

A Verification Framework for Interannual-to-Decadal Prediction Experiments

Lisa Goddard⁺, Paula Gonzalez⁺, Simon Mason⁺, Arthur Greene⁺,
and the US CLIVAR Working Group on Decadal Predictability

⁺*International Research Institute for Climate & Society (IRI), Columbia University*

1. Introduction: motivations and overview of the work

This work resulted from the efforts of the US CLIVAR Working Group on Decadal Prediction (DPWG) (<http://www.usclivar.org/wgdp.php>). The authors listed explicitly on this paper are scientists at IRI that participated in the working group. This limited lifetime working group addressed two objectives during our 2009-2011 tenure. The first focused on methodologies that attempt to separate natural from forced climate changes (Solomon *et al.* 2011). The second objective was aimed at the validation and verification of dynamical decadal hindcasts through a set of targeted metrics. The work presented here represents an illustration of the resulting hindcast verification work (see Goddard *et al.* 2012 for more complete details).

Verification of forecasts is needed not only to indicate their expected skill for those who may wish to apply them. Verification also allows improvements in prediction systems to be tracked over time, and allows for comparison across different prediction systems and/or different prediction approaches. The value of a verification framework is provision of guidelines for a common format of the results across hindcasts from different prediction centers, such as common observational data for verification, common period over which the hindcasts are verified and common period against which anomalies are calculated, common metrics, and even common graphical presentation. Through the efforts of our working group, these results are being collected on a central website: <http://dpwg-clivar.iri.columbia.edu>).

2. Data and methodology/experimental design

2a. Data

The dynamical decadal hindcasts presented here include: the perturbed physics experiments from Hadley Centre using an updated version of the DePreSys forecast system (Smith *et al.* 2010) and the hindcasts from the Canadian Climate Centre using CanCM4 (Arora *et al.* 2011, Merryfield *et al.* 2011). The verification dates are those dictated by the CMIP5 experimental design (Taylor *et al.* 2011), which contain 10 sets of hindcasts – initialized near the end of 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, and 2005. The two hindcast datasets both include more years than this nominal set, and it should be noted that skill metrics generally indicate better performance when the more complete set of hindcasts is used. However, since the common denominator for the CMIP5 hindcasts are these start dates, this is what is shown. Visit the DPWG verification site for more complete information.

The verification is performed for air temperature and precipitation, with the following observational data sets: (1) Air temperature: Hadley Centre/Climate Research Unit Temperature version 3 variance - adjusted (HadCRUT3v; available on a 5°x5° grid). Preference is given to the HadCRUT3v data because missing data is indicated as such. This can make verification over time more difficult, but it also provides a more realistic view of where forecasts can be verified with gridded data. (2) Precipitation: Global Precipitation Climatology Centre version 4 (GPCPv4). This dataset covers the period 1901-2007, at a resolution of 2.5°x2.5° grid, although the data is provided also at higher resolutions.

The model 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 done at a resolution of 5x5 degrees resolution for temperature and 2.5x2.5 degrees resolution for precipitation. We also do the verification analysis on a smoothed version of the model and observational data. A balance between skill improvement

signal-to-noise retention suggests that 5° latitude x 5° longitude represents a reasonable scale for smoothing precipitation, and 10° latitude x 10° longitude for temperature (Räsänen 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. Both grid-scale and spatially-smoothed verification contain useful information, and in combination provide guidance on the robustness of signals on different spatial scales.

2b. Methodology: Verification metrics

Verification metrics are chosen to answer specific questions regarding the quality of the forecast information. Our questions address the accuracy in the forecast information (Q1) and the representativeness of the forecast ensembles to indicate forecast 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 forecast uncertainty on average?

In both cases skill scores are used, which allow for comparison against a chosen baseline. The first question (Q1) is assessed using the mean squared skill score (MSSS, Murphy 1988). When the baseline is the climatological average (common baseline in seasonal-to-interannual predictions), the MSSS can be decomposed into the square of the correlation and the square of the conditional bias. To answer Q1, the baseline skill is defined by the uninitialized climate change projections from the same model as the initialized hindcasts. The MSSS thus quantifies the improvement in the mean squared error (MSE) for the initialized versus uninitialized hindcasts.

To address Q2, we use a measure of probabilistic quality: the continuous ranked probability skill score (CRPSS). 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 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 forecasts, particularly the relative risk of attaining or exceeding some threshold. Following Q2, we assess whether a model's average ensemble spread is suitable for quantifying forecast uncertainty compared to the standard error of the mean forecast, once corrected for conditional bias.

2c. Methodology: Statistical significance

Statistical significance must be estimated, and is of particular importance when sampling uncertainty is likely to be large, such in this case with only 10 hindcasts over a 45-year period for phenomena with a timescale of 20+ years. There is no single way to assess significance. The DPWG verification framework uses a non-parametric bootstrap approach that takes serial autocorrelation into account. (see Goddard *et al.* 2012 for more details)

3. Results

Deterministic (Q1):

The MSSS comparing the initialized and uninitialized temperature hindcasts is shown at the top of Figure 1 based on the spatially smoothed data for the forecast target of the average over years 2-9, or equivalently a 1-year lead-time for a decadal-average prediction. Additionally, the MSSS for each of those hindcasts relative to a climatology baseline is shown in the lower panels. For both prediction systems, the MSSS for temperature from the initialized predictions and the uninitialized projections show positive values over much of the map, suggesting that the trend plays an important role in the MSSS when using a climatological reference forecast. Most of the places where the MSSS is worse (negative or blue areas in the figure) than the reference forecast of climatology (Figure 1, middle and bottom row) is where the temperature trend has been weak or negative. However, many of the regions of negative MSSS referenced against climatology is where the conditional bias (Figures 2&3, right), is large; these are areas where the strength of the model response is too large compared to the observations for a given correlation.

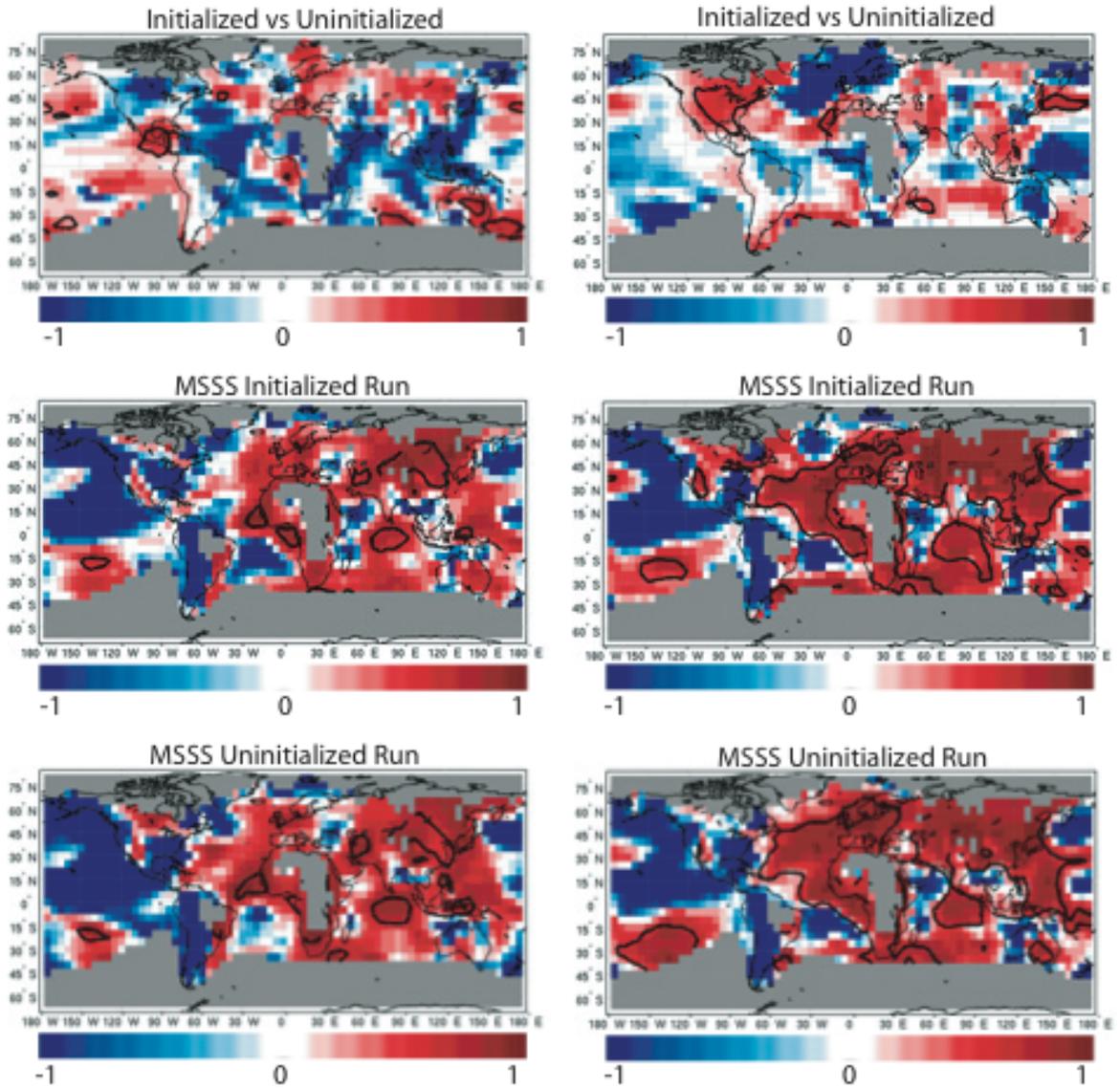


Fig. 1 Mean squared skill score (MSSS) for decadal temperature hindcasts from the DePreSys prediction system of the Hadley Centre (left) and the CanCM4 prediction system of the Canadian Climate Centre (right). Top row: MSSS comparing the initialized hindcasts (“forecasts”) and the uninitialized hindcasts (“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 1961-2006 (*i.e.* 10 hindcasts). Contour line indicates statistical significance that the MSSS is positive at the 95% confidence level.

The MSSS of the initialized hindcasts referenced to the uninitialized ones shows that areas of improved skill due to initialization differ between the two models (Figures 1, top panels, the positive or red areas). For example, over the Atlantic the initialized DePreSys hindcasts for temperature improve over the uninitialized hindcasts in the North Atlantic, whereas in the CanCM4 temperature hindcasts the improvement is seen in the tropical Atlantic. That different prediction systems differ in where they are skillful is a common situation in seasonal-to-interannual prediction. It should also be noted that in the case of the Atlantic neither of these improvements are deemed statistically significant, which is shown by the heavy contour line enclosing the

positive skill areas. For the CMIP5 experimental design, very few cases are available and thus it cannot be ruled out that some differences of skill between the two systems result from sampling variability.

The MSSS for the precipitation hindcasts (not shown) are not significantly better than the reference forecast of climatology, anywhere. There are regions where the MSSS of the initialized hindcasts are significantly better than the uninitialized ones, but these areas are small, and though point-wise significant may still be related to the small sample size. Even in regions where improvement between the initialized and uninitialized hindcasts is seen, this improvement must be viewed together with the actual skill (*i.e.* relative to climatology) from the initialized hindcasts.

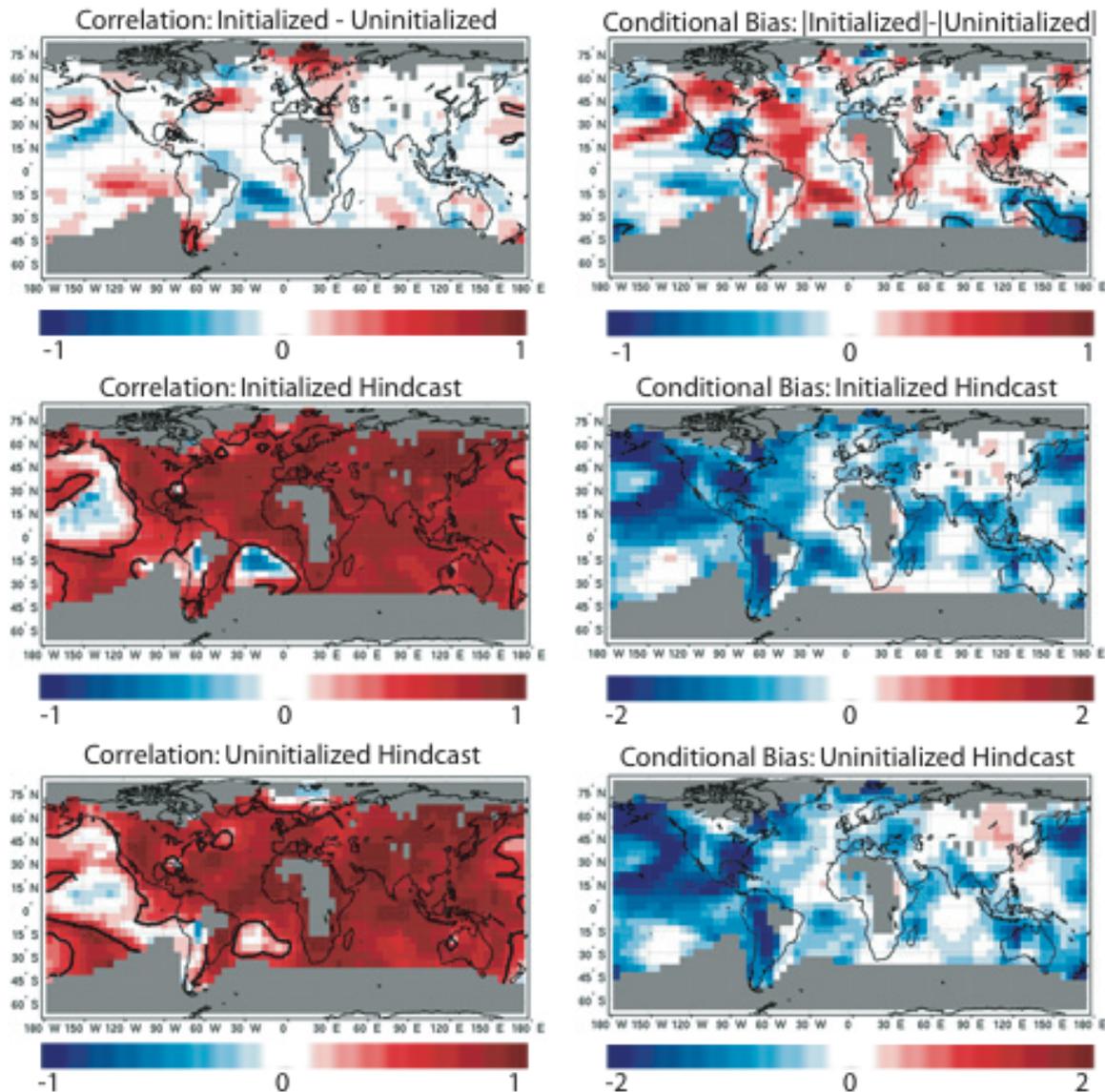


Fig. 2 Skill metrics related to MSSS decomposition for DePreSys temperature hindcasts. Left: Anomaly correlation coefficients with top row depicting the difference between the correlation of the initialized hindcasts (middle row) and that of the uninitialized hindcasts (bottom). Right: Conditional bias, with top row depicting the decrease 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 1961-2006 (*i.e.* 10 hindcasts). Contour line on the correlation maps indicates statistical significance that the value is positive at the 95% confidence level.

For temperature the areas where MSSS shows greater accuracy in the initialized hindcasts compared to the uninitialized (red areas, top of Fig. 1) comes from a reduction in conditional bias (blue areas, Figs. 2&3 right) rather than an increase in correlation (red areas, Figs. 2&3 left). This suggests that at least for this forecast target and in these prediction systems, increased accuracy is not due to the capture of signals in climate variability. For precipitation, improvements may come from both increased correlation and decreased bias, though again the improvements are not typically associated with forecasts that are skillful in their own right.

Probabilistic (Q2):

The CRPSS of the temperature hindcasts show very similar patterns (not shown) whether one estimates the uncertainty in a given forecast from the average ensemble spread or the standard error of the mean. Negative CRPSS values dominate the comparative metric that tests the uncertainty from the ensemble members against the uncertainty from the standard error. This indicates that the use of the ensemble spread leads to less reliable forecasts. The CRPSS for the precipitation forecasts yields very similar results.

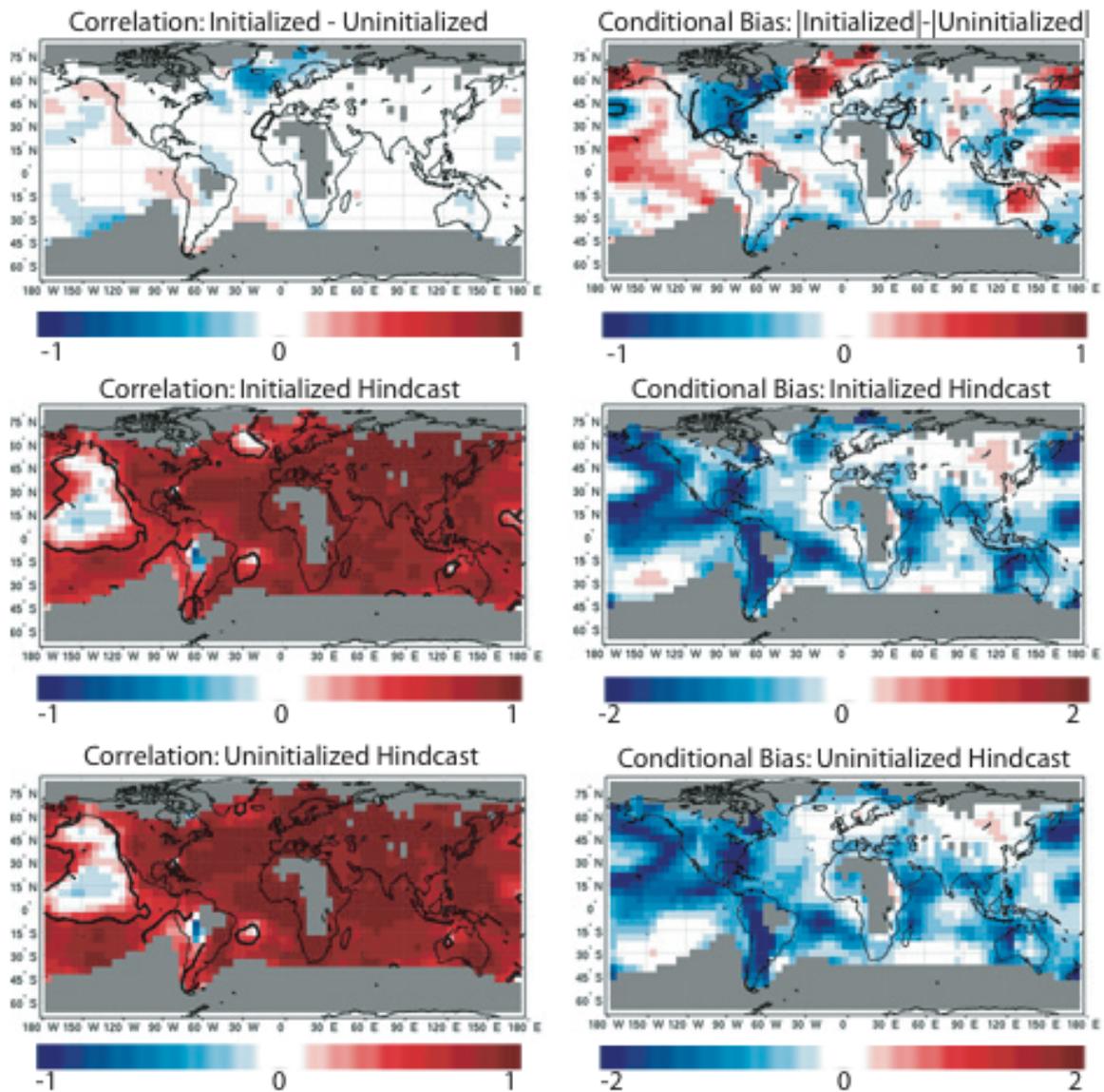


Figure 3. Same as Figure 4, but for CanCM4 hindcasts.

4. Concluding remarks /discussions

US CLIVAR Working Group on Decadal Predictability has developed a framework for verification of decadal hindcasts that allows for common observational data, metrics, temporal structure, spatial scale, and presentation. The framework addresses specific questions of the hindcast quality and offers suggestions for how they might be used. Considerable complementary research has aided this effort in areas of bias and forecast uncertainty, spatial scale of the information, and stationarity impacts on reference period.

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 general, precipitation is a more localized variable in both space and time, and thus subject to larger noise-like variability that is not predictable. Second, forecasts from different prediction systems often differ in where they perform well. Some common areas of good and poor performance are seen in both prediction systems. However, many differences exist as well, especially for precipitation, and also for the impact of initialization.

Although these results may be sobering, they should not be viewed as a conclusion that there is no decadal predictability. Decadal prediction is very much an experimental activity, including how best to initialize the predictions. 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. Those interested in these predictions should also visit the DPWG verification website to examine whether other time horizons might have more useable information. Even the more detailed paper that outlines the framework (Goddard *et al.* 2012) 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. It is also possible that gains in prediction quality may be made by multi-model ensembling, as has been realized for seasonal prediction. Preliminary results based on just the two models used in this study show mixed results (not shown). Statistical post-processing, or calibration, of model predictions may also improve forecast quality. However, to do that robustly will require larger ensemble sizes and more forecast cases (*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 should be expected to improve the quality of decadal-scale climate information.

5. References

- Arora, V., and Co-authors, 2011: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophys. Res. Lett.*, **38**, L05805, doi:10.1029/2010GL046270.
- Goddard, L., and Co-authors, 2012: A verification framework for interannual-to-decadal prediction experiments, *Clim. Dyn.*, submitted.
- Merryfield, W. J., and coauthors, 2011: The Second Coupled Historical Forecasting Project (CHFP2): I. Models and Initialization. In preparation.
- Murphy, A.H., 1988: Skill scores based on the mean squared error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417-2424.
- Räisänen, J., and J.S. Ylhäisi, 2011: How much should climate model output be smoothed in space? *J. Climate*, **24**, 867-880.
- Smith, D. M., and Co-authors, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nature Geoscience*, doi: 10.1038/NGEO1004.
- Solomon, A., and Co-authors, 2010: Distinguishing the roles of natural and anthropogenically forced decadal climate variability: Implications for prediction. *Bull. Amer. Meteor. Soc.*, doi: 10.1175/2010BAMS2962.1.
- Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2011: An overview of CMIP5 and the experiment design, *Bull. Amer. Meteorol. Soc.*, doi: <http://dx.doi.org/10.1175/BAMS-D-11-00094.1>.