A Verification Framework for Interannual-to-Decadal Prediction Experiments

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

The International Research Institute for Climate and Society, Earth Institute, Columbia University. *goddard@iri.columbia.edu

ABSTRACT

One of the two main objectives of the US CLIVAR Working Group on Decadal Predictability (DPWG) is to develop a metrics framework for verification of the decadal hindcast experiments will be part of the Fifth Assessment of the IPCC. Many outside the climate community are eager to use the hindcasts and forecasts for impacts studies and sectoral forecasts that use climate data as an input. However, the climate community is still investigating how to produce and assess the predictions that may or may not contain information on climate variability in addition to climate change. It is therefore crucial that there be a coordinated assessment of the prediction skill of these experiments that can guide their use.

The purpose for coordinated verification is twofold. The primary reason is to make the skill assessments across forecast systems comparable, in terms of which observations are used for verification, what period(s) are used, and how that information is displayed. This will certainly not be the only forecast verification work done by the centers, or the scientific community, but it serves as a minimum set of metrics.

DATA

Canadian Centre Climate Hindcasts:
CanCM4(T42 resolution); Full-field initialization; 9 ensemble members
CMIP5 design for start dates = 10 cases (every 5 years, starting/end 1960/beg. 1961)

Hadley Centre Hindcasts
DePreSys (3.75x2.5 deg resolution); Anomaly initialization; 9 ensemble members
Start dates for every year, but using CMIP5 design for start dates

Observations:
Air Temperature: Hadley Centre/CRUT3v; available on a 5° longitude by 5° latitude grid
Precipitation: GPCCv4; available at a resolution of 2.5° longitude by 2.5° latitude grid

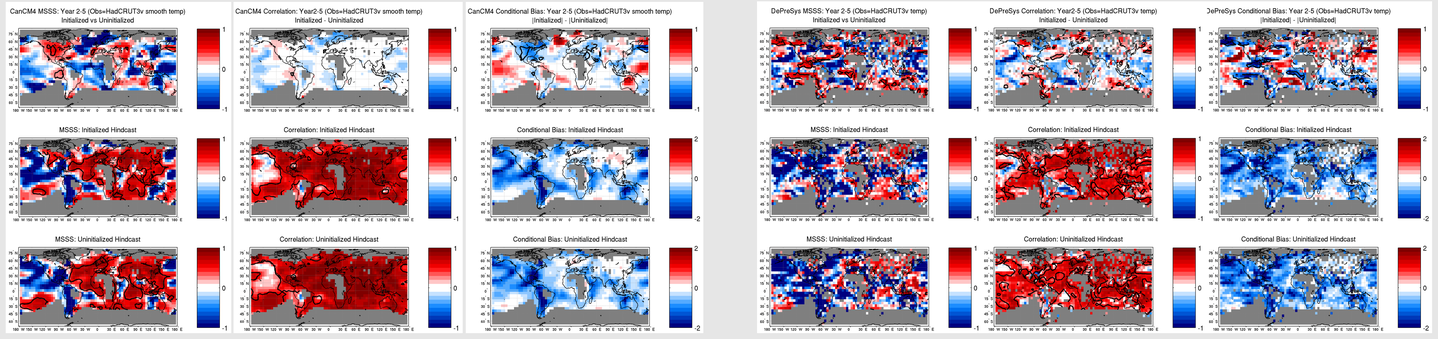
Q1: Do the initial conditions in the hindcasts lead to more accurate predictions of the climate?

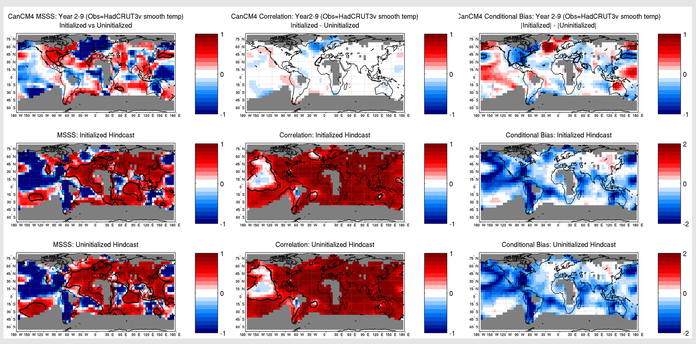
Deterministic

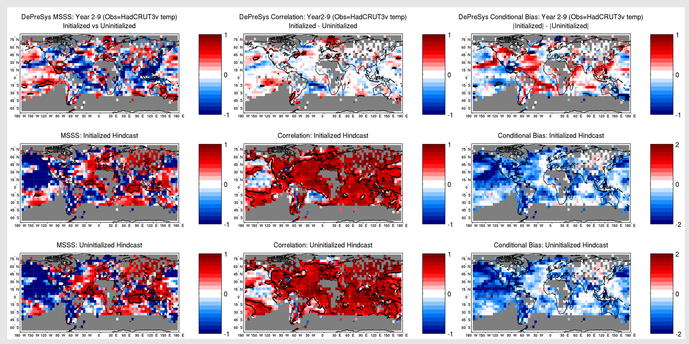
This question, regarding whether the initial conditions provide an improved signal and thus greater accuracy in the predictions, can be addressed using deterministic metrics. We advocate the use of the mean squared skill score (MSSS) and its decomposition following Murphy (1988). The MSSS is based on the mean squared error (MSE) of the forecast under test, Y, which is the initialized hindcasts, the observations, X, and of the reference forecast, W, which is the initialized hindcasts, and of the reference forecast which is the uninitialized hindcasts. All hindcasts considered have been corrected for mean bias (ICPO 2011). Therefore, it can be shown that the form of the MSSS relevant to the decadal verification framework can be written as:

$$MSSS(Y,W,X) = 1 - \frac{MSE_Y}{MSE_W} \qquad MSSS(Y,W,X) = \frac{r_{XY}^2 - \left[r_{XY} - \left(s_Y/s_X\right)\right]^2 - r_{XW}^2 + \left[r_{XW} - \left(s_W/s_X\right)\right]^2}{1 - r_{XW}^2 + \left[r_{XW} - \left(s_W/s_X\right)\right]^2} \qquad MSSS(Y,W,X) = \frac{MSSS_Y - MSSS_W}{1 - MSSS_W}$$

The MSSS is a summary metric; it combines: (1) the square of the correlation coefficient, and (2) the square of the conditional forecast bias. If the MSSS is positive, it indicates that the initialized hindcasts are more accurate than the uninitialized hindcasts. The maximum value of MSSS is 1.0, but it is not bounded on the negative skill side. Given the role of the correlation and the conditional bias in determining the MSSS, those deterministic metrics are presented as well. The Pearson's correlation coefficient that is used here is the linear association between the forecast mean and the observations. As such, it gives a measure of potential skill because the translation between the forecast value and the observed value must still consider the biases inherent in the forecasts. The conditional bias is related to the regression line of the observations (given the forecasts) and the forecasts. Given a positive correlation, negative values of the conditional bias generally indicate that the slope of the regression is larger than one, and smaller observed values would be expected relative to that indicated by the forecast mean.







Q2: Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?

Probabilistic

In addition to establishing the level of accuracy in the ensemble mean forecast, one is often interested in quantifying the range of possibilities or uncertainty about that forecast value. The purpose of the probabilistic metrics here is not to ascertain skill of the forecast relative to the uninitialized projections; that would be largely redundant information to that of the deterministic metrics. Here we pose the question of whether the ensemble spread in the forecast is, on average, adequate to represent the forecast uncertainty. Again, a skill score is used to determine the probabilistic quality of the forecast spread from the ensemble members relative to some reference approach. The measure of probabilistic quality is the cumulative ranked probability score, which is analogous to the MSE for probabilistic forecasts. By definition, the CRPS is:

$$CRPS(Y_j, X_j) = \int_{-\infty}^{\infty} \left(G(Y_j) - H(X_j)\right)^2 dy$$

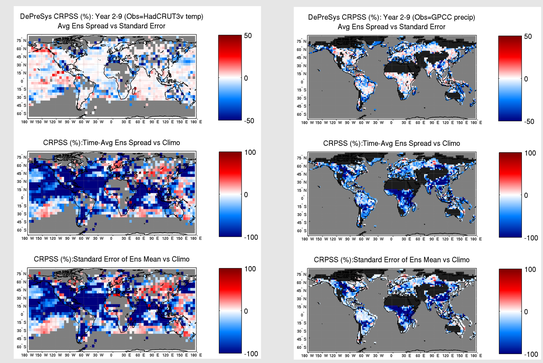
where G and H represent the cumulative distribution functions of the forecast and the observations, respectively. In this case, where X, represents the observations, the cumulative function H is the Heavyside function. If the predictive distribution is a Gaussian with mean Y and variance σ², then it follows that (Gneiting and Raftery, 2007):

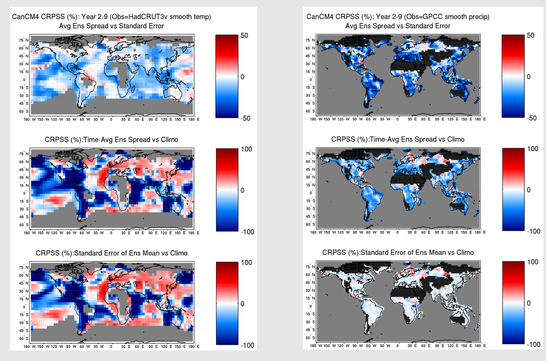
$$CRPS\left(N\left(\hat{Y}_j, \sigma_{\hat{Y}_j}^2\right), X_j\right) = \sigma_{\hat{Y}_j}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{X_j - \hat{Y}_j}{\sigma_{\hat{Y}_j}}\right) - \frac{X_j - \hat{Y}_j}{\sigma_{\hat{Y}_j}}\left(2\Phi\left(\frac{X_j - \hat{Y}_j}{\sigma_{\hat{Y}_j}}\right) - 1\right)\right]$$

where Φ and Φ represent the probability distribution function (pdf) and cumulative distribution function (cdf) of a standard Gaussian variable. Note that the forecast value is not necessarily identical to Y_j used for the deterministic metrics, which has had only the mean-bias removed. is the ensemble mean forecast value that has been corrected for the conditional bias, as diagnosed through the deterministic metrics. The slope of the regression line between the observations (given the forecasts) and the forecasts is $(s_Y/s_X)r_{YX}$, which is the scaling used to correct the forecasts for the conditional bias. Given the $CRPS_Y$ for the forecast distribution, and the $CRPS_W$ for the reference distribution, the corresponding skill score can be defined as:

$$CRPSS = 1 - \frac{\sum_{j=1}^{n} CRPS_{Y_j}}{\sum_{j=1}^{n} CRPS_{X_j}}$$

The "forecast" distribution is assumed Gaussian, with the mean given by the corrected ensemble mean and the variance given by the average ensemble variance (i.e. averaged over all hindcasts). Since we are only testing the uncertainty in the forecasts, the mean of the distribution is the same for both the forecast under test and the reference forecast (i.e. $W_j = Y_j$). The "reference" distribution has a variance given by the standard error variance of the hindcasts' ensemble mean compared to the observations.





Discussion

A framework for verification of interannual-to-decadal predictions has been described and illustrated for two prediction systems and for a specific forecast target of multi-year averages of annual means. The framework is not exhaustive, nor is it intended to be. 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 can be quantified. Equally important, the framework provides information on forecast quality across prediction systems that puts the verification of each on equal footing – observational verification data, verification period, spatial and temporal averaging, and even graphical presentation – such that relative comparisons can be 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 correction of mean and conditional biases, and consideration of how to best approach forecast uncertainty.

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**, and the dominant signal is due to the upward trends, which are captured reasonably well by both systems over most of the world. However, there are **large conditional biases** that suggest caution in using the model data directly. 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. 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. Additionally, more improvement is seen in the prediction systems if we consider all start years (i.e. Hadley Centre hindcasts, not shown) rather than the CMIP5 nominal design of initial conditions taken ever 5 years. 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). 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.

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

References

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, C. Deser, S. Mason, B. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, S. Kharin, W. Merryfield, G. Hegerl, C. Ferro, D. Stephenson, G.A. Meehl, T. Stockdale, R. Burgman, A. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori[14], D. Vimont, T. Delworth. 2011: A verification framework for interannual to decadal prediction experiments. Clim. Dyn., submitted.

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

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

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.

The authors of this work are members of the Decadal Predictability Working Group sponsored by U.S. CLIVAR. We appreciate the support from the U.S. CLIVAR office. Goddard, Gonzalez and Greene received funding from a NOAA grant (NA08OAR4320912) for work on this project.

VISIT DPWG VERIFICATION PAGE: HTTP://CLIVAR-DPWG.IRI.COLUMBIA.EDU