Metrics and Protocols for Evaluating NCEP Climate Models and Forecasts

July 26, 2013

This document provides NCEP metrics for evaluating improvements in climate model performance to NCEP operational Climate Forecast System version 2 (CFSv2) and for evaluating forecast tools (either statistical or dynamical) compared to Climate Prediction Center (CPC) operational forecasts or some other common baselines (such as, persistence). The forecast metrics include deterministic, categorical, and probabilistic measures. The model performance metrics are intended to include measures of model performance in simulating mean climate, variability (i.e., ENSO, MJO), and key physical processes (e.g., convection, surface fluxes). Researchers or tool developers can select appropriate metrics, variables, or hindcast periods from this document accordingly.

1. Climate model performance metrics

The metrics are designed along following four sets of diagnostics during the model development and evaluation process. The proposed diagnostics includes two aspects (a) characteristics of climate variability in free simulations (both atmosphere alone and coupled), and (b) initialized weather and seasonal predictions. The strategy behind model evaluation is to develop a hierarchical evaluation procedure with appropriate decision points to better utilize available resources.

- **AMIP (~30 year run)** to determine basic characteristics of mean climate and its variability on various time scales
  - Mean and biases (u, v, T, q, SST, soil moisture, snow, sea ice, precipitation, surface fluxes,…)
  - Characteristics of atmospheric modes of variability (PNA, NAO, MJO, …)
  - ENSO tele-connection

- **CMIP (~30 year run)**
  - Mean and biases (u, v, T, q, SST, soil moisture, snow, sea ice, precipitation, surface fluxes,…)
  - Characteristics of atmospheric modes of variability (PNA, NAO, MJO, …)
  - ENSO variability
  - ENSO tele- connection

- **Short to medium range** weather forecast performance evaluations
  The performance of model upgrades in the atmosphere component of CFS (e.g., GFS) on short to medium range weather forecasts is assessed using the Global NWP Model Verification Package developed at NCEP EMC. The evaluation requires two years of retrospective medium-range forecasts. The main metrics include
  - Anomaly correlations, biases, RMSE (u, v, T, P, SLP, q, cloud)
  - Hurricane track and intensity errors

- **Initialized climate prediction evaluations** (October and May initial conditions) The performance of model upgrades in the CFS monthly and seasonal forecast will be assessed based on hindcast over selected periods.
    - Skill of SSTs
    - ENSO skill
    - Surface temperature and precipitation skill

2. Climate forecast evaluation metrics

The forecast evaluation metrics are for evaluations after the model development is finalized and hindcasts over the appropriate period are available. The metrics include different aspects of verifications including error/accuracy, skill scores, conditional statistics, reliability, and biases for three types of forecasts -- deterministic, categorical, and probabilistic. The metrics are consistent with those used in the CPC real-time verification system and are used at many places elsewhere, for example, the WMO Standardized Verification System for Long-Range Forecasts (http://www.bom.gov.au/wmo/lrfvs/).

- Deterministic/Continuous Forecasts:
    - Anomaly Correlation (AC)
    - Root Mean Square Error (RMSE)
    - Mean Absolute Error
    - Amplitude
    - Biases
- Categorical:
    - Contingency Table
    - Heidke Skill Score
- Probabilistic:
    - Brier Skill Score (BSS)
    - Rank Probability Skill Score (RPSS)
    - Reliability

3. Climate Forecast and Hindcast Evaluation Protocol

The hindcast requirements mainly follow the current NMME protocol that has a focus on monthly/seasonal time scales and needs to be modified for subseasonal forecast (see Appendix). It should be noted that the predictability and scientific basis for subseasonal forecasts need to be evaluated. For testing use of new forecast tools in NCEP/CPC operations, developers are required to provide hindcast data to CPC/CTB for systematic evaluations. Hindcast start times must include all 12 calendar months. Real-time prediction system must be identical to the system used to produce hindcast

- Hindcast period:
    - Monthly and seasonal forecasts: minimum 30 years (1982-2012)
    - Subseasonal forecasts: minimum 10-15 years (preferably during 1999-2012)
- Forecast lead time:
    - Monthly and seasonal forecasts: 1- 9 months
    - Intra-seasonal forecasts: 1- 32 days (under discussion)

- Number of ensemble members: To be decided by the tool developer
- Basic data:
  - Monthly and seasonal forecasts: Monthly mean of T2m, Prate, Z200 and SST
  - Subseasonal forecasts: Daily data of OLR, U and V at 250 and 850 hPa, Z500, Z700, T2m, Prate
- Data distributed must include each ensemble member and total uncorrected fields.
- Data format: Grid 1x1
- Domain: Global

4. Contact Information

NCEP Climate Test Bed Director, Jin Huang (Jin.Huang@noaa.gov)

Appendix: NMME Hindcast and Real-time Experimental Prediction Protocol

The CY2011 NMME experimental predictions have been made in real-time since August 2011. As part of the development of the real-time capability, the NMME partners agreed on a hindcast and real-time prediction protocol. Some of the key elements of this protocol include:

- Real-time ISI prediction system must be identical to the system used to produce hindcasts. This necessarily includes the procedure for initializing the prediction system. The number of ensemble members per forecast, however can be larger for the real-time system.

- Hindcast start times must include all 12 calendar months, but the specific day of the month or the ensemble generation strategy is left open to the forecast provider.

- Lead-times up to 9 months are required, but longer leads are encouraged.

- The target hindcast period is 30 years (typically 1981-2010).

- The ensemble size is left open to the forecast provider, but larger ensembles are considered better.

- Data distributed must include each ensemble member (not the ensemble mean). Total fields are required (i.e., systematic error corrections to be coordinated by MME combination lead, NOAA/CPC). Forecast providers are welcome to also provide bias-corrected forecasts and to develop their own MME combinations.

- Model configurations – resolution, version, physical parameterizations, initialization strategies, and ensemble generation strategies – are left open to forecast providers.

- Required output is monthly means of global grids of SST, T2m, and precipitation rate. More fields will be added based on experience and demand. It is also recognized that higher frequency data is desirable and this will be implemented as feasible.

- Routine real-time forecast data must be available by the 8th of each month.