Clim Dyn (2013) 40:245–272
DOI 10.1007/s00382-012-1481-2

A verification framework for interannual-to-decadal predictions experiments

L. Goddard · A. Kumar · A. Solomon · D. Smith · G. Boer · P. Gonzalez · V. Kharin · W. Merryfield · C. Deser · S. J. Mason · B. P. Kirtman · R. Msadek · R. Sutton · E. Hawkins · T. Fricker · G. Hegerl · C. A. T. Ferro · D. B. Stephenson · G. A. Meehl · T. Stockdale · R. Burgman · A. M. Greene · Y. Kushnir · M. Newman · J. Carton · I. Fukumori · T. Delworth

Received: 10 October 2011 / Accepted: 31 July 2012 / Published online: 24 August 2012

Abstract Decadal predictions have a high profile in the climate science community and beyond, yet very little is known about their skill. Nor is there any agreed protocol for estimating their skill. This paper proposes a sound and coordinated framework for verification of decadal hindcast experiments. The framework is illustrated for decadal hindcasts tailored to meet the requirements and specifications of CMIP5 (Coupled Model Intercomparison Project phase 5). The chosen metrics address key questions about the information content in initialized decadal hindcasts. These questions are: (1) Do the initial conditions in the hindcasts lead to more accurate predictions of the climate, compared to un-initialized climate change projections? and (2) Is the prediction model's ensemble spread an appropriate representation of forecast uncertainty on average? The first question is addressed through deterministic metrics that compare the initialized and uninitialized hindcasts. The second question is addressed through a probabilistic

L. Goddard (✉) · P. Gonzalez · S. J. Mason · A. M. Greene
International Research Institute for Climate and Society,
The Earth Institute of Columbia University, Palisades, NY, USA
e-mail: goddard@iri.columbia.edu

A. Kumar
Climate Prediction Center, National Centers for Environmental
Prediction, NOAA, Silver Spring, MD, USA

A. Solomon · M. Newman
Earth System Research Laboratory, NOAA,
University of Colorado, Boulder, CO, USA

D. Smith
UK Met Office, Hadley Centre, Exeter, UK

G. Boer · V. Kharin · W. Merryfield
Canadian Centre for Climate Modelling and Analysis,
Environment Canada, Victoria, BC, Canada

C. Deser · G. A. Meehl
National Center for Atmospheric Research, Boulder, CO, USA

B. P. Kirtman
Rosentiel School for Marine and Atmospheric Science,
University of Miami, Miami, FL, USA

R. Msadek · T. Delworth
NOAA's Geophysical Fluid Dynamics Laboratory,
Princeton, NJ, USA

R. Sutton · E. Hawkins
NCAS-Climate, Department of Meteorology,
University of Reading, Reading, UK

T. Fricker · C. A. T. Ferro · D. B. Stephenson
University of Exeter, Exeter, UK

G. Hegerl
University of Edinburgh, Edinburgh, UK

T. Stockdale
European Centre for Medium-Range Weather Forecasts,
Reading, UK

R. Burgman
Florida International University, Miami, FL, USA

Y. Kushnir
Lamont-Doherty Earth Observatory, The Earth Institute
of Columbia University, Palisades, NY, USA

J. Carton
University of Maryland, College Park, MD, USA

I. Fukumori
Jet Propulsion Laboratory, NASA, Pasadena, CA, USA

metric applied to the initialized hindcasts and comparing different ways to ascribe forecast uncertainty. Verification is advocated at smoothed regional scales that can illuminate broad areas of predictability, as well as at the grid scale, since many users of the decadal prediction experiments who feed the climate data into applications or decision models will use the data at grid scale, or downscale it to even higher resolution. An overall statement on skill of CMIP5 decadal hindcasts is not the aim of this paper. The results presented are only illustrative of the framework, which would enable such studies. However, broad conclusions that are beginning to emerge from the CMIP5 results include (1) Most predictability at the interannual-to-decadal scale, relative to climatological averages, comes from external forcing, particularly for temperature; (2) though moderate, additional skill is added by the initial conditions over what is imparted by external forcing alone; however, the impact of initialization may result in overall worse predictions in some regions than provided by uninitialized climate change projections; (3) limited hindcast records and the dearth of climate-quality observational data impede our ability to quantify expected skill as well as model biases; and (4) as is common to seasonal-to-interannual model predictions, the spread of the ensemble members is not necessarily a good representation of forecast uncertainty. The authors recommend that this framework be adopted to serve as a starting point to compare prediction quality across prediction systems. The framework can provide a baseline against which future improvements can be quantified. The framework also provides guidance on the use of these model predictions, which differ in fundamental ways from the climate change projections that much of the community has become familiar with, including adjustment of mean and conditional biases, and consideration of how to best approach forecast uncertainty.

Keywords Decadal · Prediction · Verification · Uncertainty · CMIP5

1 Context and motivation for a verification framework

Decadal prediction carries a number of scientific and societal implications. Information that could be provided by interannual-to-decadal predictions could advance planning towards climate change investment, adaptation, as well as the evaluation of those efforts (Vera et al. 2010; Goddard et al. 2012). In addition to the impact on people and economies, decadal-scale variability can also impact the perception or expectations of anthropogenic climate change. By using information on the initial state of the climate system in addition to the changes due to atmospheric composition the goal of decadal climate predictions is to capture the natural low-frequency climate variability that evolves in combination with climate change. Unlike climate change projections, however, the science of decadal climate predictions is new and is considered experimental.

The decadal prediction experiments that are contributing to CMIP5 (the Coupled Model Intercomparison Project phase 5), which is the suite of model experiments designed to advance climate science research and to inform the IPCC (Intergovernmental Panel on Climate Change) assessment process, use dynamical models of the coupled ocean–atmosphere–ice-land system, and possibly additional components of the Earth system. The informed use of such predictions requires an assessment of prediction quality and a sound theoretical understanding of processes and phenomena that may contribute to predictability on this time-scale. The process of building confidence in decadal predictions will ultimately involve both verification, which examines whether the prediction system can capture some predictable element of the climate system, and model validation, which examines whether the model representation of that element of the climate system is physically sound. Both topics are currently receiving attention from the climate research community.

Coordinated verification, or the assessment of the skill of a prediction system using a common framework, serves a number of purposes:

- comparison of the performance of prediction systems across modeling centers;
- evaluation of successive generations of the same prediction system and documenting improvements with time;
- use in multi-model ensemble techniques that depend on the model performance in predicting past history;
- feedback to the modelers on model biases;
- guidance on appropriate use of prediction or forecast information; and
- use in managing user expectations regarding the utility of the prediction information.

The skill of a prediction system is generally assessed from a set of hindcasts (i.e. predictions of the past cases), and their comparison with the observational counterpart. Skill assessment of the quality of information across different prediction systems, however, requires a level of standardization of observational datasets for validation, verification metrics, hindcast period, ensemble size, spatial and/or temporal smoothing, and even graphical representation. Such efforts have been initiated for seasonal predictions, for example, development of a standardized verification system for long-range forecasts (SVSLRF) within the purview of the World Meteorological Organization (WMO) (Graham et al. 2011).

A verification framework for decadal predictions is proposed here that employs a common, minimal set of metrics to assess prediction quality at the interannual-to-decadal timescale. This framework is intended to facilitate a sound and coordinated verification effort, results of which can be made available through a central website and can provide information to prediction and modeling centers on the relative performance of their system, and also to users of the data. A coordinated verification effort will also serve as a starting point for a wider set of prediction verification and model validation efforts that undoubtedly will be conducted on the initialized decadal prediction experiments, such as those that are part of the protocol for CMIP5.

The elements of the verification framework, which was developed by the US CLIVAR Working Group on Decadal Predictability and collaborators, builds on lessons learned from verification at the seasonal-to-interannual timescale. For seasonal-to-interannual predictions, skill mainly derives from anomalies in the ocean state, which are represented in the initial conditions used for the model prediction. On the other hand, the predictability on longer timescales is more intimately connected to anthropogenic climate change. Thus, a main issue addressed by the verification framework is the extent to which the initialized decadal prediction experiments provide better quality climate information than the climate change projections from the same models.

1.1 Feasibility of initialized decadal prediction

Research on decadal predictability began in the late 1990s with "perfect model predictability studies" (Griffies and Bryan 1997; Collins et al. 2006) and focused on the study of predictability for decadal-scale natural variations in the Atlantic meridional overturning circulation (AMOC). In the typical experimental design of such studies a coupled model is initialized with an ocean state taken from a particular time in a long model integration, together with a suite of atmospheric initial states taken from the same control simulation but from different times representing initial perturbations. The goal is for the perturbed ensemble members to 'predict' the evolution of the control run. The initial studies, and those to follow, suggested predictability of the AMOC to 10 or 20 years, and also indicated that some extreme states may be more predictable (Griffies and Bryan 1997; Collins et al. 2006; Pohlmann et al. 2004; Msadek et al. 2010; Teng et al. 2011). Other "diagnostic" studies (e.g. Boer 2004) relied on analysis of variance in long model integrations.

Prior to the development of decadal hindcasts, such entirely model-based predictability studies provided the only available estimates of predictability on decadal and longer time scales as there are no reliable methods applicable to the short instrumental record of observations. Perfect model and diagnostic studies have been used to identify regions where potential predictability is located and to identify which variables might be the more predictable on decadal time scales (e.g. Collins et al. 2006; Boer and Lambert 2008; Branstator and Teng 2010; Msadek et al. 2010). These have also been used to identify when the response to external forcing will emerge above the "climate noise" (e.g. Branstator and Teng 2010; Teng et al. 2011). Such approaches, however, only provide model-based estimates of decadal predictability, and may overestimate the skill in the setting of real predictions for the same model. There is no substitute for confronting a model with assimilated observations, and making actual predictions, for estimating the skill.

Recent studies have shown some similarity between model simulated decadal variability and observed variability, which further suggests feasibility of providing useable decadal-scale predictions, at least tied to Atlantic variability. Dynamical models of the climate system do produce multi-decadal scale fluctuations in the strength of their AMOC, and an associated SST pattern that closely resembles the observed pattern of SSTs, as well as other patterns of terrestrial climate during positive AMV conditions (Knight et al. 2005). When prescribed to an atmospheric model, the heat fluxes associated with positive AMV conditions produce many of the teleconnections that have been empirically identified from the observations (Zhang and Delworth 2006), such as wetter conditions over India and the Sahel during their monsoon season (Giannini et al. 2003) and increased Atlantic hurricane activity (Goldenberg et al. 2001). These results suggest that if the changes in the AMOC could be predicted, then the associated SSTs might be predictable, which in turn could lead to predictability of the associated terrestrial teleconnections. Several practical issues such as model biases, however, may be limiting factors and thus only through experimental predictions of past decades can we begin to assess how much predictability might be realized in the current state-of-the-art forecast systems.

In addition to taking advantage of the predictability of naturally occurring decadal variability suggested by the above evidence, initialized decadal predictions are likely to better maintain the near-term (i.e. next couple years to next couple decades) evolution of the forced response by better initializing the climate change commitment imparted to the ocean by increased greenhouse gasses. For example, in the pioneering decadal prediction paper of Smith et al. (2007), most of the improvement in prediction quality in the initialized hindcasts compared to the uninitialized climate change projections was found in regions such as the Indian Ocean and western Pacific, where the twentieth century temperature changes are likely dominated by radiative forcing from increasing greenhouse gasses (Ting et al. 2009; Du and Xie 2008).

1.2 The need for verification

Real-time predictions and forecasts,[1] for any timescale, must be accompanied with estimates of their skill to guide appropriate and justifiable use. For predictions of seasonal climate anomalies, for example, skill estimates are obtained from a set of hindcasts for which the observations are also available (Barnston et al. 2010; Wang et al. 2010). The use of the forecast information without reference to the skill of similar forecasts for past cases invites undue confidence in the forecast information, whereas information about how skillful past predictions or forecasts have been provides historical context that can assist users to incorporate real-time forecasts into their decision making process.

Although highly desirable, obtaining reliable estimates of skill for decadal predictions faces many challenges—one of the most formidable of these is the relatively short length of the hindcast period. The statistical significance of skill depends critically on the length of the verification time-series with progressively more robust estimates requiring verifications over longer and longer period (Kumar 2009). The required verification period also becomes longer as the time-scale of the phenomena of interest becomes longer. As the premise for skillful decadal predictions requires accurate specification of the initial state for the slowly varying components of the climate system, for example oceans and sea-ice, and because of a lack of observational systems extending back in time, a long history of decadal hindcasts is generally not feasible. Consequently, decadal hindcasts that are based on a short history skill estimates will be affected by sampling issues.

An additional issue that is related to estimates of past skill, and their applicability in the context of real-time decadal predictions, is the conditional nature of the prediction skill. In the context of seasonal prediction, it is well known that skill of seasonal predictions depends on the El Niño-Southern Oscillation (ENSO) variability, with skill being higher during large amplitude ENSO events (e.g. Goddard and Dilley 2005). Conditional skill has also been seen in decadal predictability as a function of initial state types (Collins et al. 2006; Branstator and Teng 2010). However, reliable estimates of skill conditional on specific circumstances are even harder to determine due to a smaller sample for verification.

2 A metrics framework to assess initialized decadal predictions

2.1 Data and hindcast procedures

A key element to the proposed verification framework is consistency in the use of observational data sets, even if more than one dataset is chosen for a particular variable. This includes consistency in the verification period, such as the span of years of the hindcasts and the specific initial condition dates within that span. Another important element is the treatment of the hindcast data, such as how bias adjustment is done (e.g. Meehl et al. 2012). As part of the verification, the model data may need to be re-gridded to the observational grids, and further spatial smoothing may also be applied to both model and observed fields. These parts of the verification framework are discussed below.

2.1.1 Observational data for hindcast verification

We advocate using a uniform and relatively small set of observational datasets for the assessment of decadal climate predictions and simulations. Observational data carry uncertainty, and the quantitative verification from a particular prediction system will vary when measured against different observational datasets. A conclusion that one prediction system is superior to another should therefore be based on the verification against more than one observational analysis. For the illustrative purposes of demonstrating the verification framework, we focus on air temperature and precipitation. The following data sets are chosen for hindcast verification:

1. Air temperature: Hadley Centre/Climate Research Unit Temperature version 3 variance-adjusted (HadCRUT3v; available for the period 1850–2011 on a 5° longitude by 5° latitude grid) (Brohan et al. 2006). Preference is given to the HadCRUT3v data because missing data is indicated as such. This can make verification more technically difficult. However, it also provides a more realistic view of where hindcasts can be verified with gridded data, and the resulting skill estimates are more trustworthy.
2. Precipitation: Global Precipitation Climatology Centre version 4 (GPCCv4; Schneider et al. 2008; Rudolf

[1] Our use of the terms "predictions" and "forecasts" follows the NRC report on intraseasonal-to-interannual predictability (NRC 2011) in which predictions are the outputs of models, and forecasts are the final product disseminated with the intention to inform decisions, and which are based on one or more prediction inputs that may include additional processing based on past performance relative to the observations (e.g. Robertson et al. 2004; Stephenson et al. 2005). These terms are both distinct from "projection", which is future climate change information from a model or models and depend mostly on the repones to external forcing, such as increasing greenhouse gasses.

> et al. 2010). This dataset covers the period 1901–2007, at a resolution of 2.5° longitude by 2.5° latitude grid, although the data is provided also at higher resolutions. We note though that GPCC does periodically update their dataset, and currently version 5 is available that extends through 2009. GPCC also provides a monitoring product that is typically available within 2 months or so of realtime. For studies requiring global coverage, the global precipitation climatology project version 2 (GPCPv2) available on a 2.5° longitude by 2.5° latitude grid) incorporates satellite measurements to provide additional coverage over the oceans.

The model hindcast data are first interpolated to the resolution of the observations prior to the calculation of verification metrics. Thus, the "grid-scale" analysis shown in the results is at a resolution of 5° × 5° for temperature and 2.5° × 2.5° for precipitation.

In the verification examples provided here, SST is not considered separately from air temperature. However, for SST specific verification, we suggest the use of either the Hadley Centre sea ice and SST version 1 (HadISST1; available on a 1° longitude by 1° latitude grid) (Rayner et al. 2003) or the National Oceanic and Atmospheric Administration Extended Reconstructed SST version 3 (ERSSTv3b; available on a 2° longitude by 2° latitude grid) (Smith et al. 2008). Both datasets are based on a blend of satellite and in situ data for the period since 1982, and employ quality-control and statistical procedures to construct globally-complete fields for each month (see references above for more detail).

All data described above can be accessed through the IRI Data Library in addition to their source institutions. The Decadal Verification web page contains links to downloading these data (http://clivar-dpwg.iri.columbia.edu, follow the Observational Dataset link under the Sample Code tab).

2.1.2 Model data used in this assessment

The ability to replicate observed climate variability depends on the prediction system, which includes the model as well as the data assimilation system used to initialize it. To illustrate the differences in skill that can arise between different prediction systems, the results for two different hindcast prediction experiments are presented in this paper. The first is the perturbed physics hindcasts from Hadley Centre using an updated version of the DePreSys prediction system (Smith et al. 2010). The second is the set of hindcasts from the Canadian Climate Centre using CanCM4 (Merryfield et al. 2011). These models are just two of those participating in the CMIP5 decadal experiment suite, although the Hadley Centre is using a slightly different experimental set-up for CMIP5. The assessment of these two models serves as an illustrative example of the verification framework, and allows for interpretive discussion of the metrics. Further, use of a minimum of two models illustrates differences in skill that can occur across different models. Additional contributions to the coordinated verification by other modeling centers, which is already occurring, will enable more informed use of the CMIP5 experiments.

The Met Office Decadal Prediction System (DePreSys, Smith et al. 2007) is based on the third Hadley Centre coupled global climate model (HadCM3, Gordon et al. 2000) with a horizontal resolution of 2.5° × 3.75° in the atmosphere and 1.25° in the ocean. The hindcasts assessed here (not the same as those for CMIP5), are from an updated version of DePreSys (Smith et al. 2010) that employs an ensemble of nine variants of the model, sampling parameterization uncertainties through perturbations to poorly constrained atmospheric and surface parameters. HadCM3 was also updated to include a fully interactive representation of the sulphur cycle, and flux adjustments to restrict the development of regional biases in sea surface temperature and salinity (Collins et al. 2010). Initial conditions for hindcasts for each variant were created by relaxation to the atmospheric (ERA-40 and ECMWF) and ocean (Smith and Murphy 2007) analyses. In this, observed values were assimilated as anomalies in order to minimize model initialization shock after the assimilation is switched off. The hindcasts consist of nine-member ensembles (one for each model variant) starting from the first of November in every year from 1960 to 2005, and extend 10 years from each start time.

CCCma decadal predictions are based on the CanCM4 climate model, which is similar to the CanESM2 earth-system model employed for the long-range projection component of CMIP5 (Arora et al. 2011), except that the latter includes an interactive carbon cycle involving terrestrial and ocean ecosystem models. Atmospheric model resolution is approximately 2.8° × 2.8° with 35 levels, and ocean model resolution is approximately 0.94° × 1.4°, with 40 levels. Initial conditions for the 10-member hindcast ensemble were obtained from a set of assimilation runs, one for each ensemble member, begun from different initial conditions drawn from a multi-century spin-up run. Atmospheric and surface fields in these runs were constrained to remain close to the ECMWF (ERA 40 and ERA-Interim) atmospheric, HadISST 1.1 sea ice and NCEP (ERSST and OISST) sea surface temperature analyses beginning in 1958. Temperatures from the NCEP GODAS (1981 to present) or SODA (before 1981) ocean analyses were assimilated off-line using the a method similar to that of Tang et al. (2004), after which salinities were adjusted as in Troccoli et al. (2002). In contrast to DePreSys, all assimilation is based on full-field observed values rather than anomalies. The 10-year hindcasts were initialized at the beginning of each January from 1961 until present.

Different modeling centers have generated their hindcasts with varying start dates and ensemble sizes, as evidenced by the two hindcast sets described above. The coverage of these hindcast sets exceeds that of the initial CMIP5 experimental design (Taylor et al. 2012), which called for 10-year hindcasts started every 5 years beginning in late 1960/early 1961 with at least 3 ensemble members. For the sake of a unified comparison across different sets of hindcasts in CMIP5, the standard verification is restricted here to this initial CMIP5 experimental design with the exception that all available ensemble members are used. However, this framework could be configured to apply to any collection of predictions. It should be noted that based on preliminary verification studies, CMIP5 now recommends that hindcasts be initialized every year. An extension of the verification analysis will be applied to those more complete hindcast sets as they become available, and posted to http://clivar-dpwg.iri.columbia.edu.

2.1.3 Adjustment for mean bias of prediction systems

Because climate models are imperfect, there are systematic differences between model simulations and observations. Since some model biases can be as large as the signal one wants to predict, model biases must be accounted for in some way in order to create prediction anomalies, and to assess skill from a set of hindcasts. There are two main approaches for reducing mean, or climatological, biases of models in decadal climate predictions, which depend on the methodology used for initializing decadal predictions, i.e. the full field initialization or the anomaly initialization (see ICPO 2011).

In full field initialization, initial conditions of the predictions are created by constraint of model values to be close to the observed analysis. During the prediction period, the model will inevitably drift away from the specified observed initial state towards its preferred climatology. Based on a set of hindcasts, the drift can, in principle, be estimated as a function of lead-time and calendar month. This estimate can then be subtracted from the model output to yield bias-corrected predictions (Stockdale 1997; ICPO 2011).

In anomaly initialization models are initialized by adding observed anomalies to the model climatology (e.g. Pierce et al. 2004; Smith et al. 2007). Observed and model climatologies are computed for the same historical period, with model climatologies obtained from simulations that include anthropogenic and natural external forcing but do not assimilate observations. In this case the initialized model predictions only deviate from the model's preferred climatology within the bounds of random variability so that, in principle, there is no systematic drift in the predictions.

There are technical problems with both initialization approaches. For example, neither approach overcomes potential drifts due to an incorrect model response to anthropogenic or natural external forcings. Because such a bias is non-stationary, simple removal of the mean hindcast drift may not adequately correct for it. Furthermore, while anomaly initialization attempts to overcome drifts that are present in the long-term integrations of the climate change projections, it does not necessarily avoid initialization shocks (abrupt changes at the beginning of the forecast due to dynamical imbalance in various fields, for example, pressure gradients and ocean currents) or non-linear interaction between drift and evolution of the quantity being predicted. In addition, the observed anomalies might not be assimilated at optimal locations relative to features such as the Gulf Stream if these are offset in models compared to reality. Errors in estimating the model bias will directly contribute to errors in the prediction. Ideally therefore a large set of hindcasts, which samples different phases of the variability to be predicted, should be employed for bias adjustment in order to reduce sampling errors.

All decadal hindcasts used in this analysis have had their mean biases removed following the methods outlined in ICPO (2011).

2.1.4 Temporal and spatial averaging

A disconnect often exists between the predictable space and time scales of the climate information and the scales at which individuals wish to use it. Spatially, for example, common use of the information relies on grid-scale data, or further downscaling to even higher spatial resolution. However, local-scale variability that may be unrelated to larger-scale climate variability adds noise, and thus reduces the prediction skill. Similar logic applies to the temporal scale. Spatial smoothing has been used in most previous decadal prediction studies (Smith et al. 2007; Keenlyside et al. 2008) although the scale of the smoothing varies from study to study. The smoothing is beneficial in skill assessment due to reduction of the unpredictable grid-scale noise (Räisänen and Ylhäisi 2011).

We advocate verifying on at least two spatial scales: (1) the observational grid scales to which the model data is interpolated, and (2) smoothed or regional scales. The latter can be accomplished by suitable spatial-smoothing algorithms, such as simple averages or spectral filters. Given that precipitation derives from more localized processes, the recommended smoothing is over smaller scales than temperature. Although other criteria could be used, a balance between skill improvement and signal-to-noise retention suggests that 5° latitude × 5° longitude represents a reasonable scale for smoothing precipitation, and 10° latitude × 10° longitude for temperature (Räisänen and Ylhäisi 2011). At these scales, grid-scale noise is reduced while retaining the strength of the climate signal and increasing the skill of the verification. It should be noted that many of the observational datasets discussed above already contain some spatial smoothing.

The verification of the temporal information in this framework is provided at different scales: year 1, years 2–5, years 6–9, and years 2–9. This set of temporal smoothing choices may seem somewhat arbitrary, but it represents a small set of cases that can illustrate the quality of the information for different lead times and temporal averaging (e.g. Smith et al. 2007). As with spatial smoothing, temporal smoothing will typically reduce higher frequency noise and increase skill. The reason to show different averaging periods is that one may be tempted to look at skill for decadal averages from these hindcasts and assume that level of quality applies throughout the period. However, the 4-year average forecasts (years 2–5 and 6–9) within the decade are likely to have lower skill, and there are potential differences between those also. Thus these four cases are a minimum to show skill dependence on averaging and lead time. The first year of the prediction represents overlap with currently available seasonal-to-interannual predictions, and should be most predictable owing to its proximity to the observed initial conditions. The year 2–5 average still represents the interannual timescale, but it discards the initial year for which the imprint of initial conditions is strong; it is likely still dominated by year-to-year variability and less by the climate change signal. The year 2–9 average represents decadal-scale climate and excludes the relatively large contribution to skill from the first year of the prediction. This approximately decadal period is the common time horizon of the CMIP5 decadal prediction experiments.[2] The 6–9 year average predictions are also verified, and the skill is compared with the skill of 2–5 year average predictions to understand dependence of skill on lead time.

Following ICPO (2011) report on mean bias adjustment, through changing the variable names to better reflect what they stand for, the initialized hindcasts are represented by $H_{ij\tau}$, where $i = 1, n$ is the set of ensemble members, run at each initial time $j = 1, n$, and extending over a prediction range of $\tau = 1, m$. In the nominal experimental design of CMIP5 (Taylor et al. 2012) the start dates of the prediction experiments are every 5 years, from late 1960/beginning 1961 to 2005/2006, yielding n = 10 hindcasts, and each of these hindcasts predicts 10 years out (i.e. m = 10) from the initial date. The ensemble mean prediction averages over the ensemble members, $H_{j\tau} = \frac{1}{ne}\sum_{i=1}^{ne} H_{ij\tau}$. Similarly, the temporal average between the initial year, YRi, and the final year, YRf, of a particular hindcast is achieved by summing over those years in the prediction range, $H_j = \frac{1}{(YRf-YRi+1)}\sum_{\tau=YRi}^{YRf} H_{j\tau}$. An example of a prediction of temperature and precipitation anomalies from initial conditions at the end of 1995 for lead time years 2–9 (i.e. 1997–2004) compared to observations for the same period is shown in Fig. 1, with the spatial smoothing discussed above.

2.2 Assessing the quality of prediction experiments

Verification metrics are chosen to answer specific questions regarding the quality of the prediction information. The metrics can identify where errors or biases exist in the predictions to guide more effective use of them. The proposed questions address the accuracy in the prediction information (Q1) and the representativeness of the prediction ensembles to indicate uncertainty (Q2). Specifically, the questions are:

- **Q1**: Do the initial conditions in the hindcasts lead to more accurate predictions of the climate? If so, on what time scales?
- **Q2**: Is the model's ensemble spread an appropriate representation of prediction uncertainty on average?

2.2.1 Deterministic metrics

The question of whether the initialization provides greater accuracy in the predictions can be addressed using deterministic metrics. The primary deterministic verification metric chosen for the framework is the mean squared skill score (MSSS). The MSSS is based on the mean squared error (MSE) between a set of paired predictions (or hindcasts),[3] Hj, and observations, Oj, over $j = 1, n$ years or start dates, following the formulation (though not exact notation) of Murphy (1988). Here, the ensemble mean prediction and the corresponding observation are given for a specific target lead time, or average of lead times, as anomalies relative to their respective climatologies (which is equivalent to removal of mean bias). MSE is given by

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(H_j - O_j\right)^2 \tag{1}$$

since $\mathrm{H_j}$ and $\mathrm{O_j}$ in (1) are anomalies, the MSE as written represents only the error variance but does not include the bias error component. The MSSS represents the MSE, or

[2] Some of the decadal prediction experiments extend to a full 10 calendar years after the start date, but not all. For example, one started in Nov 1960, might only extend to Oct 1969.

[3] In this section hindcasts are equivalent to the set of predictions, following Murphy (1988).

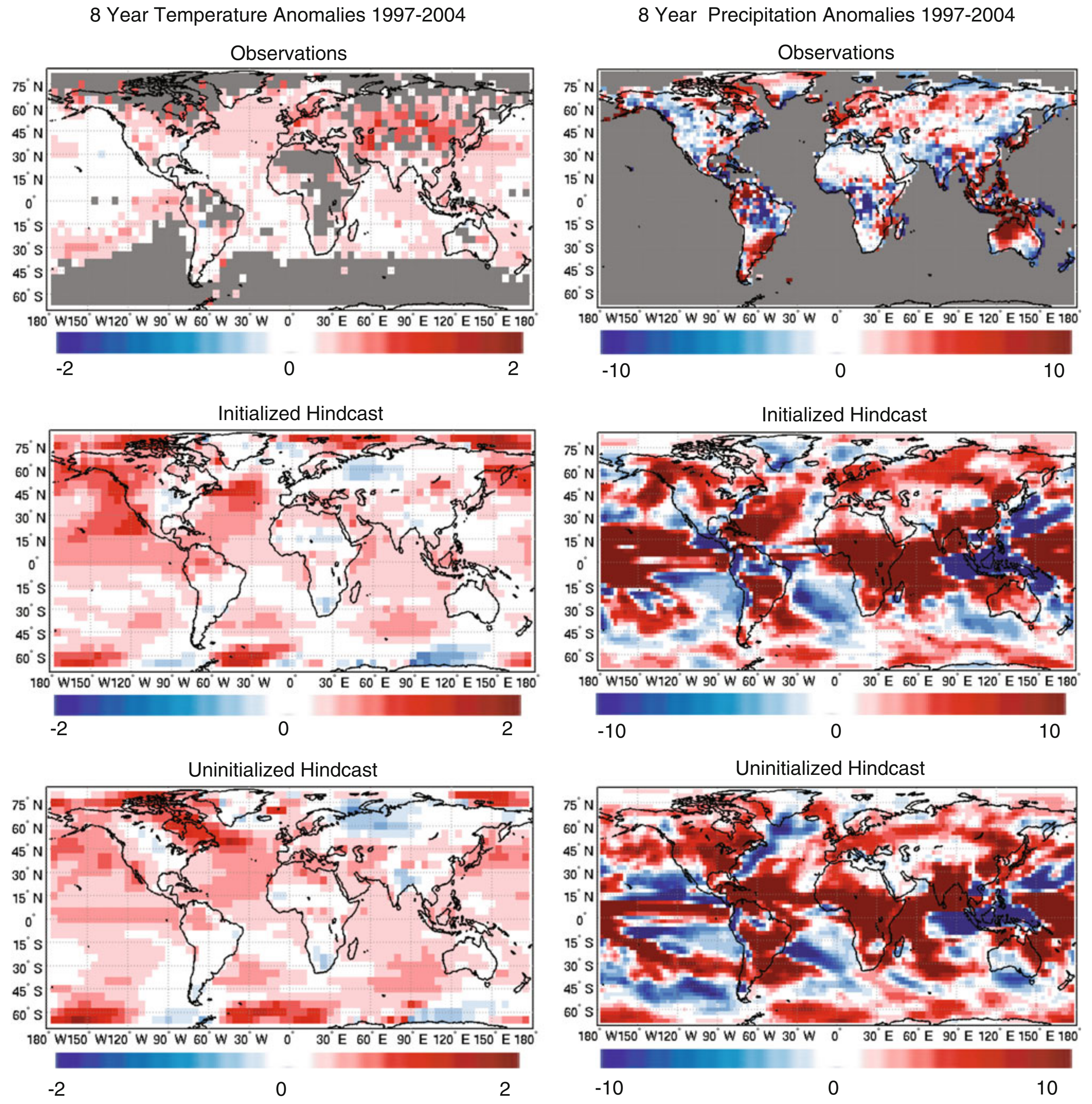

8 Year Temperature Anomalies 1997-2004
8 Year Precipitation Anomalies 1997-2004
Observations
Initialized Hindcast
Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180° W150° W120° W 90° W 60° W 30° W 0° 30° E 60° E 90° E 120° E 150° E 180°
-2
0
2
-10
0
10

Fig. 1 Eight year averages (1997–2004) for temperature smoothed over 10° × 10° boxes (*left*) and precipitation smoothed over 5° × 5° boxes (*right*). *Top row* observations based on HadCRUT3v temperature anomalies and GPCC precipitation anomalies; *middle row* initialized hindcasts from DePreSys starting with assimilated observations in November 1995; *bottom row* uninitialized hindcasts from DePreSys

accuracy skill, of a "test prediction" against some reference prediction, such as the climatological average, $\bar{O} = \frac{1}{n}\sum_{j=1}^{n} O_j$, or a zero anomaly forecast. The MSSS is defined as

$$MSSS(H, \bar{O}, O) = 1 - \frac{MSE_H}{MSE_{\bar{O}}} \quad (2)$$

Thus the MSSS is a function of the prediction one wants to evaluate, the reference prediction, and the observations. If the reference prediction is the climatological average taken over the same period as the hindcasts to be assessed, then the MSSS can be expanded as:

$$MSSS(H, \bar{O}, O) = r_{HO}^2 - [r_{HO} - (s_H/s_O)]^2 \quad (3)$$

where r_{HO} is the correlation coefficient between the observed and hindcast time series, and $s_H^2 = \left(1/_n\right) \sum_{j=1}^{n} \left(H_j\right)^2$ and $s_O^2 = \left(1/_n\right) \sum_{j=1}^{n} \left(O_j\right)^2$ are the population variances of the hindcasts and the observations, respectively (Murphy 1988). The MSSS is a summary metric; it combines: (1) the square of the correlation coefficient (first term on right hand side of Eq. 2), and (2) the square of the conditional prediction bias (second term on right hand side of Eq. 3).

The correlation coefficient is a scale-invariant measure of the linear association between the predicted mean and the observations (e.g. Murphy 1988). As such, it gives a measure of potential skill. The biases inherent in the forecasts affect the translation between the predicted value and the observed value and thus the MSE (or MSSS). If the forecast contained no conditional biases the $MSSS(H, \bar{O}, O)$ would be determined by the correlation coefficient alone.

The correlation coefficient is a measure of relative association; relative magnitude of the time series is not considered. The conditional bias does consider the magnitude, or expected value, of the observation given the prediction. As an example, consider the case where the climate evolves as a simple linear trend in temperature. A mean bias, with no conditional bias, would refer to a mere offset of that trend, but the rate of change over time in the predictions would match the observations (Fig. 2a). Alternatively, conditional bias with no mean bias occurs when the predicted and observed temporal means are approximately the same, but the rate of change is different (Fig. 2b). Initialization occurring at points within the timeline, will bring the prediction closer to the observation. This in itself may reduce conditional bias. As the prediction evolves, however, its response to increasing greenhouse gasses may tend toward that of the uninitialized hindcasts over the course of an adjustment period during which the influence of the initial conditions is "forgotten" (Kharin et al. 2012). For both sets of hindcasts the correlation with the observations would be 1.0 at all lead times, but they are not accurate because the model has a bias in the magnitude of its response to the forcing. The regression between the initialized hindcasts and the observations is much closer to unity near the start of the forecast, than would be the case for uninitialized forecasts, even if they over respond to the forcing or subsequently drift away. Both correlation and conditional bias are important elements of the relative accuracy of a prediction system.

Care must be taken in interpreting the MSSS with a climatological reference prediction when there is a trend in the observations. In the context of this verification the climatology period is taken as the entire hindcast period (1961–2006). In the presence of a trend, the MSE of the climatological prediction increases with longer verification periods as more of the trend is sampled as part of the prediction. This phenomenon may be desirable in the case of uninitialized test predictions, as the ability to predict the trend is indeed an important aspect of skill. If on the other hand the test predictions are a series of initialized hindcasts at a fixed lead time, the inflation of the MSSS caused by the trend may be spurious. The reason is that the long-term trend will be part of the initial conditions, but there is no actual prediction of the trend except its initialization. For example, even predictions produced by simply persisting the initial conditions will achieve a high MSSS in the presence of a trend, given a long enough verification period.

The above problem of trend inflation of the MSSS is mitigated if, instead of using the MSSS to compare a single prediction system to a climatological reference, it is used to compare two competing prediction systems that both follow the trend. In the proposed verification framework the MSSS is used to address Q1, for which the "test" predictions are the initialized decadal hindcasts denoted as *Hj*, and the "reference" predictions are the uninitialized climate change projections, denoted as P_j. It may still be possible that an uninitialized prediction with a correct sign of trend, but wrong magnitude, has a higher MSE than initialized hindcasts that better captures the trend due to the initialization. As a consequence, the initialized hindcasts may appear more skillful in comparison even though the additional skill originates solely from the persistence of the initial conditions rather than from their subsequent evolution. An alternative representation that should be investigated, though not a part of this framework, would be to verify the incremental change from initialization time to target time.

The MSSS comparing the test (initialized) hindcasts, *H*, against the reference (uninitialized) projections, *P*, can be written as:

$$MSSS(H, P, O) = 1 - \frac{MSE_H}{MSE_P} \tag{4}$$

$$MSSS(H, P, O) = \frac{r_{HO}^2 - [r_{HO} - (s_H/s_O)]^2 - r_{PO}^2 + [r_{PO} - (s_P/s_O)]^2}{1 - r_{PO}^2 + [r_{PO} - (s_P/s_O)]^2} \tag{5}$$

$$MSSS(H, P, O) = \frac{MSSS_H - MSSS_P}{1 - MSSS_P} \tag{6}$$

A perfect MSSS of 1.0 would require $MSE_H = 0$ and $MSE_p \neq 0$. The MSSS represents the improvement in accuracy of the test predictions (or hindcasts) *H* over the reference predictions (or projections) *P*. While a positive MSSS suggests the test predictions are more accurate than

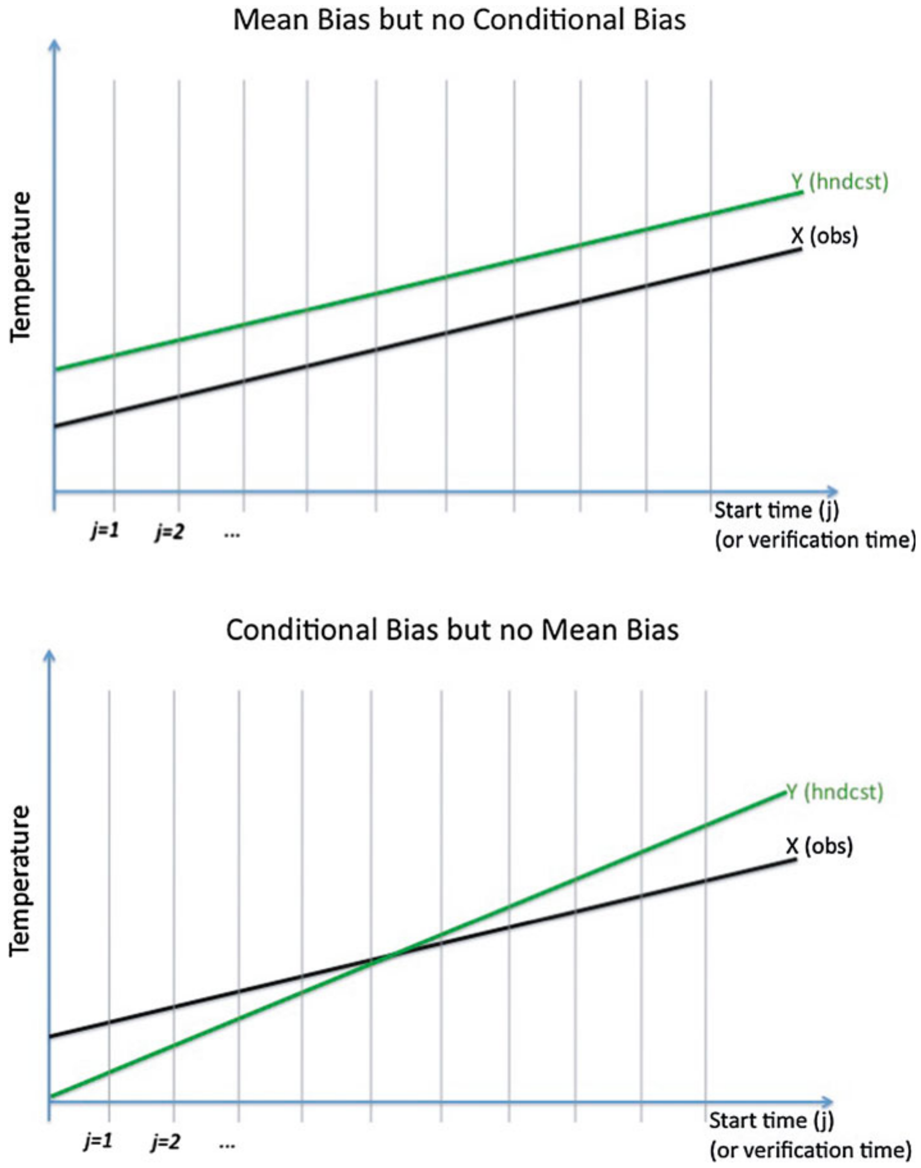


Fig. 2 Graphical illustration of the concept of mean and conditional bias bias between hindcasts and the observations. **a** Mean bias is positive, but no conditional bias exists because the magnitude of the trends are the same. **b** Mean bias is zero, but the conditional bias is negative because correlation is 1.0 and the variance of the hindcast is larger than that for the observations. This conditional bias would exist in a model that over-responded to increasing greenhouse gasses, for example. *Vertical grey bars* along the start time axis represent the start times of the CMIP5 decadal hindcast experiments and are spaced 5 years apart

the reference predictions and a negative MSSS suggests the opposite, the MSSS is not symmetric about zero, in that it does not satisfy MSSS(H, P,O) = −MSSS(P,O). Thus the absolute value of a negative MSSS does not have the same interpretation as a positive MSSS of the same magnitude.

The results presented in Figs. 3, 4, 5, 6, 7 and 8 are based on the spatially smoothed data that reduces grid scale noise (see Sect. 2.1.4). Maps of the MSSS for the DePreSys and CanCM4 decadal hindcasts are shown for temperature (Fig. 3) and for precipitation (Fig. 4). These are decadal-scale predictions that cover years 2–9, or equivalently a 1-year lead-time for a decadal-average prediction. If the MSSS is positive it indicates that the initialized hindcasts under test are more skillful than the corresponding reference hindcasts. The MSSS in the upper panels is for the initialized hindcasts with uninitialized hindcasts as the reference, and red areas indicate that the initialized hindcasts are more accurate than the uninitialized hindcasts, and blue areas denote areas where the initialized hindcasts are less accurate. For both prediction systems, the MSSS for temperature from the initialized predictions (middle row) and the uninitialized projections (bottom row) show positive values over much of the map, illustrating the point made above about the trend playing an important role in the MSSS when using a climatological reference prediction. Most of the places where the MSSS is worse (negative or blue areas in the figure) than the reference prediction of climatology (Fig. 3, middle and bottom row) are where the temperature trend has been weak or negative. Many of these regions of negative MSSS (referenced against climatology) are where the conditional bias is large; these are

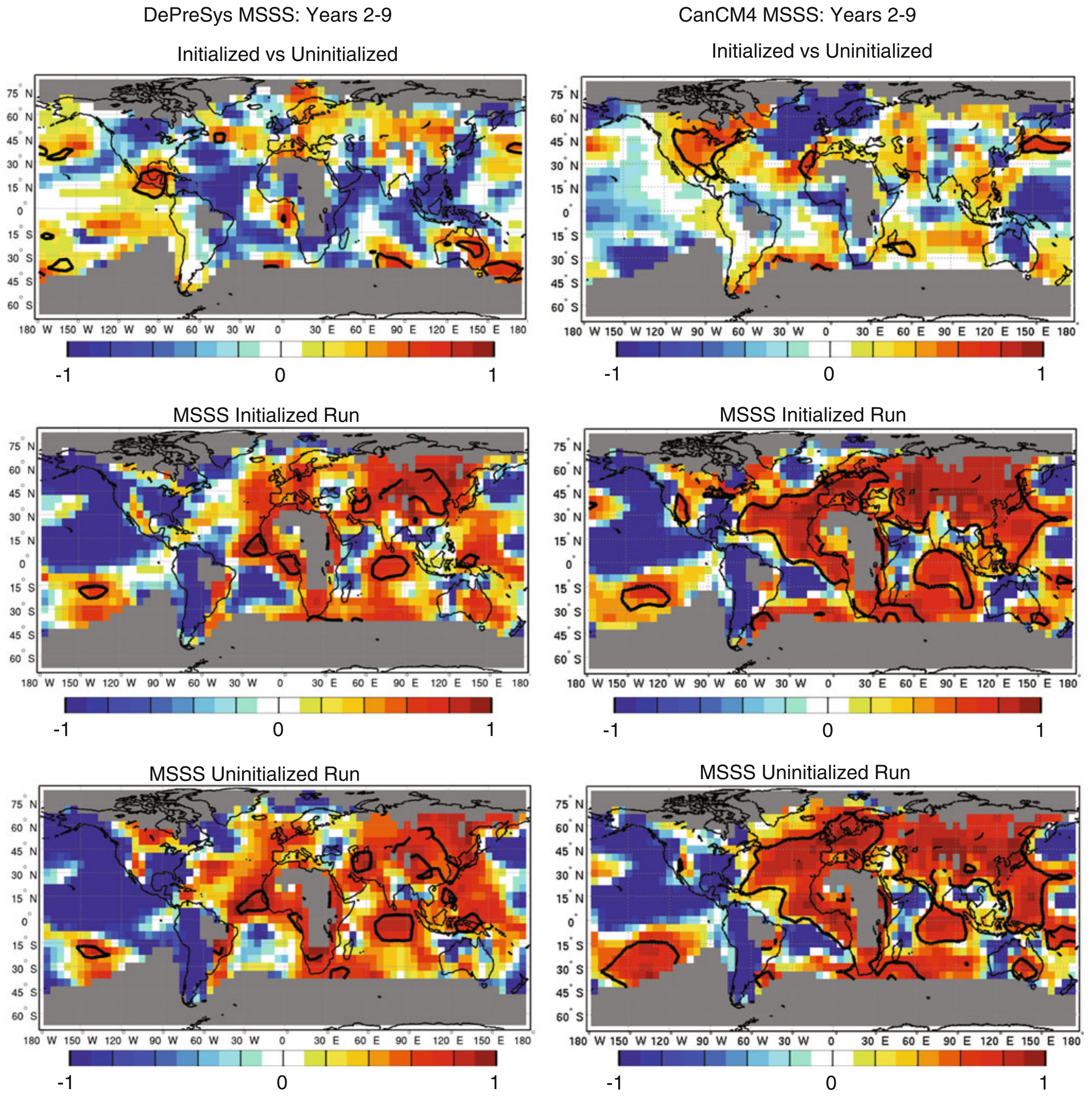


Fig. 3 Mean squared skill score (MSSS) for decadal temperature hindcasts from the DePreSys prediction system of the Hadley Centre (*left*) and the CanCM4 prediction system of the Canadian Climate Centre (*right*). *Top row* MSSS comparing the initialized hindcasts ("forecasts") and the uninitialized hindcasts ("reference") as predictions of the observed climate; *middle row* MSSS comparing the initialized hindcasts ("forecasts") and the climatological mean ("reference"); *bottom* MSSS between the uninitialized hindcasts ("forecasts") and the climatological mean ("reference"). Observed and model data has been smoothed as described in text. The forecast target is year 2–9 following the initialization every 5 years from 1961 to 2006 (i.e. 10 hindcasts). *Contour line* indicates statistical significance that the MSSS is positive at the 95 % confidence level

typically areas where the strength of the model response is too large compared to the observations for a given correlation.

The MSSS of the initialized hindcasts relative to the uninitialized hindcasts shows that areas of improved skill due to initialization differ between the two models (Figs. 3, 4, top panels, the positive or red areas). For example, the initialized DePreSys hindcasts for temperature improve over the uninitialized hindcasts in the North Atlantic, whereas in the CanCM4 temperature hindcasts the

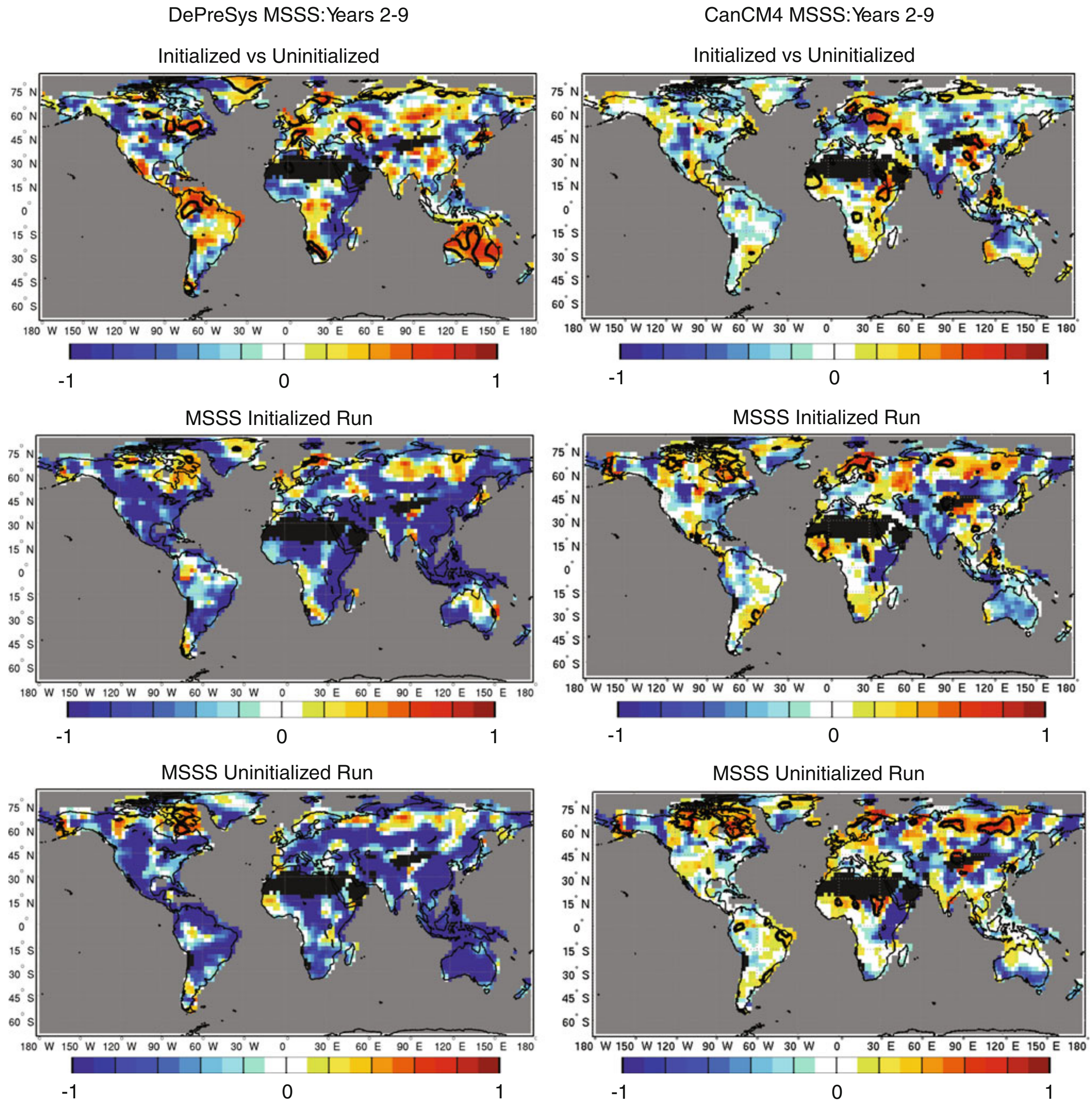


Fig. 4 Same as Fig. 3, but for precipitation hindcasts

improvement is seen in the tropical Atlantic. That different prediction systems differ in where they are estimated to be more skillful is a common situation in seasonal-to-interannual prediction, and has been the basic premise for multi-model seasonal prediction systems. It should also be noted that in the case of the Atlantic neither of the models' improved skill is deemed statistically significant (see "Appendix 2" for methodology), which is shown by the heavy contour line enclosing the positive skill areas. These differences in skill therefore may be due to sampling errors, given the limited number of cases in the CMIP5 experimental design.

The MSSS for the precipitation hindcasts (Fig. 4, middle and bottom row) are not significantly better than the reference prediction of climatology, anywhere. There are regions where the MSSS of the initialized hindcasts are estimated to be significantly better than the uninitialized ones, but these areas (indicated by significance contours)

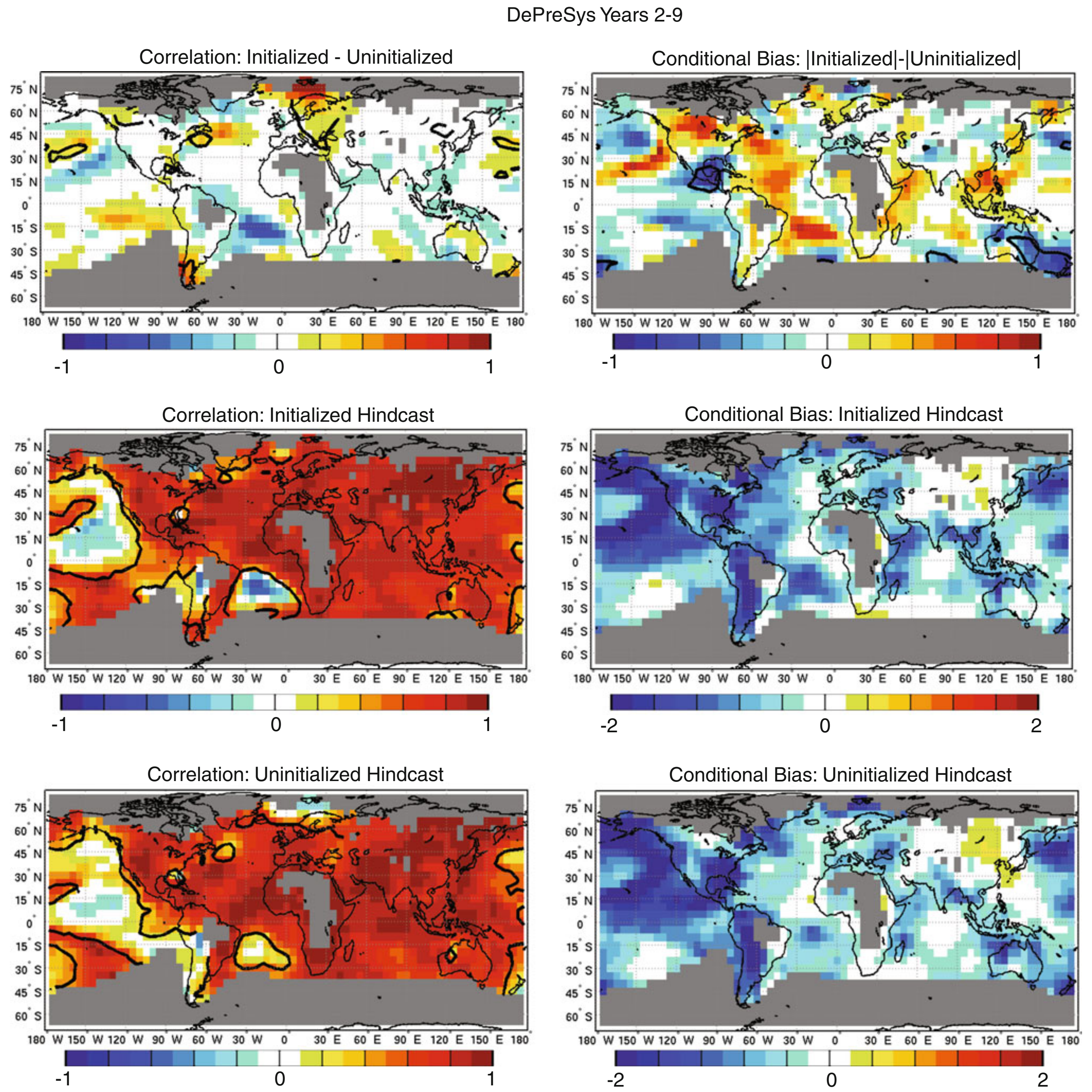

DePreSys Years 2-9
Correlation: Initialized - Uninitialized
Conditional Bias: |Initialized|-|Uninitialized|
Correlation: Initialized Hindcast
Conditional Bias: Initialized Hindcast
Correlation: Uninitialized Hindcast
Conditional Bias: Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1
-2
0
2

Fig. 5 Skill metrics related to MSSS decomposition for DePreSys temperature hindcasts. *Left* Anomaly correlation coefficients with top row depicting the difference between the correlation of the initialized hindcasts (*middle row*) and that of the uninitialized hindcasts (*bottom*). *Right* Conditional bias, with top row depicting the change in magnitude of conditional bias between the initialized hindcasts (*middle*) relative to that of the uninitialized hindcasts (*bottom*). Observed and model data has been smoothed as described in text. The forecast target is year 2–9 following the initialization every 5 years from 1961 to 2006 (i.e. 10 hindcasts). Contour line on the correlation maps indicates statistical significance that the value is positive at the 95 % confidence level

are small, and point-wise significance may still be related to the small sample size. Note that we only test the improvements for significance. Even in regions where improvement between the initialized and uninitialized hindcasts is seen (top panels), this improvement must be viewed together with the actual skill from the initialized hindcasts. For example, in the case of northern South America (Fig. 3, upper row left) the improvement occurs over a region where the actual accuracy of the initialized hindcasts is on par with climatology. For example, in the case of eastern Africa (Fig. 3, upper right) the region of improvement is one where the initialized hindcasts may be better than the unutilized ones but are still much worse than climatology.

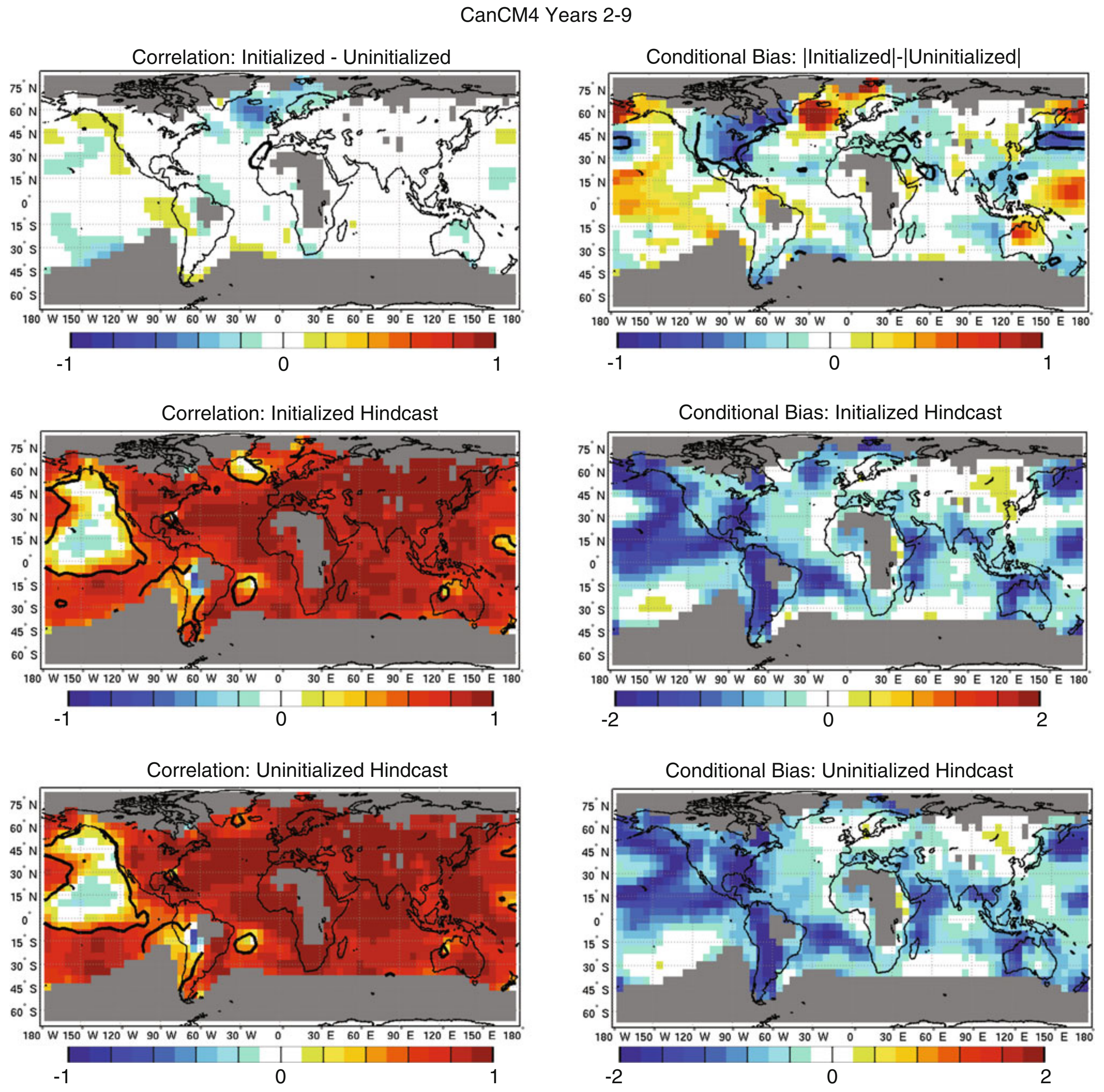

CanCM4 Years 2-9
Correlation: Initialized - Uninitialized
Conditional Bias: |Initialized|-|Uninitialized|
Correlation: Initialized Hindcast
Conditional Bias: Initialized Hindcast
Correlation: Uninitialized Hindcast
Conditional Bias: Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1
-2
2

Fig. 6 Same as Fig. 5, but for CanCM4 hindcasts

Given that both the correlation and conditional bias determines the MSSS, those deterministic metrics are presented as well (Figs. 5, 6, 7, 8). These are shown as metrics in their own right, not in the squared version in which they appear in the MSSS equation.

As expected from the MSSS maps, the correlation for temperature (Figs. 5, 6 left, middle and lower maps) is high from both the initialized and uninitialized predictions over most areas, with the notable exception of the ENSO-related tropical Pacific area where the year-to-year variability is large and trends to date are small. Improvements in temperature predictions due to initialization are most notable in the north Atlantic and north Pacific for the DePreSys hindcasts, but are of small spatial extent (Fig. 5, top left) and are effectively non-existent for the CanCM4 hindcasts (Fig. 6, top left). That the correlation differences are small suggests that for this forecast target (i.e. year 2–9 annual means)[4] there is little additional predictive skill derived from the initialization.

[4] Verification has been done for some seasonal means as well, but not included in this manuscript. See http://clivar-dpwg.iri.columbia.edu.

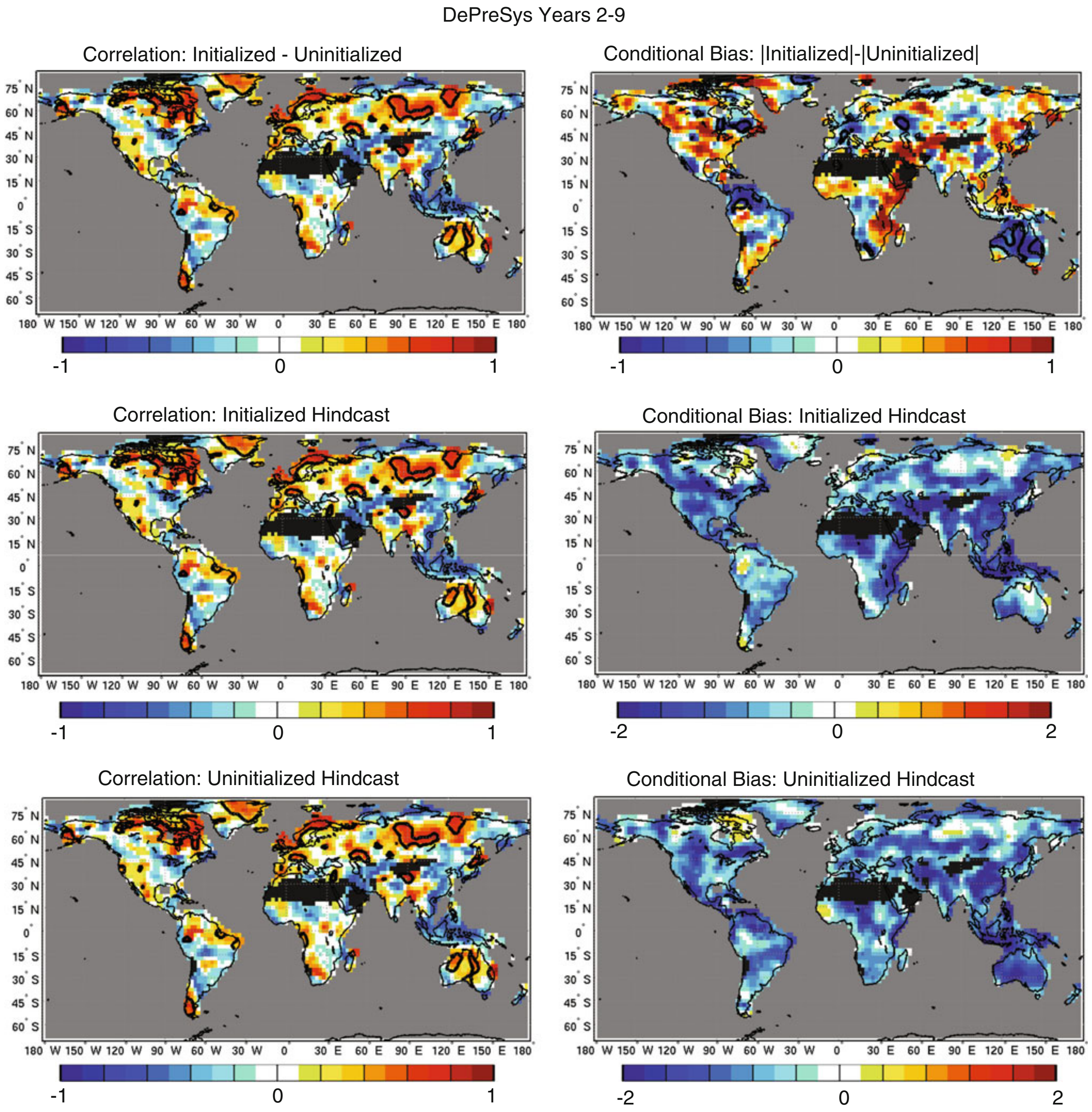

DePreSys Years 2-9
Correlation: Initialized - Uninitialized
Conditional Bias: |Initialized|-|Uninitialized|
Correlation: Initialized Hindcast
Conditional Bias: Initialized Hindcast
Correlation: Uninitialized Hindcast
Conditional Bias: Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1
-2
0
2

Fig. 7 Same as Fig. 5, but for precipitation hindcasts

For decadal-scale precipitation hindcasts (Figs. 7, 8, left) both sets of hindcasts from the two prediction systems show positive anomaly correlations over the high latitudes of the northern hemisphere. The CanCM4 hindcasts also show high correlations throughout much of the tropics. However, the improvement in correlation for precipitation due to initialization is meager at best. As was seen for the MSSS for precipitation, the areas of statistically significant improvement are small and, though point-wise significant, may still be the result of sampling issues. i.e. given that statistical significance is assessed in so many places, some areas will be assessed as locally statistically significant by chance (Livezey and Chen 1983; Power and Mysak 1992).

A negative conditional bias in temperature is seen in both sets of hindcasts (Figs. 5, 6 right, middle and lower panels) in regions where the correlation has been weak, and by implication the variance of the forecast is too large relative to the observations and the correlation coefficient. The conditional bias is typically, though not always, more negative when the anomaly correlations are small

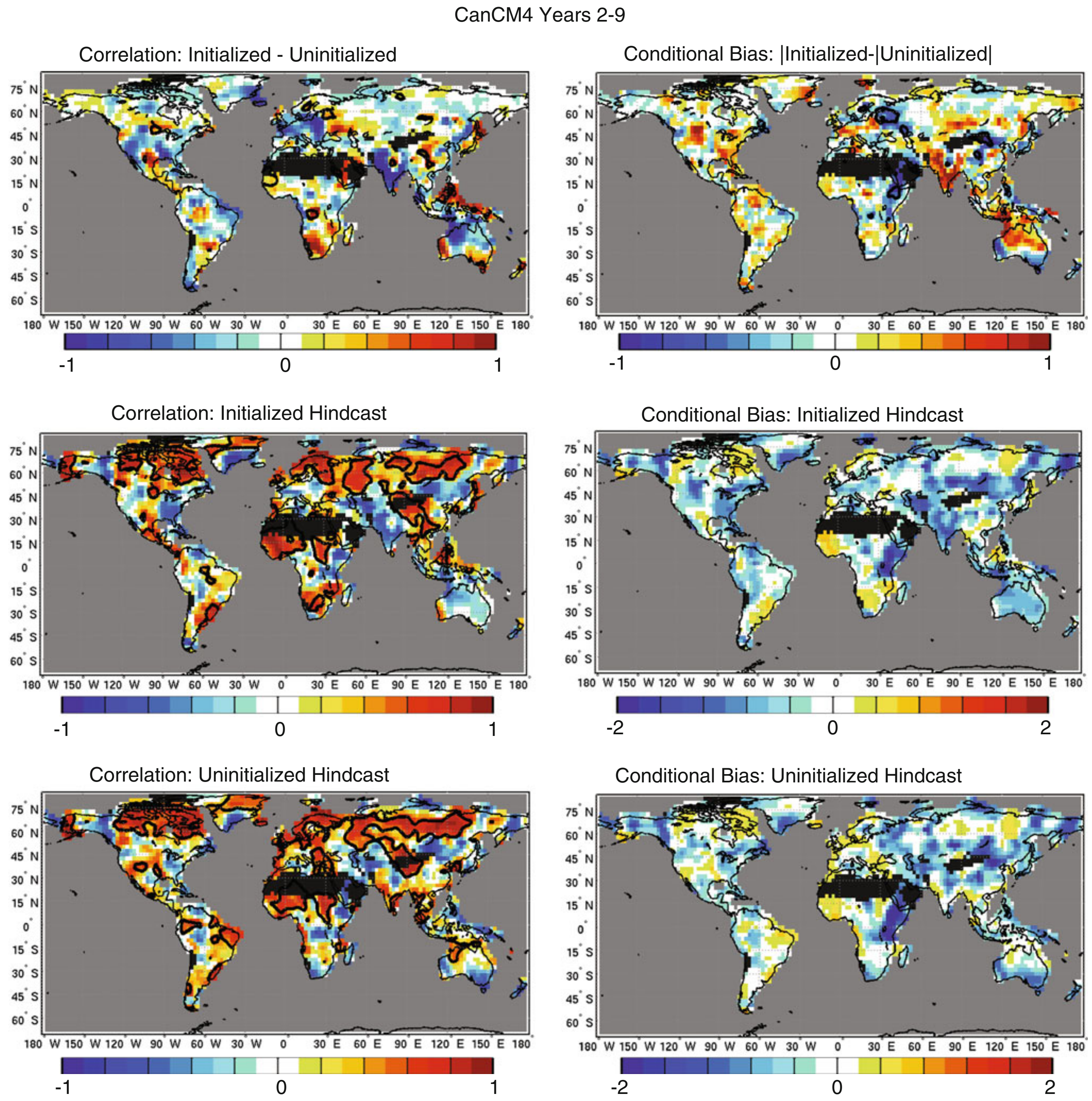

CanCM4 Years 2-9
Correlation: Initialized - Uninitialized
Conditional Bias: |Initialized-|Uninitialized|
Correlation: Initialized Hindcast
Conditional Bias: Initialized Hindcast
Correlation: Uninitialized Hindcast
Conditional Bias: Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1
-2
2

Fig. 8 Same as Fig. 6, but for precipitation hindcasts

(e.g. Fig. 7, middle and lower panels) than when the correlation is large and significant (compare Fig. 8, middle and lower maps). For the reduction in bias between the initialized predictions and uninitialized projections (Fig. 5, 6, 7, 8, upper right maps), blue areas are the regions of apparent improvement; these are the areas where the magnitude of the bias has been reduced because of initialization.

Returning to the question of what aspects of the predictive accuracy (i.e. *MSSS*(*H*,*P*,*O*)) are improved by the initialization, one discovers that for temperature most of the improvement is related to reduction of the conditional bias. Since the same model is used for both the initialized predictions and the uninitialized projections, this is likely due to the initialization itself being closer to the observed state. For precipitation, comparison of the upper panels of Figs. 7 and 8 with the upper panels of Fig. 3 suggest that improved forecast quality is due to both increased local correlation as well as reduction in conditional bias.

The results discussed above (Figs. 3, 4, 5, 6, 7, 8) are based on the spatially smoothed data that reduces grid scale noise. However, users of the decadal prediction experiments who require the climate data for applications or decision models may need the data at grid scale, or downscale it to even higher resolution. Thus it is useful to provide verification at the grid scale (Fig. 9), which here is chosen to be the scale of the gridded global observations. The correlation coefficients are much noisier and generally lower, as expected. Similar results hold for the other verification measures and other prediction systems (not shown).

2.2.2 *Probabilistic metrics*

In addition to establishing the level of accuracy in the ensemble mean prediction, one is often interested in

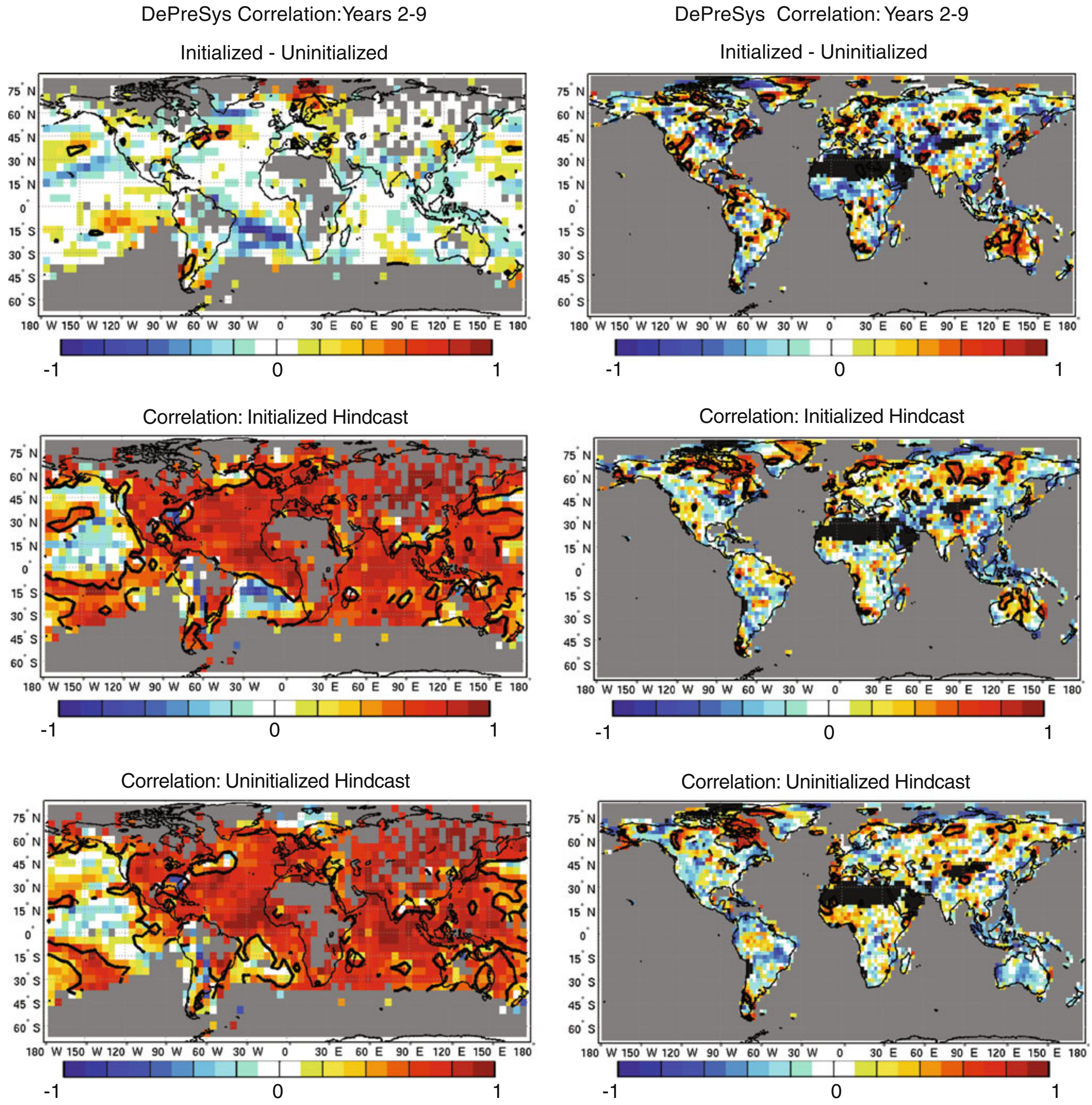

DePreSys Correlation:Years 2-9
Initialized - Uninitialized
DePreSys Correlation: Years 2-9
Initialized - Uninitialized
Correlation: Initialized Hindcast
Correlation: Initialized Hindcast
Correlation: Uninitialized Hindcast
Correlation: Uninitialized Hindcast
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1

Fig. 9 Anomaly correlation coefficient for DePreSys hindcasts (*left* temperature, *right* precipitation) with top row depicting the difference between the correlation of the initialized hindcasts (*middle row*) and that of the uninitialized hindcasts (*bottom*). Calculations are performed at the gridscale of the observations, which is 5° × 5° for temperature and 2.5° × 2.5° for precipitation. The forecast target is year 2–9 following the initialization every 5 years from 1961 to 2006 (i.e. 10 hindcasts). Contour line indicates statistical significance that the value is positive at the 95 % confidence level

quantifying the uncertainty, or the range of possibilities, in the prediction. This assessment requires the use of probabilistic metrics. The purpose of the probabilistic metric in this framework is not to ascertain skill of the forecast per se, but to test whether the ensemble spread in the prediction is adequate to quantitatively represent the range of possibilities for individual predictions over time. This is particularly important if the predictions are to be used for any quantitative assessment of climate-related risk.

Again, a skill score is used to determine the probabilistic quality of the prediction spread relative to some reference approach. The measure of probabilistic quality is the continuous ranked probability skill score (CRPSS, see "Appendix 1"). The CRPSS is based on the continuous ranked probability score, analogous to the relationship between MSSS and MSE. The CRPS is a measure of squared error in probability space. A commonly used probabilistic metric of forecast quality in seasonal prediction is the ranked probability skill score (RPSS), which looks at the squared error between observations and probabilistic categorical forecasts. A continuous score is preferable to a categorical score in the context of a non-stationary climate, where trends may lead to a chronic forecast of, say above-normal temperatures, and offer little discrimination among predictions, particularly the relative risk of attaining or exceeding some threshold.

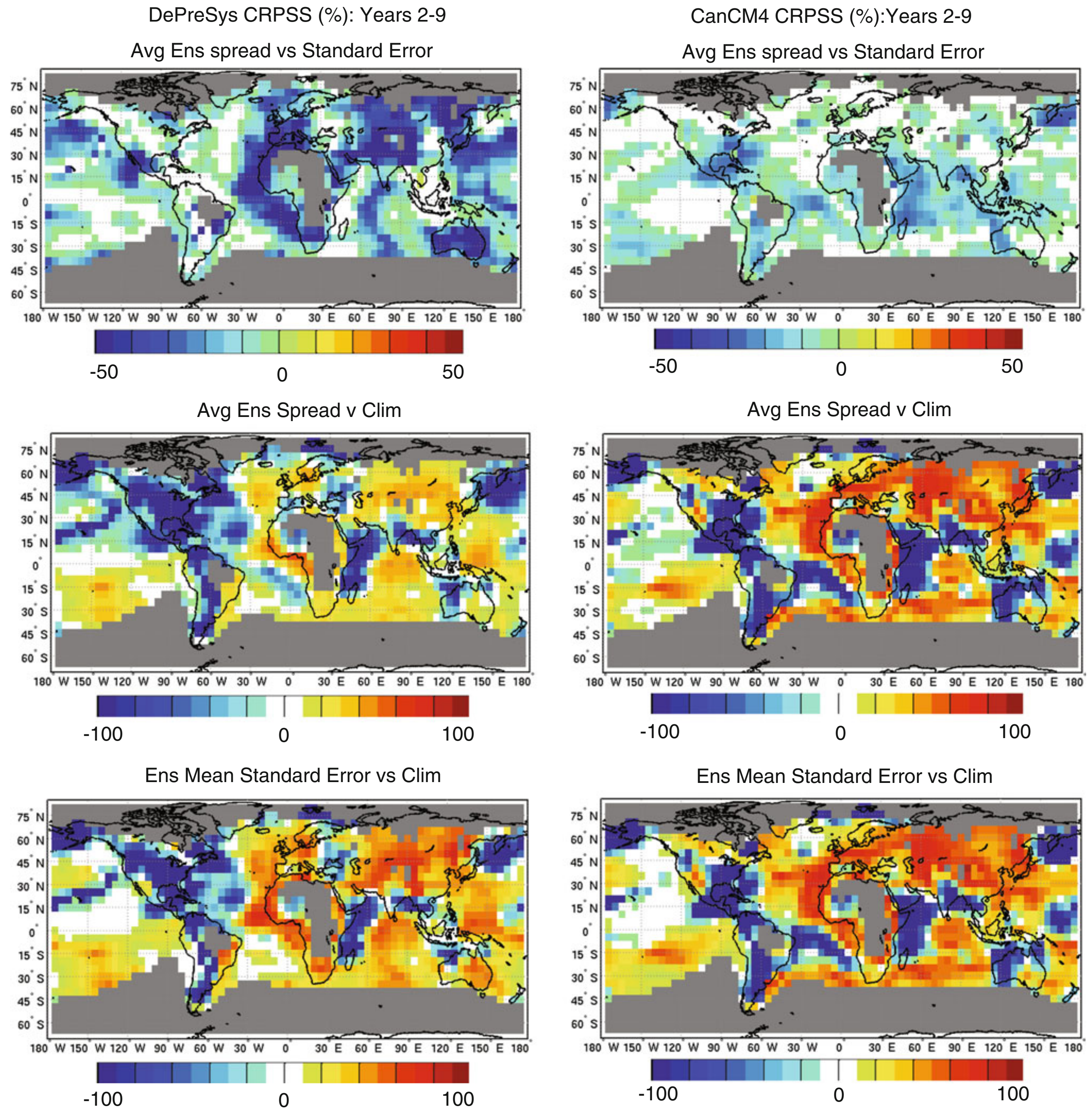
Following Question 2, we assess whether a model's average ensemble spread is suitable for quantifying prediction uncertainty compared to the standard error of the mean prediction, once corrected for conditional bias. The standard error is estimated as the standard deviation of the residuals from a simple least squares regression between the observations and the ensemble mean hindcasts. For a perfectly calibrated prediction system with sufficiently large ensembles, the spread of the ensemble members should represent the true range of possibilities for the future climate being predicted. In other words, over time the standard error of the ensemble mean prediction, relative to the observations, and the average spread of the ensemble members, as measured by the average standard deviation about the ensemble mean, should be the same. However, the models predictions are not necessarily well calibrated. The use of the standard error to represent the forecast uncertainty will be preferable if the spread of the ensemble members is systematically too narrow or too broad.

Since prediction uncertainty is under test here, conditional bias is removed to allow the biases in prediction spread to be assessed. If the conditional bias in the ensemble mean prediction were not first removed, it would dominate the probabilistic error in this metric. The average ensemble spread is used rather than the ensemble spread of each prediction because the experimental design of CMIP5, which calls for a nominal set of three ensemble members, is very small and sampling limitations are likely to dominate the uncertainty in the estimate of ensemble spread. Even prediction systems with order 10 ensemble members will face uncertainty in the ensemble mean and ensemble variance for a given prediction. However, negligible differences were found using the actual time-varying ensemble spread versus the average value for these 9-member ensemble predictions (not shown). It should also be noted that for the purposes of this exercise, the removal of biases and the estimation of the standard error were not conducted using cross-validation, which is how they should be implemented for use in real forecast systems. The reason for this is that the limited number of hindcast cases leads to so much sampling error in the adjustments, that the resulting cross-validated "unbiased" forecasts are noticeably worse than the raw biased predictions. More work is needed to identify what level of bias adjustments is possible given a particular experimental design of hindcasts and predictions.

Both the Hadley Centre and Canadian Climate Centre prediction systems exhibit considerable spatial structure for CRPSS of temperature (Fig. 10). As a squared error measure, the CRPS (and CRPSS) can be quite unforgiving; the presence of a small number of poor forecasts can greatly reduce the CRPS. The fact that regions of negative CRPSS appear in the comparison of the hindcasts with the climatological odds, even where there are significant positive correlations, suggests that the non-cross-validated bias adjustment still suffers from the problem of small sample size. The CRPSS of the temperature hindcasts show similar patterns whether one estimates the uncertainty in a given forecast from the average ensemble spread (Fig. 10, middle panels) or the standard error of the mean (Fig. 10, lower panels). The dominance of negative (blue) values in the comparative metric that tests the uncertainty from the ensemble members against the uncertainty from the standard error illustrates clearly that the use of the ensemble spread for individual predictions leads to less reliable forecasts (Fig. 10, top).

For precipitation, the probabilistic skill compared to the climatological distribution is at levels comparable to what is seen for seasonal forecasts that treat conditional biases (Goddard et al. 2003). As was the case with temperature, the use of standard error leads to improvement in probabilistic skill (Fig. 11). Looking at the comparison of CRPS from the different approaches to uncertainty indicates that the use of standard error to estimate forecast uncertainty is better for both temperature and precipitation from both sets of hindcasts (Figs. 10, 11, upper left panels).

Again, these results are based on non-cross-validated conditional bias adjustments and estimates of standard



DePreSys CRPSS (%): Years 2-9

Avg Ens spread vs Standard Error

CanCM4 CRPSS (%):Years 2-9

Avg Ens spread vs Standard Error

Avg Ens Spread v Clim

Avg Ens Spread v Clim

Ens Mean Standard Error vs Clim

Ens Mean Standard Error vs Clim

75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-50
0
50
-100
0
100

Fig. 10 Continuous ranked probability skill score (CRPSS) testing uncertainty quantification for (*left* DePreSys, *right* CanCM4). *Top row* CRPSS between the initialized temperature hindcasts with uncertainty given by the average ensemble spread against and the same hindcasts but with uncertainty given by the standard error of the ensemble mean; middle: CRPSS comparing hindcasts with uncertainty given by the average ensemble spread against the climatological distribution, and *bottom* CRPSS with uncertainty given by the standard error of hindcasts against the climatological distribution. The hindcasts have been adjusted for conditional bias. Observed and model data has been smoothed as described in text. Forecast target is year 2–9 after initialization every 5 years from 1961 to 2006 (i.e. 10 hindcasts)

error. Therefore the main conclusion that can be drawn here is that, as is common to seasonal-to-interannual model predictions, the spread of the ensemble members is not necessarily a good representation of forecast uncertainty. If nothing better can be done to recalibrate the ensemble predictions, this shortcoming must at least be acknowledged, and the possible implications for quantitative risk estimated.

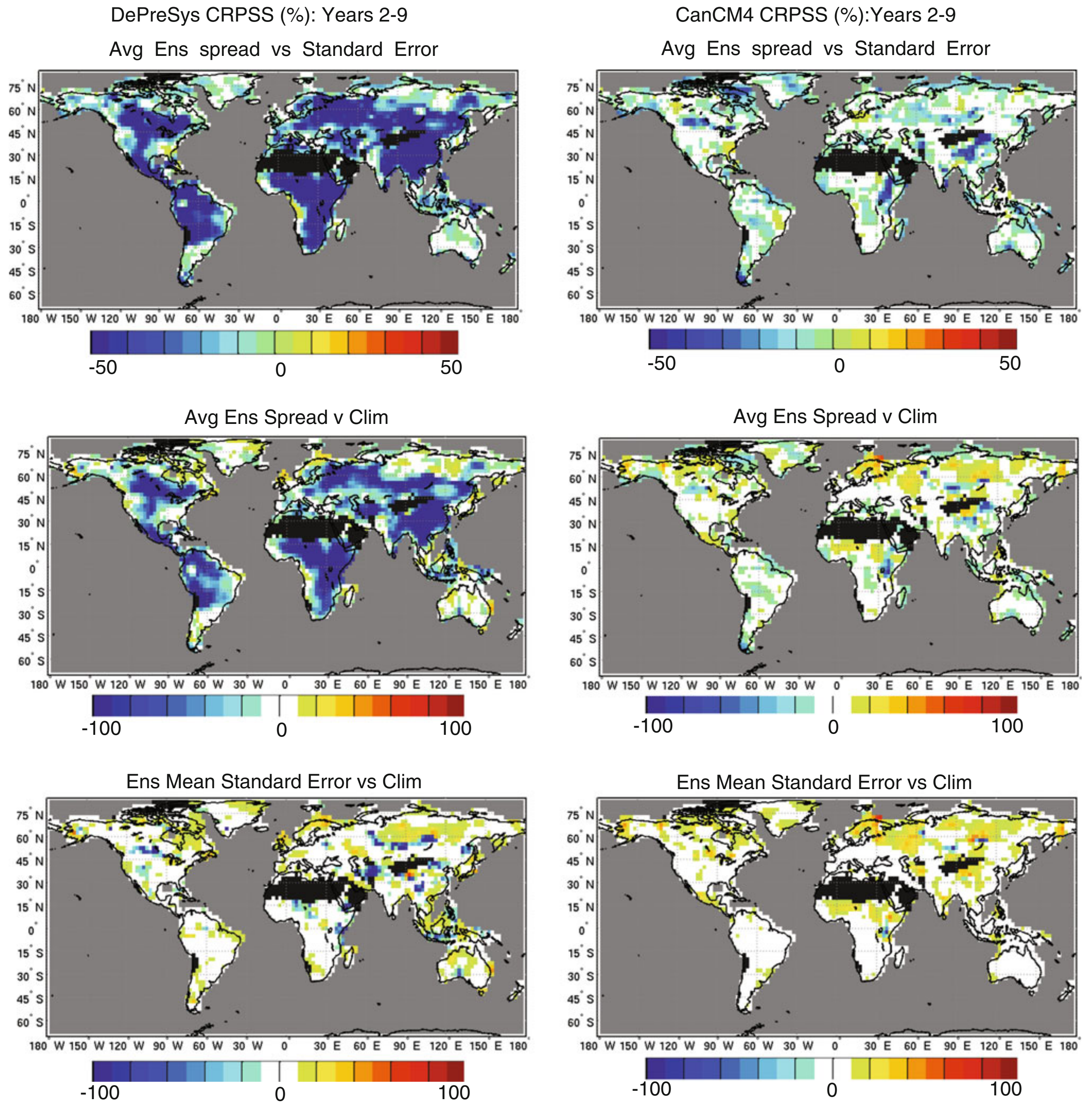

DePreSys CRPSS (%): Years 2-9
Avg Ens spread vs Standard Error
CanCM4 CRPSS (%):Years 2-9
Avg Ens spread vs Standard Error
-50
0
50
Avg Ens Spread v Clim
-100
0
100
Ens Mean Standard Error vs Clim
75° N
60° N
45° N
30° N
15° N
0°
15° S
30° S
45° S
60° S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180

Fig. 11 Same as Fig. 9, but for precipitation hindcasts

2.2.3 *Conclusions from examples of verification assessment*

The results from the hindcast verification performed on the two prediction systems yield some features that are also common to seasonal-to-interannual predictions. First, temperature is better predicted than precipitation. In this case the dominant signal is due to the upward trends, which are captured reasonably well by both systems over most of the world. In addition, precipitation is less coherent in space and time, and thus subject to larger noise-like variability that is not predictable. Second, different prediction systems often differ in where they perform well. Some common areas of good and poor performance shown through the verification scores are seen in both prediction systems. However, many differences exist, especially for precipitation, although these may be related to sampling issues. Additionally differences exist between prediction systems regarding the skill added by initialization.

Although these results may be sobering, they should not be viewed as a conclusion that there is no decadal predictability. As stated earlier, decadal prediction is very much an experimental and nascent activity, and how best to initialize and verify the predictions are matters of ongoing research. One positive result is the reduction in conditional bias that is seen for some areas in the initialized predictions, which is improved information about anthropogenic climate change, although much of this may be due to the initialization.

This paper outlines the framework and cannot show all the results, but there are instances of statistically significant skill obtained at the 1-year lead or 2–5 year period that do not appear in the decadal-scale results shown here (see http://clivar-dpwg.iri.columbia.edu/). Also, skill estimates are better estimated, and therefore show more coherent regions of significant skill, when more samples (i.e. start dates) are used compared to the every-5-year start dates originally mandated by the CMIP5 experimental design (e.g. Fig. 12, compare with left panels of Figs. 3, 4). This added robustness of skill estimates comes from reduced sampling errors in adjustment of the mean bias, as well as better sampling of the climate variability. The hindcasts still face a relatively short history in face of decadal-scale variability Finally, it is possible that gains in prediction quality may be made by multi-model ensemble approaches (e.g. Hagedorn et al. 2005), as has been the case for seasonal prediction (Robertson et al. 2004; Graham et al. 2011). Preliminary results based on just the two models used in this study show mixed results (Fig. 13). Statistical post-processing, or calibration, of model predictions may also improve forecast quality (e.g. Tippett et al. 2005). However, to do that robustly will require larger ensemble sizes and more samples (i.e. more start dates) than was mandated for CMIP5. Finally development of improved models, and improved understanding of the processes that must be modeled well, is ongoing throughout the scientific community, and is expected to improve the quality of decadal-scale climate information.

3 Summary and discussion

A framework for verification of interannual-to-decadal predictions has been described and illustrated for two prediction systems and for a specific prediction target of annual means over the years 2–9 following initialization. Similar analyses have been performed for multi-year averages over specific seasons with comparable results (see http://clivar-dpwg.iri.columbia.edu). The framework is not exhaustive, nor is it intended to be prescriptive, but rather it addresses a couple of fundamental questions about the initialized decadal prediction experiments. Given the truly experimental nature of the decadal prediction effort, the set of metrics from such a framework provides a useful baseline against which future improvements prediction system components can be quantified, including advances in the observing system, the assimilation methods used to map those observations into model initial conditions, and improvements in the models themselves. Equally important, the information on prediction quality across prediction systems provided by the framework puts the verification of prediction system on equal footing—observational verification data, verification period, spatial and temporal averaging, and even graphical presentation—such that relative assessments can be easily made. Additionally the framework provides guidance on the use of these model predictions, which differ in fundamental ways from the climate change projections that much of the community has become familiar with. This guidance includes adjustment of mean and conditional biases, and consideration of how to best approach forecast uncertainty.

Decadal prediction shares common scientific and methodological issues with both seasonal-to-interannual prediction and with climate change projection. Common to seasonal prediction, it is possible (and necessary) to examine the skill of past variability and events. Common to climate change projections, how variability and the mean climate might change subject to anthropogenic changes to the climate system are very difficult to separate and to test. While mean skill metrics will always be an important dimension of evaluating any prediction system, it is important to recognize that for decadal predictions, such metrics are subject to four important limitations:

1. The shortness of the observational record and dearth of ‘climate quality data’ (NRC 1999) leads to sampling uncertainty, which may be very large for decadal variability;
2. The observing system has varied considerably on decadal timescales, leading to large variations in knowledge of initial states, which can influence prediction bias (for an example of dependence of forecast bias on initial conditions in seasonal predictions see Kumar et al. 2012);
3. The climate itself is not stationary due to natural variability and anthropogenic radiative forcings;
4. There is considerable evidence that the predictability of climate on decadal timescales, as on shorter timescales, is state dependent.

For all these reasons it is difficult, if not impossible, to accurately quantify the mean skill of hindcasts. Therefore, the estimates obtained from the hindcasts may provide a poor, and even misleading, guide to the future performance of the decadal prediction systems.

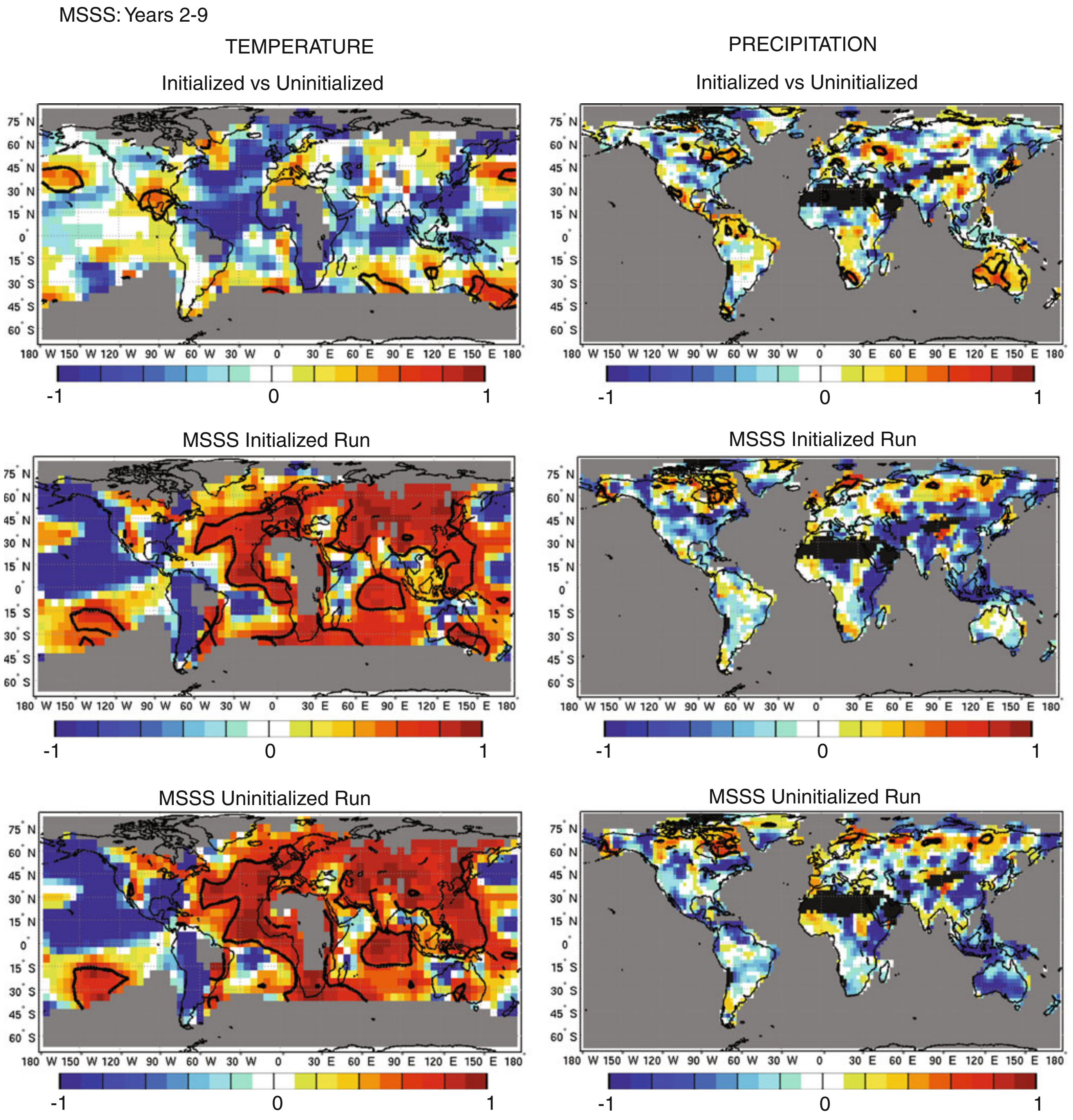


Fig. 12 Mean squared skill score (MSSS) for decadal temperature (*left*) and precipitation (*right*) hindcasts from the DePreSys prediction system of the Hadley Centre. *Top row* MSSS comparing the initialized hindcasts ("forecasts") and the uninitialized hindcasts ("reference") as predictions of the observed climate; middle row: MSSS comparing the initialized hindcasts ("forecasts") and the climatological mean ("reference"); bottom: MSSS between the uninitialized hindcasts ("forecasts") and the climatological mean ("reference"). Observed and model data has been smoothed as described in text. The forecast target is year 2–9 following the initialization every year (40 cases: 1960–2001 start dates for (1962–1969) to (2003–2010) decadal hindcast periods). *Contour line* indicates statistical significance that the MSSS is positive at the 95 % confidence level

Yet another reason that the skill of hindcasts may not be indicative of future performance is related to bias adjustment. As described in Sect. 2.1.3, mean bias adjustment of a model (regardless of the initialization procedure) is performed over a climatological reference period. If this reference period overlaps with the hindcast period, then the hindcasts are contaminated with observational data and the experiment cannot be considered "out of sample"

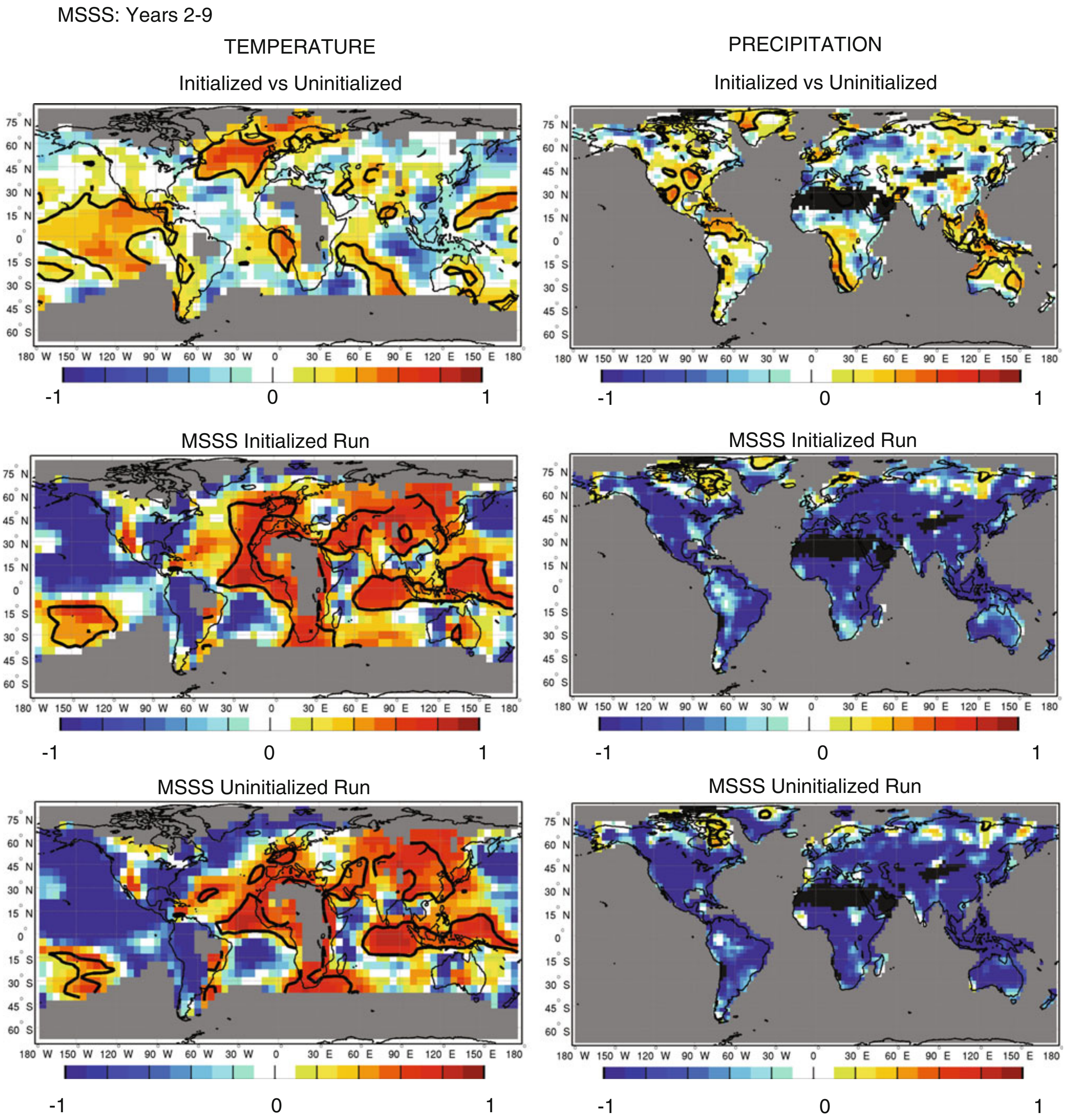

MSSS: Years 2-9
TEMPERATURE
Initialized vs Uninitialized
MSSS Initialized Run
MSSS Uninitialized Run
PRECIPITATION
Initialized vs Uninitialized
MSSS Initialized Run
MSSS Uninitialized Run
75°N 60°N 45°N 30°N 15°N 0 15°S 30°S 45°S 60°S
180 W 150 W 120 W 90 W 60 W 30 W 0 30 E 60 E 90 E 120 E 150 E 180
-1
0
1

Fig. 13 Mean squared skill score (MSSS) for decadal temperature (*left*) and precipitation (*right*) from the multi-model ensemble formed by combining the ensembles of DePreSys and CanCM4 hindcasts. *Top row* MSSS comparing the initialized hindcasts ("forecasts") and the uninitialized hindcasts ("reference") as predictions of the observed climate; *middle row* MSSS comparing the initialized hindcasts ("forecasts") and the climatological mean ("reference"); *bottom* MSSS between the uninitialized hindcasts ("forecasts") and the climatological mean ("reference"). Observed and model data has been smoothed as described in text. The forecast target is year 2–9 following the initialization every 5 years from 1961 to 2006 (i.e. 10 hindcasts). Contour line indicates statistical significance that the MSSS is positive at the 95 % confidence level

verification. For example, it can be shown that the expected value of the MSE of a prediction system is smaller within the climatological reference period than outside it. However, such effects may be difficult to distinguish from the effect of the inevitable sampling variation that occurs for different reference periods. This is illustrated in Fig. 14, which shows the effect of the length of reference period on the MSSS for initialized/uninitialized DePreSys hindcasts of global mean temperature. The MSSS varies considerably as the reference period is extended, but it is not clear how much of this effect is from the changing degree of reference/hindcast overlap and how much is from sampling variation. Note though, that the MSSS for global mean temperature is always positive, suggesting that the conclusion that the initialized model performs better than the uninitialized for this variable is robust (Smith et al. 2007).

Finally, note that to create confidence in the interannual-to-decadal predictions, the model processes ultimately must be validated. The relative roles of oceanic, atmospheric and coupled processes in specific events must be analyzed in observations and across prediction systems. This is a natural extension of the verification analysis, and an important complement. A complementary approach to judging hindcasts through mean skill metrics is model validation through the case study approach, which seeks to confirm that models produce climate variability for the right reasons. A crucial dimension of the case study approach is that the assessment must be process-based. The purpose is not merely to assess whether the event was predicted, but whether the mechanisms captured in the prediction system are consistent with those that were responsible for the change in the real world. Developing a clear understanding of these processes is therefore an essential first step. The idea is to identify events in the historical record when unusually large change occurred on decadal time scales, and then to focus detailed analysis on assessing the performance of the prediction system for the event(s) in question. Such events are rare, and may be viewed as "surprises"; arguably it is the importance of providing advanced warning of such surprises that motivates the need for a decadal prediction system. Examples of such events include the 1976 Pacific "climate shift" (Trenberth and Hurrell 1994; Meehl et al. 2009, 2010), the rapid cooling of the North Atlantic Ocean in the 1960s (Thompson et al. 2010), and the rapid warming of the North Atlantic in the mid 1990s (Robson et al. 2012). The case study approach in the validation for decadal prediction systems is an important complement to overall metrics of skill, such as those proposed in this framework.

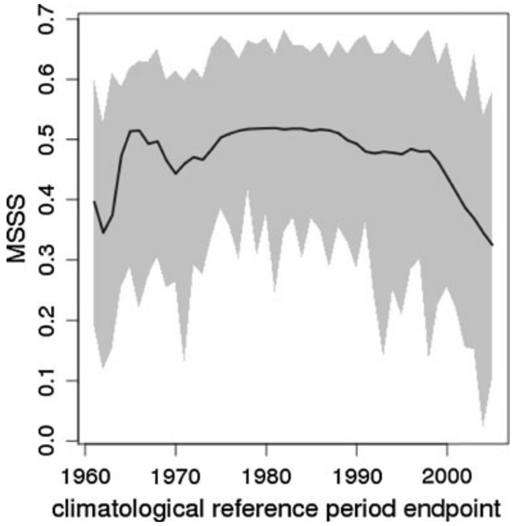


Fig. 14 MSSS for initialized versus uninitialized Hadley Centre hindcasts of global mean temperature, plotted against the end date of the climatological reference period (black line). The climatological reference period always starts in 1960. The gray band shows pointwise 90 % confidence intervals estimated using the bootstrap method outlined in "Appendix 2"

In the meantime, and for those interested in using decadal climate prediction experiments, the verification framework provides some guidance and an initial baseline for the capabilities of current prediction systems.

Acknowledgments The authors of this paper are members of the Decadal Predictability Working Group sponsored by US CLIVAR. We thank the two reviewers of this paper, who offered considerable constructive suggestions toward the improvement of this work. The Working Group appreciates the support from the US CLIVAR office. Goddard, Gonzalez and Greene received funding from a NOAA grant (NA08OAR4320912) for work on this project. Amy Solomon received support form NOAA OAR CVP and NSF AGS #1125561. Doug Smith was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101) and the EU ENSEMBLES project. Deser and Meehl were supported by the Office of Science (BER), US Department of Energy, Cooperative Agreement No. DE-FC02-97ER62402, and the National Science Foundation that sponsors the National Center for Atmospheric Research. Matthew Newman was supported by the NOAA OAR CVP program. Sutton and Hawkins' contributions were covered by support form support from the UK National Centre for Atmospheric Science. Fricker, Ferro, and Stevenson contributions were supported by byNERC Directed Grant NE/H003509/1. Hegerl's contribution was funded by EQUIP, NERC NE/H003533/1. Kushir's contribution to this work was funded by NOAA Grant NA0OAR4320912.

Appendix 1: Probabilistic metrics

The measure of probabilistic quality applied here is the continuous ranked probability score (CRPS). The ranked probability score is commonly used to assess probabilistic forecasts (e.g. Goddard et al. 2003; Barnston et al. 2010), but is typically used with categorical forecasts. Since the changing background climate subverts the usefulness of categorical forecasts, we wish to cast the hindcasts in terms of a continuous, quantitative, analytical distribution with a mean and standard deviation determined from the hindcast ensemble, although clearly both of these parameters are subject to substantial sampling errors with the small nominal ensemble sizes requested for CMIP5. The notation H_{ij} is used for the initialized hindcasts, with ensemble mean, H_j, where $i = 1,N$ represents the ensemble member and $j = 1,n$ represents the start time.

By definition, the CRPS is:

$$CRPS(H_{ij}, O_j) = \int_{-\infty}^{+\infty} \left(\mathcal{G}(H_j) - \mathcal{H}(O_j)\right)^2 dy \qquad (7)$$

where $\mathcal{G}$ and $\mathcal{H}$ represent the cumulative distribution functions of the hindcast (as a cumulative Gaussian distribution) and the observations O_j (as the Heaviside function), respectively. The subscript on H nominally represents the ensemble in general terms. Here, $\mathcal{G}(H_j)$ represents the cumulative version of a Gaussian distribution centered on the mean H_j with a spread determined by the variance of the ensemble members about that mean, but $\mathcal{G}(H_j)$ could alternatively represent the empirical cumulative distribution function. The CRPS is very much like the mean squared error, but in probability space. If the hindcast distribution is identical to the observed distribution for all times, which would also coincide with an accurate deterministic prediction for the observed outcome in every case, the CRPS would be 0.

In the present case where $\mathcal{H}\ (O_j)$ is the Heaviside function and $\mathcal{G}(\hat{H}_j)$ is the Gaussian distribution with mean $\hat{H}_j$ and ensemble variance about that mean, $\sigma^2_{\hat{H}}$, then it follows that (Gneiting and Raftery 2007):

$$CRPS(N(\hat{H}_j, \sigma^2_{\hat{H}_j}), O_j) = \sigma_{\hat{H}_j}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{O_j - \hat{H}_j}{\sigma_{\hat{H}_j}}\right) - \frac{O_j - \hat{H}_j}{\sigma_{\hat{H}_j}}\left(2\phi\left(\frac{O_j - \hat{H}_j}{\sigma_{\hat{H}_j}}\right) - 1\right)\right] \qquad (8)$$

where φ and ϕ represent the probability distribution function (pdf) and cumulative distribution function (cdf) of a standard Gaussian variable, respectively. Note that the hindcast value $\hat{H}_j$ is not necessarily identical to H_j used for the deterministic metrics, which has only the mean-bias removed; $\hat{H}_j$ has been corrected also for the conditional bias as diagnosed through the deterministic metrics. The slope of the regression line between the observations (given the hindcasts) and the hindcasts is $\left({s_O}/{s_H}\right) r_{HO}$, which is the scaling used to correct the hindcasts for the conditional bias (Murphy 1988), where s_O and s_H are the square roots of the sample variances of the observations and predicted ensemble means, respectively, and r_{HO} is the correlation between the observations and ensemble mean hindcasts. The corrected ensemble mean predictions are given by

$$\hat{H}_j = \left({s_O}/{s_H}\right) r_{HO}\, H_j \qquad (9)$$

Thus each ensemble member for a given prediction is shifted by the same amount, as determined by $\hat{H}_j - H_j$. This correction should be cross-validated by holding out the set of ensemble members being corrected, but that is not done in the analysis here.

Given the $\mathrm{CRPS_F}$ for the hindcast distribution, and the $\mathrm{CRPS_R}$ for the reference distribution the corresponding skill score can be defined as:

$$CRPSS = 1 - \frac{\sum_{j=1}^{n} CRPS_{F_j}}{\sum_{j=1}^{n} CRPS_{R_j}} \qquad (10)$$

The hindcast distribution at this stage is assumed Gaussian, with the mean given by the corrected ensemble mean and the variance given by the ensemble variance. Since we are only testing the uncertainty in the hindcasts, the mean of the distribution is the same for both the hindcast under test and the reference hindcast.

With the CRPSS, the question addressed is: Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?

For the hindcast distribution, the variance of the hindcast distribution is calculated from the average variance over n hindcasts of the N ensemble members:

$$\bar{\sigma}^2_F = \frac{1}{n}\sum_{j=1}^{n} \sigma^2_{\hat{H}_j} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{H}_{ij} - \hat{H}_j\right)^2 \qquad (11)$$

The variance for the reference distribution is given by the root mean squared error between the hindcast ensemble mean and the observations:

$$\bar{\sigma}^2_R = \frac{\sum_{j=1}^{n}\left(\hat{H}_j - O_j\right)^2}{n-2} \qquad (12)$$

It should be noted that if mean or conditional biases remain in the hindcasts, the standard error between the hindcast mean and the observations, $\bar{\sigma}_R$, may actually be larger than the climatological variance of the observations (Johnson and Bowler 2009). This is another reason that hindcast data should be used judiciously.

Appendix 2: Statistical significance estimation

Statistical significance of verification scores and differences between scores is an important component of any verification assessment. The extremely small set of hindcasts leads to sampling issues. The presence of a trend leads to auto-correlation within the hindcasts and the observed time series, and reduces the degrees of freedom further, though not necessarily consistently across variables or over space for a particular variable. These issues violate many assumptions contained in standard tables of significance for a given number of data values.

In the verification framework presented in this manuscript, significance of the verification scores was assessed through a non-parametric bootstrapping approach. The initialized hindcasts are given by, H_{ij}, with ensemble mean, H_j, where $i = 1,N$ represents the ensemble member and $j = 1,n$ represents the start time. The 'reference' hindcasts, which here are the uninitialized hindcasts (i.e. climate change projections), are given by P_{ij}, with ensemble mean, P_j.

A set of $k = 1,M$ re-samplings of the initialized and uninitialized simulations is created in order to obtain probabilistic information for the scores. In each sample, a new ensemble is created as follows: a set of n start times, $J(j,k)$, are randomly selected from the set of years with replacement; followed by the random selection with replacement of N ensemble members taken from each start time, $I(J)$. So, for a given sample k and a given start time I the simulated ensembles are built as:

$$\tilde{H}_{ij}(k) = H_{I(J)J(j,k)} \tag{13}$$

$$\tilde{P}_{ij}(k) = P_{I(J)J(j,k)} \tag{14}$$

Due to the ensemble resampling, it is clear that $\tilde{H}_j(k) \; / = H_{J(j,k)}$ because even though they represent the same start time, they are likely to be comprised of a different set of N ensemble members.

An additional step has been used in the application of this methodology to the problem of the decadal hindcasts. Given the likely trends in the time series due to anthropogenic forcing and thus temporal auto-correlation, the start time re-sampling actually takes pairs of start times. Since the nominal experimental design dictates start times every 5 years, we only consider these neighboring 5-year pairs. Verification using data with more frequent start times, may want to use longer runs over time. In our case, for example, if in our first bootstrapped time series we choose I = 5 (realtime: 1980) for i = 1, then we will take I = 6 (realtime: 1985) for i = 2, then draw randomly for i = 3, and choose the following start time for i = 4, etc. For each I, a new random set of ensembles (J's) are drawn. Once this M ensemble mean time series of bootstrapped hindcasts are created, they can be used to create distributions for the different scores.

Correlation coefficients

In the case of the correlation coefficients, three different statistics are tested:

r_{HO}: correlation between the initialized hindcasts, H_j, and the corresponding observations

r_{PO}: correlation between the uninitialized (or reference) hindcasts, P_j, and the corresponding observations

$\Delta r = r_{HO} - r_{PO}$: the difference between the two.

Using the previous bootstrap-generated series, the three previous statistics can be calculated for each sample, k, as:

$\tilde{r}(k)_{HO}$: correlation between the bootstrapped initialized hindcast sample, $\tilde{H}_j(k)$, and the corresponding observations

$\tilde{r}(k)_{PO}$: correlation between the bootstrapped uninitialized (or reference) hindcast sample, $\tilde{P}_j(k)$, and the corresponding observations

$$\Delta\tilde{r}(k) = \tilde{r}(k)_{HO} - \tilde{r}(k)_{PO}$$

The significance of only positive values is assessed, since a positive r_{HO} or r_{PO} accounts for a positive linear relationship between the hindcasts and observations, and in the case of a positive value of Δr, it represents an improvement in the representation of the observations due to model initialization.

To test the significance of any of these metrics, for example r_{HO}, the distribution of the M $\tilde{r}_{HO}$ values is used. The fraction of observed negative values serves as a p value for the test and is compared to the chosen significance level α. If the p value is lower that or equal to α, r_{HO} is considered significant for the $(1-\alpha) \times 100$ % confidence level. The same approach is applied to r_{PO} using the observed frequencies for $\tilde{r}(k)_{PO}$, and to Δr using the distribution obtained for $\Delta\tilde{r}(k)$.

Mean squared skill score

The mean squared skill score is the primary metric used to measure the improvement of the hindcast due to model initialization. The same bootstrap-generated time series can be used to generate a distribution of values for the MSSS, based on calculating $MS\tilde{S}S(k)$ with $k = 1,M$. The fraction of observed negative values again serves as the p value for the test and is compared to the chosen significance level α. If the p value is lower or equal than α, MSSS is considered significant for the $(1-\alpha) \times 100$ % confidence level.

Conditional bias

In addition, the decrease in magnitude of the conditional bias, db, is also used to assess the relative improvement in forecast accuracy due to model initialization.

$$db = |bP| - |bH|$$

where bP is the conditional bias of the uninitialized (reference) projections, and bH is the conditional bias of the initialized hindcasts.

A distribution of M values for the fractional decrease can be created by calculating this metric for the bootstrapped time series.

$$d\tilde{b}(k) = \left|b\tilde{H}\right| - \left|b\tilde{P}\right|$$

Based on this population of values for the fractional decrease in conditional biases, the p value is given by the fraction of negative values and compared with the significance level α. The decrease is considered significant if the p value is lower or equal than α.

References

Arora V, Scinocca J, Boer G, Christian J, Denman KL, Flato G, Kharin V, Lee W, Merryfield W (2011) Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. Geophys Res Lett:L05805. doi:10.1029/2010GL046270

Barnston AG, Li S, Mason SJ, DeWitt DG, Goddard L, Gong X (2010) Verification of the first 11 years of IRI's seasonal climate forecasts. J App Meteor Climatol 49(3):493–520. doi:10.1175/2009JAMC2325.1

Boer GJ (2004) Long time-scale potential predictability in an ensemble of coupled climate models. Clim Dyn 24:29–44

Boer GJ, Lambert SJ (2008) Multi-model decadal potential predictability of precipitation and temperature. Geophys Res Lett 35:L05706. doi:10.1029/2008GL033234

Branstator G, Teng H (2010) Two limits of initial-value decadal predictability in a CGCM. J Clim 23:6292–6311

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. J Geophys Res 111:D12106. doi:10.1029/2005JD006548

Collins M et al (2006) Interannual to decadal climate predictability in the North Atlantic: a multi model-ensemble study. J Clim 19:1195–1203

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2010) Climate model errors, feedbacks and forcing: a comparison of perturbed physics and multi-model ensembles. Clim Dyn. doi:10.1007/s00382-101-0808-0

Du Y, Xie S-P (2008) Role of atmospheric adjustments in the tropical Indian Ocean warming during the 20th century in climate models. Geophys Res Lett 35:L08712. doi:10.1029/2008GL033631

Giannini A, Saravanan R, Chang P (2003) Oceanic forcing of Sahel rainfall on interannual to interdecadal timescales. Science 302:1027–1030

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Amer Stat Assoc 102:359–378. doi:10.1198/016214506000001437

Goddard L, Barnston AG, Mason SJ (2003) Evaluation of the IRI's "net assessment" seasonal climate forecasts: 1997–2001. Bull Amer Meteor Soc 84:1761–1781

Goddard L, Dilley M (2005) El Nino: catastrophe or opportunity. J Clim 18:651–665

Goddard L, Hurrell JW, Kirtman BP, Murphy J, Stockdale T, Vera C (2012) Two time scales for the price of one (almost). Bull Amer Meteor Soc 93:621–629. doi:10.1175/BAMS-D-11-00220.1

Goldenberg SB, Landsea CW, Mestas-Nuñez AM, Gray WM (2001) The recent increase in Atlantic hurricane activity: causes and implications. Science 293:474–479

Gordon C et al (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16:147–168

Graham RJ, Yun W-T, Kim J, Kumar A, Jones D, Bettio L, Gagnon N, Kolli RK, Smith D (2011) Long-range forecasting and the global framework for climate services. Clim Res 47:47–55. doi:10.3354/cr00963

Griffies SM, Bryan K (1997) Predictability of North Atlantic multidecadal climate variability. Science 275:181–184

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A 57:219–233. doi:10.1111/j.1600-0870.2005.00103.x

ICPO (International CLIVAR Project Office) (2011) Decadal and bias correction for decadal climate predictions. January. International CLIVAR Project Office, CLIVAR Publication Series No. 150, 6 pp. Available from http://eprints.soton.ac.uk/171975/1/150_Bias_Correction.pdf

Johnson C, Bowler N (2009) On the reliability and calibration of ensemble forecasts. Mon Wea Rev 137:1717–1720

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. doi:10.1038/nature06921

Kharin VV, Boer GJ, Merryfield WJ, Scinocca JF, Lee W-S (2012) Statistical adjustment of decadal predictions in a changing climate. Geophys Res Lett. (in revision)

Knight JR, Allan RJ, Folland CK, Vellinga M, Mann ME (2005) A signature of persistent natural thermohaline circulation cycles in observed climate. Geophys Res Lett 32:L20708. doi:10.1029/2005GL024233

Kumar A (2009) Finite samples and uncertainty estimates for skill measures for seasonal predictions. Mon Wea Rev 137:2622–2631

Kumar A et al (2012) An analysis of the non-stationarity in the bias of sea surface temperature forecasts for the NCEP Climate Forecast System (CFS) version 2. Mon Wea Rev (to appear). http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-11-00335.1

Livezey RE, Chen WY (1983) Statistical field significance and its determination by Monte Carlo methods. Mon Weather Rev 111:46–59

Meehl GA, Hu A, Santer BD (2009) The mid-1970s climate shift in the Pacific and the relative roles of forced versus inherent decadal variability. J Clim 22:780–792

Meehl GA, Hu A, Tebaldi C (2010) Decadal prediction in the Pacific region. J Clim 23:2959–2973

Meehl GA et al. (2012) Decadal climate prediction: an update from the trenches. Bull Amer Meteorol Soc. (Submitted)

Merryfield WJ et al. (2011) The Canadian seasonal to interannual prediction system (CanSIPS). Available on-line from http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/op_systems/doc_opchanges/technote_cansips_20111124_e.pdf

Msadek R, Dixon KW, Delworth TL, Hurlin W (2010) Assessing the predictability of the Atlantic meridional overturning circulation and associated fingerprints. Geophys Res Lett:37. doi:10.1029/2010GL044517

Murphy AH (1988) Skill scores based on the mean squared error and their relationships to the correlation coefficient. Mon Wea Rev 116:2417–2424

NRC (1999) Adequacy of climate observing systems. The National Academies Press, Washington, DC, p 66

Pierce DW, Barnett TP, Tokmakian R, Semtner A, Maltrud M, Lysne J, Craig A (2004) The ACPI project, element 1: initializing a coupled climate model from observed initial conditions. Clim Change 62:13–28

Pohlmann H, Botzet M, Latif M, Roesch A, Wild M, Tschuc P (2004) Estimating the decadal predictability of a coupled AOGCM. J Clim 17:4463–4472

Power SB, Mysak LA (1992) On the interannual variability of arctic sea-level pressure and sea ice. Atmos Ocean 30:551–577

Räisänen J, Ylhäisi JS (2011) How much should climate model output be smoothed in space? J Clim 24:867–880

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407. doi:10.1029/2002JD002670

Robertson AW, Lall U, Zebiak SE, Goddard L (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon Wea Rev 132:2732–2744

Robson J, Sutton R, Lohmann K, Smith D, Palmer MD (2012) Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. J Clim 25:4116–4134. doi:10.1175/JCLI-D-11-00443.1

Rudolf B, Becker A, Schneider U, Meyer-Christoffer A, Ziese M (2010) GPCC Status Report December 2010. GPCC, December 2010, 7 pp

Schneider U, Fuchs T, Meyer-Christoffer A, Rudolf B (2008) Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre (GPCC), DWD, Internet Publikation, pp 1–12

Smith DM, Murphy JM (2007) An objective ocean temperature and salinity analysis using covariances from a global climate model. J Geophys Res 112:C02022. doi:10.1029/2005JC003172

Smith D, Cusack S, Colman A, Folland C, Harris G, Murphy J (2007) Improved surface temperature prediction for the coming decade from a global circulation model. Science 317:796–799

Smith TM, Reynolds RW, Peterson TC, Lawrimore J (2008) Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). J Clim 21:2283–2296

Smith DM, Eade R, Dunstone NJ, Fereday D, Murphy JM, Pohlmann H, Scaife AA (2010) Skilful multi-year predictions of Atlantic hurricane frequency. Nat Geosci. doi:10.1038/NGEO1004

Stephenson DB, Coelho CAS, Doblas-Reyes FJ, Balmaseda M (2005) Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. Tellus A 57:253–264. doi:10.1111/j.1600-0870.2005.00110.x

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

Tang Y, Kleeman R, Moore AM, Vialard J, Weaver A (2004) An off-line, numerically efficient initialization scheme in an oceanic general circulation model for El Nino-Southern Oscillation prediction. J Geophys Res 109:C05014. doi:10.1029/2003JC002159

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Amer Meteorol Soc 92:485–498. doi:10.1175/BAMS-D-11-00094.1

Teng H, Branstator G, Meehl GA (2011) Predictability of the Atlantic overturning circulation and associated surface patterns in two CCSM3 climate change ensemble experiments. J Clim 24:6054–6076. doi:10.1175/2011JCLI4207.1

Thompson DWJ, Wallace JM, Kennedy JJ, Jones PD (2010) An abrupt drop in Northern Hemisphere sea surface temperature around 1970. Nature 467:444–447. doi:10.1038/nature09394

Ting M, Kushnir Y, Seager R, Li C (2009) Forced and internal 20th century SST trends in the North Atlantic. J Clim 22(6): 1469–1481. doi:10.1175/2008JCLI2561.1

Tippett MK, Goddard L, Barnston AG (2005) Statistical-dynamical seasonal forecasts of Central Southwest Asia winter precipitation. J Clim 18:1831–1843

Trenberth KE, Hurrell JW (1994) Decadal atmosphere-ocean variations in the Pacific. Clim Dyn 9:303–319

Troccoli A, Balmaseda MA, Segschneider J, Vialard J, Anderson DLT, Haines K, Stockdale T, Vitart F, Fox AD (2002) Salinity adjustments in the presence of temperature data assimilation. Mon Wea Rev 130:89–102

Vera C, Barange M, Dube OP, Goddard L, Griggs D, Kobysheva N, Odada E, Parey S, Polovina J, Poveda G, Seguin B, Trenberth K (2010) Needs assessment for climate information on decadal timescales and longer. Procedia Environ Sci 1:275–286. doi:10.1016/j.proenv.2010.09.017

Wang W, Chen M, Kumar A (2010) An assessment of the CFS real-time seasonal forecasts. Wea Fcsting 25:950–969

Zhang R, Delworth TL (2006) Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. Geophys Res Lett 33:L17712. doi:10.1029/2006GL026267