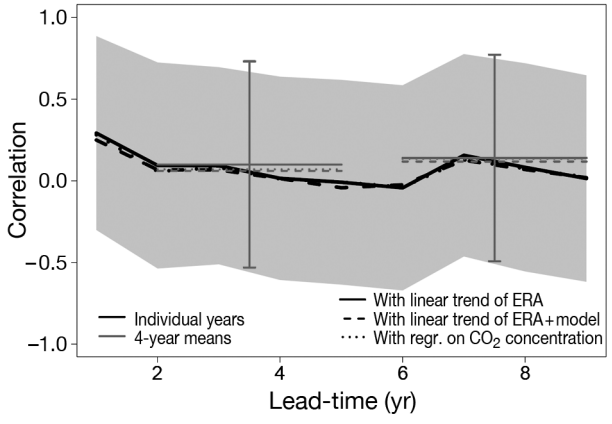

Fig. 11. Correlation of de-trended decadal predictions of 2 m temperature calculated at each grid-point (multi-model mean) and then averaged over the globe, with de-trended observations as a function of lead-time. The same procedures were applied as in Fig. 10. See Fig. 10 legend for explanation of lines and shading. Dotted lines partly obscured by other lines

higher values for the Years 6 to 9 compared to 2 to 5. These fluctuations may be the consequence of sampling uncertainty (note that each of the correlation values is based on only 9 samples, a sample size which requires that correlation is >0.67 or <–0.67 in order to be significantly different from zero on the 5% level). The surprisingly high skill value at Year 7 may be interpreted as a statistical feature in a sense of an outlier, particularly since there is no known physical process that could serve as an explanation for such a sudden outburst of skill after 7 yr of integration, but more research is necessary to arrive at a definite conclusion. One may wonder why the skill variability is so much larger here as compared to the corresponding result without de-trending (Fig. 8). The reason is the reduced predictability of the de-trended data, which implies higher uncertainty when estimating a sample correlation (demonstrated with toy model simulations; not shown here). Indeed, any subset of a set of perfectly correlated data-pairs would again be perfectly correlated (i.e. no sampling uncertainty), while a small random subset of a set of uncorrelated data-pairs could in principle yield any sample correlation (i.e. high sampling uncertainty). In comparison to this sampling uncertainty, the choice of de-trending method is clearly of minor importance. A smoother picture is obtained for the average grid-point validation in Fig. 11, where 360 × 175 individual correlation values have been averaged. The differences to the corresponding result without de-trending (Fig. 9) can be summarized as follows. (1) The jackknife estimates of grid point skill are, except from in Year 1, close to zero. (2) Averaging over 4 yr has hardly any effect on skill, which is an indication that in today's forecast systems there is little predictable signal on grid-point level visible after noise reduction. (3) The uncertainties resulting from the choice of de-trending method are about an order of magnitude lower than the uncertainty resulting from sampling uncertainty and thus negligible.

But can one conclude from this that there is no predictability at all on decadal time-scales? Fig. 12 shows the spatial distribution of the skill values obtained by grid-pointwise verification. The maps compare the correlation of observations with forecast averages for Years 2–5 and 6–9 before and after a linear trend has been removed. While the skill is clearly higher before de-trending, after trend removal a few regions still exhibit positive values, such as the North Atlantic and the Indian Ocean. Their extent and location depend to some extent on the average period considered, and on a global average these positive values are partly compensated by negative values in other regions, such as in the Pacific and southern tropical Atlantic. The results nevertheless indicate that some areas exhibit regionally predictable processes, notably the North Atlantic.

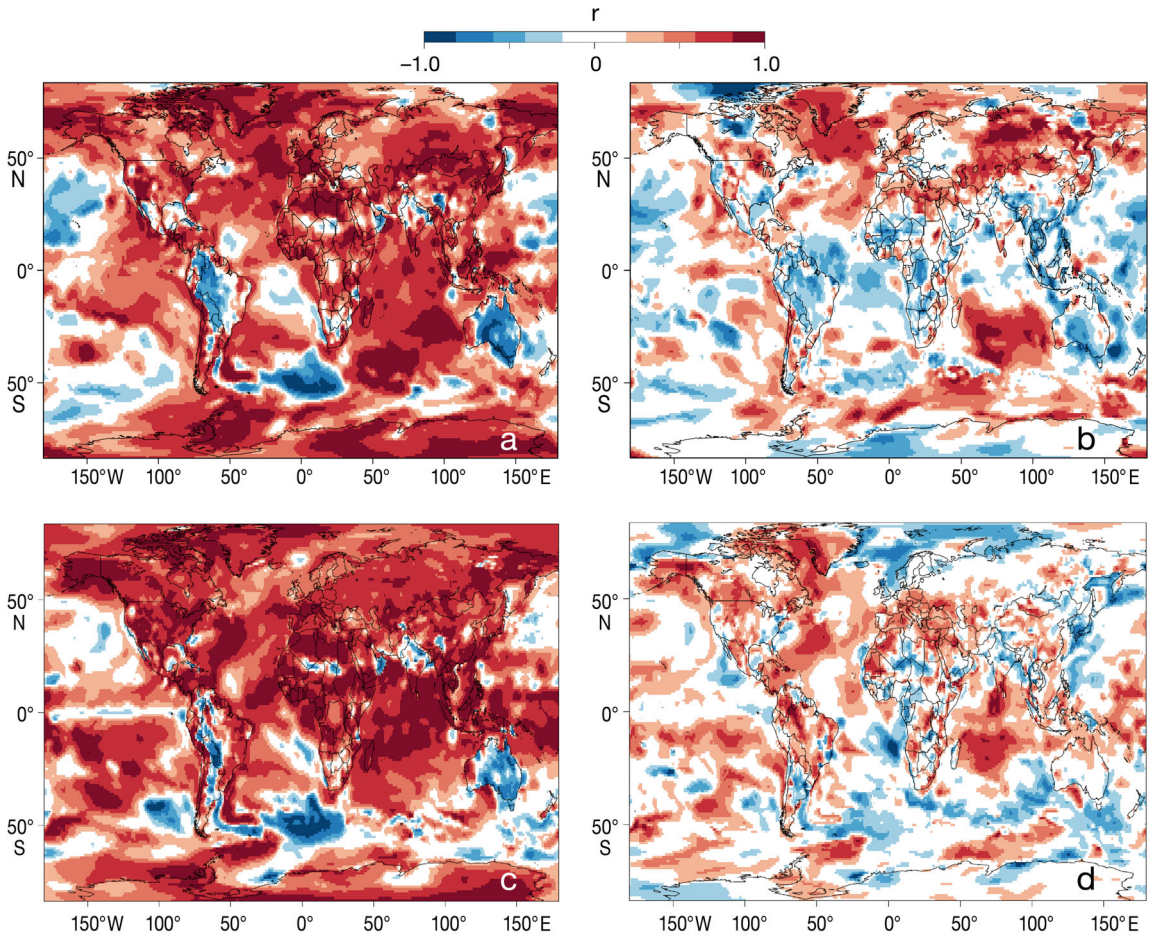

Fig. 12. Correlations of means for (a,b) Years 2–5 and (c,d) Years 6–9 with jackknifing (a,c) before and (b,d) after de-trending. These grid-point values are used for the global average shown in Figs. 9 & 11 (as grey solid lines)

All in all, these results are consistent with the findings of van Oldenborgh et al. (2012) in that—except from in the first prediction year and some regions—the trend appears to contain the larger part of the predictable temperature signal at the majority of grid-points. Again, the small sample size and the corresponding sample uncertainties do not allow us to decide, by means of statistically robust verification, whether or not today's decadal forecast systems have significant skill beyond the long-term trend.

6. CONCLUSIONS

Decadal prediction is still a very new field of research with many open issues. Among them are the understanding of the underlying processes of decadal predictability, optimum initialization and modeling techniques, but also methodological issues concerning the validation and post-processing of decadal forecasts. Most of the methodologies published so far for the validation of decadal forecasts build upon techniques that are well-established in seasonal forecasting. For instance, systematic model biases and predictability of seasonal forecasts are usually quantified on the basis of past forecasts and hindcasts. If the seasonal forecasts to be verified have been post-processed (e.g. correction of systematic model bias), the verification is usually done in cross-validation mode to avoid using the same data for the derivation of the post-processing parameters (e.g. model bias correctors) and for the validation of the forecasts. And if the prediction skill resulting from seasonal predictability is to be separated from the skill due to global warming trends, the verification data are often de-trended prior to verification. Currently there is still no consensus on the degree to which such techniques are also applicable to decadal forecasts. It has been the aim of this paper to contribute to this discussion by analyzing the effects of

drift-correction, cross-validation and de-trending for the validation of decadal hindcasts.

Our analysis was carried out on the basis of a multi-model ensemble (4 models with 3 initial condition ensemble members each taken from the ENSEMBLES project) of 10 yr hindcasts of 2 m temperature, which have been launched in 5 yr intervals from 1960 through 2005, as well as on the basis of a synthetic toy model. The key findings can be summarized as follows:

1. Model drifts are better corrected if a smooth (polynomial) drift curve is fitted along the temporal evolution instead of computing the independent averages. Conventionally, the lead time dependent mode biases due to model drift are estimated by computing the average difference between forecasts and observations for each lead-time year. However, estimates obtained that way are subject to sampling uncertainty, which can be on the order of about 40% of the variance of the potentially predictable signal. By fitting a smooth (polynomial) drift curve through the temporal evolution of the annual bias estimates, this sampling uncertainty can be reduced by about 30%.

2. Simple leave-one-out cross-validation is not recommended for the assessment of drift-corrected decadal predictions, since it may yield strongly biased skill estimates, at least if the correlation coefficient is used as a skill metric. This verification bias is particularly large if the number of verification samples is small, if the forecast variance is smaller than the observation variance (as is for example the case if several ensemble members are averaged), and if potential predictability is low. For the assessment of anomaly predictions by the leave-one-out cross-validation, other effects may be expected whose assessment is left for future research.

3. Due to the aforementioned cross-validation bias, it is difficult to estimate the actual (as opposed to potential) prediction skill of post-processed drift-corrected decadal hindcasts without bias, since alternative validation methods (such as retroactive verification or leave-one-out cross-validation) are not feasible with the small sample sizes available. One alternative option could be to only validate raw model output, i.e. ignoring biases and drift. Another option could be to focus on the potential prediction skill, i.e. the skill that would be obtained if there were enough samples to estimate the post-processing parameters without bias. Jackknifing represents a suitable technique to estimate potential skill without bias, and to quantify the underlying sampling uncertainty.

4. For the multi-model ensemble considered, the jackknife estimates of correlation indicate significant correlation skill on the order of 0.7 to 0.9 for predicting global annual mean temperature on lead-times of up to 9 yr. Skill is much lower for evaluations on a grid-point basis with an expected average correlation on the order of 0.2 to 0.4. In both cases, skill is improved by 0.1 to 0.2 if 4-yr averages are considered rather than annual means.

5. If we assume in a first order approximation that the prediction skill resulting from greenhouse gases and aerosol forcing can be removed by a linear fit prior to verification, global mean skill in the order of 0.65 is identified in the first prediction year after detrending. Beyond the first year, no clear and consistent tendency in global mean skill can be demonstrated, regardless whether individual years are considered or 4 yr averages. Regional skill is nevertheless found for longer lead times even after detrending, notably in the North Atlantic and the Indian Ocean. The uncertainty estimates are larger than for the data that have not been de-trended and the skill shows strong fluctuations with lead-time, in particular if global mean values are considered. Some additional uncertainty arises from the choice of de-trending method, but this uncertainty is an order of magnitude smaller than the sampling uncertainty.

As mentioned above, it is an ongoing discussion as to how representative the current decadal hindcasts systems are in characterizing decadal predictability. In this study we have shown that, even if the hindcasts were perfectly representative of true skill, it is not straightforward to apply validation and post-processing techniques that are well-established in seasonal forecasting to the context of decadal forecasts. Some of these issues may be partially remedied by methodological adjustments (e.g. by an improved drift-correction), and some of them may be circumvented by changing the focus of the validation (e.g. assessment of potential rather than actual prediction skill). However, as long as it is not possible to compute a larger number of independent and representative decadal hindcasts, one of the key problems for the application of classical verification approaches remains; namely, the massive sampling uncertainty. Consequently, the quantification of benefits due to improvements in decadal model physics, data sources and initialization schemes remains a great challenge. In the meantime, emphasis should be put on perfect model approaches and the assessment of fundamental model characteristics such as their capability to reproduce observed mean climatology or specific processes.

Acknowledgements. This study was supported by Swiss Re and the Swiss National Science Foundation through the National Center for Competence in Research (NCCR-Climate). We acknowledge the FP6 ENSEMBLES project for the decadal hindcasts. The valuable comments of 2 anonymous reviewers greatly improved our manuscript.

LITERATURE CITED

Balmaseda M, Anderson D (2009) Impact of initialization strategies and observations on seasonal forecast skill. Geophys Res Lett 36:L01701, doi:10.1029/2008GL035561

Barnston A, van den Dool H (1993) A degeneracy in cross-validated skill in regression-based forecasts. J Clim 6: 963–977

Collins M, Botzet A, Carril A, Drange H and others (2006) Interannual to decadal climate predictability in the North Atlantic: a multimodel-ensemble study. J Clim 19:1195–1203

Dee D, Uppala S, Simmons A, Berrisford P and others (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597

Doblas-Reyes F, Hagedorn R, Palmer T (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. II. Calibration and combination. Tellus A 57: 234–252

Doblas-Reyes F, Weisheimer A, Palmer T, Murphy J, Smith D (2010) Forecast quality assessment of the ENSEMBLES seasonal-to-decadal stream 2 hindcasts. European Centre for Medium-Range Weather Forecasts (ECMWF) Tech Memo 621

Doblas-Reyes FJ, Balmaseda MA, Weisheimer A, Palmer TN (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

Domingues CM, Church JA, White NJ, Gleckler PJ, Wijffels SE, Barker PM, Dunn JR (2008) Improved estimates of upper-ocean warming and multi-decadal sea-level rise. Nature 453:1090–1093

Dunstone NJ, Smith DM, Eade R (2011) Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. Geophys Res Lett 38:L14701, doi:10.1029/2011GL047949

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37:36–48

Fyfe JC, Merryfield WJ, Kharin V, Boer GJ, Lee WS, von Salzen K (2011) Skillful predictions of decadal trends in global mean surface temperature. Geophys Res Lett 38: L22801, doi:10.1029/2011GL049508

García-Serrano J, Doblas-Reyes FJ (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

Hagedorn R, Doblas-Reyes F, Palmer T (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. I. Basic concept. Tellus A 57:219–233

International CLIVAR Project Office (2011) Decadal and bias correction for decadal climate predictions. International CLIVAR Project Office, CLIVAR Publication Series 150

IPCC (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. Cambridge University Press, Cambridge

Ishii M, Kimoto M (2009) Reevaluation of historical ocean heat content variations with an XBT depth bias correction. J Oceanogr 65:287–299

Keenlyside NS, Latif M, Jungclaus J, Kornblueh L, Roeckner E (2008) Advancing decadal-scale climate prediction in the North Atlantic sector. Nature 453:84–88

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

Krueger O, Von Storch J (2011) A simple empirical model for decadal climate prediction. J Clim 24:1276–1283

Latif M, Collins M, Pohlmann H, Keenlyside NS (2006) Review of predictability studies of Atlantic sector climate on decadal time scales. J Clim 19:5971–5987

Latif M, Delworth T, Dommenget D, Drange H and others (2010) Dynamics of decadal climate variability and implications for its prediction. In: Hall J, Harrison DE, Stammer D (eds) Proc OceanObs09: sustained ocean observations and information for society, Vol 2. Publication WPP-306, ESA, doi:10.5270/OceanObs09.cwp.53

Lean J, Rind D (2009) How will Earth's surface temperature change in future decades? Geophys Res Lett 36:L15708, doi:10.1029/2009GL038932

Liniger MA, Mathis H, Appenzeller C, Doblas-Reyes FJ (2007) Realistic greenhouse gas forcing and seasonal forecasts. Geophys Res Lett 34:L04705, doi:10.1029/2006GL028335

Luo JJ, Masson S, Behera SK, Yamagata T (2008) Extended ENSO predictions using a fully coupled ocean-atmosphere model. J Clim 21:84–93

Mason SJ (2011) Seasonal and longer-range forecasts. In: Jolliffe IT, Stephenson DB (eds) Forecast verification: a practioner's guide in atmospheric science, 2nd edn. Wiley, Chichester, p 203–220

Mason SJ, Baddour O (2008) Statistical modeling. In: Troccoli A (ed) Seasonal climate variability: forecasting and managing risk. Springer Academic Publishers, New York, NY, p 167–206

Masson D, Knutti R (2011) Spatial-scale dependence of climate model performance in the CMIP3 ensemble. J Clim 24:2680–2692

Matei D, Pohlmann H, Jungclaus J, Müller W, Haak H, Marotzke J (2012) Two tales of initializing decadal climate prediction experiments with the ECHAM/MPI-OM model. J Clim 25:8502–8523

Meehl G, Goddard L, Murphy J, Stouffer R and others (2009) Decadal prediction: can it be skillful? Bull Am Meteorol Soc 90:1467–1485

Mochizuki T, Ishii M, Kimoto M, Chikamoto Y and others (2010) Pacific decadal oscillation hindcasts relevant to near-term climate prediction. Proc Natl Acad Sci USA 107:1833–1837

Murphy J, Kattsov V, Keenlyside N, Kimoto M and others (2010) Towards prediction of decadal climate variability and change. Procedia Environ Sci 1:287–304

Pohlmann H, Jungclaus J, Kohl A, Stammer D, Marotzke J (2009) Initializing decadal climate predictions with the GECCO Oceanic Synthesis: effects on the North Atlantic. J Clim 22:3926–3938

Scaife AA, Copsey D, Gordon C, Harris C and others (2011) Improved Atlantic winter blocking in a climate model. Geophys Res Lett 38:L23703, doi:10.1029/2011GL049573

Smith DM, Cusack S, Colman A, Folland C, Harris G, Mur-

phy J (2007) Improved surface temperature prediction for the coming decade from a global climate model. Science 317:796–799

Smith DM, Eade R, Dunstone N, Fereday D, Murphy J, Pohlmann H, Scaife A (2010) Skilful multi-year predictions of Atlantic hurricane frequency. Nat Geosci 3: 846–849

Solomon A, Goddard L, Kumar A, Carton J and others (2011) Distinguishing the roles of natural and anthropogenically forced decadal forced climate variability: implications for prediction. Bull Am Meteorol Soc 92:141–156

Stockdale TN (1997) Coupled ocean–atmosphere forecasts in the presence of climate drift. Mon Weather Rev 125: 809–818

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 92:485–498

Uppala S, Kallberg P, Simmons A, Andrae U and others (2005) The ERA-40 re-analysis. Q J R Meteorol Soc 131: 2961–3012

van der Linden P, Mitchell J (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, Exeter

van Oldenborgh GJ, Doblas-Reyes FJ, Wouters B, Hazeleger W (2012) Decadal prediction skill in a multi-model ensemble. Clim Dyn 38:1263–1280

Vera C, Barange M, Dube O, Goddard L and others (2010) Needs assessment for climate information on decadal timescales and longer. Procedia Environ Sci 1:275–286

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge

Weigel AP (2011) Ensemble forecasts. In: Jolliffe IT, Stephenson DB (eds) Forecast verification: a practioner's guide in atmospheric science 2nd edn. Wiley, Chichester, p 141–166

Weigel AP, Bowler N (2009) Comment on 'Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?'. Q J R Meteorol Soc 135: 535–539

Weigel AP, Liniger MA, Appenzeller C (2007) The discrete Brier and ranked probability skill scores. Mon Weather Rev 135:118–124

Weigel AP, Liniger MA, Appenzeller C (2008a) Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts? Q J R Meteorol Soc 134:241–260

Weigel AP, Baggenstos D, Liniger MA, Vitart F, Appenzeller C (2008b) Probabilistic verification of monthly temperature forecasts. Mon Weather Rev 136:5162–5182

Weigel AP, Liniger MA, Appenzeller C (2009) Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? Mon Weather Rev 137:1460–1479

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. International Geophysics Series, Vol. 91. Academic Press, London

Wolter K, Timlin MS (2011) El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). Int J Climatol 31: 1074–1087

Yeager S, Karspeck A, Danabasoglu G, Tribbia J, Teng H (2012) A decadal prediction case study: late twentieth-century North Atlantic Ocean heat content. J Clim 25: 5173–5189

Zhang S (2011) Impact of observation-optimized model parameters on decadal predictions: simulation with a simple pycnocline prediction model. Geophys Res Lett 38:L02702, doi:10.1029/2010GL046133

Editorial responsibility: Filippo Giorgi, Trieste, Italy

Submitted: November 17, 2011; Accepted: September 7, 2012
Proofs received from author(s): December 11, 2012