

## Homogeneous and Heterogeneous Predictability and Forecast Skill in MME

Huug van den Dool<sup>1</sup>, Emily Becker<sup>1</sup> and Malaquias Pena<sup>2</sup>

<sup>1</sup>Climate Prediction Center, NCEP/NWS/NOAA, College Park, MD

<sup>2</sup>IMSG at Environmental Modeling Center, NCEP/NWS/NOAA

### 1. Introduction

Forecast skill and potential predictability of 2 m temperature are assessed using hindcast data from Phase 1 of the National Multi-Model Ensemble (NMME) project. Forecast skill was examined using the anomaly correlation (AC) of the ensemble mean (EM) of an individual model forecast against the observed value. Predictability was considered from two angles: homogeneous, where one model is verified against a single member from its own ensemble, and heterogeneous, where a model's EM is compared to a single member from another model. This study provides insight both into the NMME and its contributing models and into the physical predictability of the 2 m temperature field.

### 2. The National Multi-Model Ensemble project

The NMME is a forecasting system consisting of coupled models from U.S. and, more recently, Canadian modeling centers. The multi-model ensemble approach has been proven to produce better prediction quality than any single model ensemble, motivating the NMME undertaking. The environmental variables included

Hindcast Situation YEAR 1							Model resident Resolutions		
	Start months available NOW	Period	Members	Arrangement of Members	Lead (months)	Atmosphere	Ocean	Reference	
NCEP-CFSv1	12	1981-2009	15	1 <sup>st</sup> 0Z +/-2days, 11 <sup>th</sup> 0Z +/-2d, 21 <sup>st</sup> 0Z +/-2d	0-9	T62L64	MOM3L40 0.30 deg Eq	Saha et al 2006	
NCEP-CFSv2	12	1982-2010	24(28)	4 members (0,6,12,18Z) every 5th day	0-9	T126L64	MOM4 L40 0.25 deg Eq	Saha et al 2012	
GFDL-CM2.1	12	1982-2010	10	All 1st of the month 0Z	0-11	2x2.5deg L24	MOM4 L50 0.25 deg Eq	Delworth et al 2006	
IRI-Echam4-f	12	1982-2010	12	All 1st of the month	0-7	T42L19	MOM3 L25 0.5 deg Eq	DeWitt MWR2005	
IRI-Echam4-a	12	1982-2010	12	All 1st of the month	0-7	T42L19	MOM3 L25 0.5 deg Eq	"	
NCAR-CCSM3.0	12	1982-2010	6	All 1st of the month	0-11	T85L26	POP L40 0.3 deg Eq	Kirtman and Min 2009	
NASA	12	1981-2010	6	1 member every 5th day as CFSv2	0-9	1x1.25deg L72	MOM4 L40 0.25 deg Eq	Rienecker et al 2008	

**Table 1** All models included in the National Multi-Model Ensemble project, year 1 of Phase 1 (August 2011 – August 2012).

in the first year of Phase I (Aug. 2011 – July 2012) were 2m surface temperature, SST, and precipitation rate; real-time and archived forecast graphics from Aug. 2011 – present are available at [www.cpc.ncep.noaa.gov/products/NMME](http://www.cpc.ncep.noaa.gov/products/NMME). Hindcast and forecast data is archived at the International Research Institute for Climate and Society (IRI), accessible from the NMME homepage. Table 1 lists the models included in the 1<sup>st</sup> year of Phase 1. All model outputs have 1.0° latitude by 1.0° longitude resolution and forecast leads of 1 – 7 months. 29 years of hindcasts (1982-2010) were available for all models except CFSv1 (28 years: 1982-2009). Model real-time forecasts are produced by no later than the 8<sup>th</sup> of each month, and graphical forecasts are available on the 9<sup>th</sup> of each month. Phase I forecasts were all delivered on time in year 1.

### 3. Forecast skill and predictability

This study assessed prediction skill, homogeneous predictability, and heterogeneous predictability for the 29 years of hindcasts for all models, using the anomaly correlation (AC). The AC is a measure of the association between the anomalies of (usually) gridpoint forecast and observed values (Wilks 1995, van den Dool 2007). By “prediction skill”, we mean one model’s ensemble mean (EM) forecast versus the observed value. The verification field for 2 m temperature (T2m) is the station observation-based GHCN+CAMS (Fan and van den Dool 2008). GHCN+CAMS has a native resolution of 0.5° latitude x 0.5° longitude, and was regridded to 1.0° x 1.0° for this study.

One common method of defining potential predictability, *i.e.* the physical extent to which a parameter can be predicted under the best of circumstances, is to evaluate one model forecast versus another (Lorenz 1982). Hence, we are testing how effective the model is at predicting itself, and therefore the limit of predictability, if we assume the model is a replica of reality. In this context, we apply the so-called ‘perfect’ model assumption, *i.e.* the forecast and proxy-observation are taken from the same world and there are no systematic errors to be corrected. Homogeneous predictability assesses one model’s EM, based on N-1 members, against the one member that is left out (the proxy-observation). Heterogeneous predictability refers to one model’s EM (based on all N members) versus one member of another model.

All values in Table 2 represent the 29-year hindcast timeseries. Leads 1-3, and all 12 start months are combined, and the area-averages are done over all land north of 23°N and south of 75°N. Model anomalies are relative to each model’s individual climatology (from the 29-year reforecast). No cross-validation is applied for the prediction skill calculations, which may be a problem, except for homogeneous predictability.

### 4. Results

The following discussion refers to Table 2. First, some comments on the size of the anomalies. The standard deviation (sd; bottom row) of all models (*i.e.* from individual model runs) agrees very well with the observations; all are in the range of 2.0° - 2.4°C. This is high praise, and different from earlier impressions (mainly from Demeter) that models are underdispersive. The sd of

TMP2m Northern Hemisphere Leads 1-3										
	cfsv1	cfsv2	echa ma	echa rrf	gfdl	nasa	ncar	obs (EM skill)	EM RMSE (C)	EM SD
cfsv1 EM	0.19	0.08	0.05	0.06	0.07	0.09	0.04	0.06	2.07	0.81
cfsv2 EM	0.09	0.27	0.09	0.08	0.16	0.19	0.01	0.19	1.98	0.77
echama EM	0.04	0.08	0.15	0.16	0.08	0.08	0.05	0.08	2.06	0.76
echamf EM	0.06	0.07	0.16	0.15	0.08	0.08	0.05	0.07	2.07	0.76
gfdl EM	0.06	0.14	0.07	0.06	0.25	0.15	0.01	0.15	2.08	1.05
nasa EM	0.07	0.14	0.07	0.05	0.15	0.27	0.00	0.14	2.06	0.93
ncar EM	0.03	0.01	0.04	0.04	-0.01	0.00	0.12	-0.01	2.24	1.07
single mem & obs SD	2.28	2.09	2.13	2.09	2.37	1.99	2.26	2.14		

**Table 2** Anomaly correlations showing 2 m temperature forecast skill (model ensemble mean (EM) verified against observations), homogeneous predictability (model EM of N-1 members verified against remaining member), and heterogeneous predictability (model EM verified against a single member of another model) for the seven models in NMME Phase 1. ACs are aggregated over leads 1-3, all start months, land 23°N – 75°N. Also shown are standard deviation for each model’s EM (EM SD, right column) and a single member (single mem & obs SD, lower row).

the model ensemble means (EM; far right column) is about 0.75° to 1.0°C, which is appropriately smaller than the sd of either the individual ensemble members or the observations. This decrease in sd follows from damping of the noise (leaving mainly the signal in the EM) by  $1/\sqrt{N}$ , where N is the number of (effectively) independent ensemble members. Models with higher N have a greater reduction of their sd.

Prediction skill, measured here by the AC (blue column), varies from -0.01 to +0.19. These are modest numbers, but +0.19 is highly significant because it is based on a huge sample. (Also, a trustworthy +0.19 tends to correspond to large areas and many targets times with little or no skill at all and a few areas and limited target times with much higher skill.) The homogeneous predictability (the yellow diagonal) ranges from 0.12 to 0.27. This is higher than the reported skill ( $\leq 0.19$ ), which means that we can do better eventually, but not hugely so. It is probably disappointing that among the seven independent opinions about predictability, none is better than 0.27, leaving not much to pick from. The heterogeneous predictability ranges from 0.0 to 0.19, exactly the range of skill already achieved. Heterogeneous predictability and the actual skill suffer equally from a mismatch in climate between model and verification – only homogeneous predictability estimates are based with justification on a perfect model assumption. The days when models predicted each other better than they predict reality appear to be over (at least for monthly means at long lead). In summary, all models predict themselves better than they predict other models (or reality).

Regarding heterogeneous predictability (black off-diagonal elements), we note that Table 2 is largely symmetric. This means that “to predict” and “to be predicted” is similar, *i.e.* if Model A (EM) can predict Model B (single member), then the reverse is also true. Curiously, the NCAR model is exceptional in a way: it has a hard time to predict anomalies in the other models, or to be predicted by the other models. In and of itself we appreciate orthogonal behavior, but since NCAR also poorly verifies against the observations, its orthogonality may be erroneous. The trio GFDL, NASA, CFSv2 correlates the most to each other, and have the highest skill against observations. That raises (unaddressed) questions of redundancy. The two IRI models predict each other almost as well as they predict themselves. These models are in fact the same, and our decision to treat the fully coupled ensemble of 12 and the anomaly coupled ensemble of 12 as two separate ensembles may be debatable. (At IRI the two sets were merged into one ensemble.) The CFSv1 and CFSv2 have a shared pedigree obviously, but in contrast to the IRI, the two NCEP model do not predict each other very well.

We finally note that prediction skill is low at present for NCAR and the two IRI-ECHAM models, and this is in part because these models have only ocean initialization, *i.e.* their atmosphere and land initial state is random and unlikely to be realistic. This impacts skill of predicting T2m negatively. CFS, GFDL and NASA attempt to have a realistic atmosphere and land initial state, in addition to a realistic initial ocean.

*Acknowledgements:* Other team members: Suranjana Saha, Peitao Peng; all data suppliers (NCAR, GFDL, NASA, IRI *etc.*), funding agents (CPO *etc.*).

## References

- Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948-present. *J. Geophys. Res.*, **113**, D01103, doi:10.1029/2007JD008470.
- Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **23**, 243-254.
- Van den Dool, H., 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 215 pp.
- Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.