A multiple model assessment of seasonal climate forecast skill for applications

David Lavers,[1,2,3] Lifeng Luo,[1,4] and Eric F. Wood[1]

Received 15 October 2009; revised 12 November 2009; accepted 17 November 2009; published 15 December 2009.

[1] Skilful seasonal climate forecasts have potential to affect decision making in agriculture, health and water management. Organizations such as the National Oceanic and Atmospheric Administration (NOAA) are currently planning to move towards a climate services paradigm, which will rest heavily on skilful forecasts at seasonal (1 to 9 months) timescales from coupled atmosphere-land-ocean models. We present a careful analysis of the predictive skill of temperature and precipitation from eight seasonal climate forecast models with the joint distribution of observations and forecasts. Using the correlation coefficient, a shift in the conditional distribution of the observations given a forecast can be detected, which determines the usefulness of the forecast for applications. Results suggest there is a deficiency of skill in the forecasts beyond month-1, with precipitation having a more pronounced drop in skill than temperature. At long lead times only the equatorial Pacific Ocean exhibits significant skill. This could have an influence on the planned use of seasonal forecasts in climate services and these results may also be seen as a benchmark of current climate prediction capability using (dynamic) couple models. **Citation:** Lavers, D., L. Luo, and E. F. Wood (2009), A multiple model assessment of seasonal climate forecast skill for applications, *Geophys. Res. Lett.*, *36*, L23711, doi:10.1029/2009GL041365.

1. Introduction

[2] Seasonal climate prediction is based on the premise that the lower-boundary sea surface temperature (SST) forcing, which evolves slowly, imparts predictability on atmospheric development [*Palmer and Anderson*, 1994]. In particular persistent SST anomalies associated with the El Niño Southern Oscillation influence atmospheric circulation, thus producing seasonal climate anomalies [*Carson*, 1998; *Stockdale et al.*, 2006]. Operational climate forecast centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF) and NOAA's National Centers for Environmental Prediction (NCEP) are now using coupled atmosphere-land-ocean models to produce their seasonal forecasts [*Palmer et al.*, 2004; *Saha et al.*, 2006]. Integrating coupled atmosphere-land-ocean models with an ensemble of different initial conditions allows predictions that consider uncertainty in the initial state, resulting in what is referred to as an ensemble forecast. Seasonal climate forecasts can be incorporated into end-user application models for determining crop yield amounts [*Cantelaube and Terres*, 2005; *Challinor et al.*, 2005] and future epidemic malaria [*Thomson et al.*, 2006]. Retrospective forecast (hindcast) datasets, such as those from the DEMETER project, give the opportunity to assess the predictive skill in current seasonal climate forecast models.

[3] Forecast quality in its complete sense can be assessed using a distributions-oriented framework [*Murphy*, 1993]. This approach uses the joint distribution of the forecasts (f) and observations (o) as this contains all of the non-time dependent information necessary for evaluating the forecast quality [*Murphy and Winkler*, 1987; *Murphy*, 1993]. For applications, one must determine the following: given a particular seasonal climate forecast, what is the conditional probability distribution of (future) seasonal climate $p(o|f)$. The extent to which the conditional seasonal distribution $p(o|f)$ varies from the climatological distribution $p(o)$ is an indication of the skill of the forecast. *Murphy and Winkler* [1987] refer to the factorization of the joint distribution into the conditional $p(o|f)$ and marginal $p(f)$ distributions as the 'calibration-refinement factorization'. Furthermore, this can also be done within a Bayesian framework that will spatially downscale and bias correct the seasonal climate forecasts, making them relevant for applications [*Luo et al.*, 2007].

[4] The predictability of 2-meter air temperature (hereafter, temperature) and precipitation is a multidimensional variable that can vary with geographical location (x, y), lead-time (τ), season (t) and with temporal (T) and spatial (L) scales. A thorough literature review of seasonal climate forecast quality assessment suggests a paucity of published papers on evaluation of monthly predictions, a fact also noted by *Weigel et al.* [2008]. To address this gap, we assess 1) the actual or realizable, and 2) the idealized predictability of monthly temperature and precipitation hindcasts from the NCEP Climate Forecast System (CFS) [*Saha et al.*, 2006] and seven models from the DEMETER project [*Palmer et al.*, 2004]. The analysis shows the current predictive capability in the "actual" and "model" climate systems.

2. Data and Methodology

[5] DEMETER was a European Union (EU) funded project that created a multi-model ensemble hindcast dataset containing seven models each with nine ensemble members. The models are from climate centers around Europe and their acronyms are: CERFACS, ECMWF, INGV, LODYC, METEO-FRANCE, MPI, and UKMO. The DEMETER models were initialized on 1st February, 1st May, 1st

[1]Environmental Engineering and Water Resources, Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.
[2]Centre for Ecology and Hydrology, Wallingford, UK.
[3]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK.
[4]Now at Department of Geography, Michigan State University, East Lansing, Michigan, USA.

0094-8276/09/2009GL041365$05.00

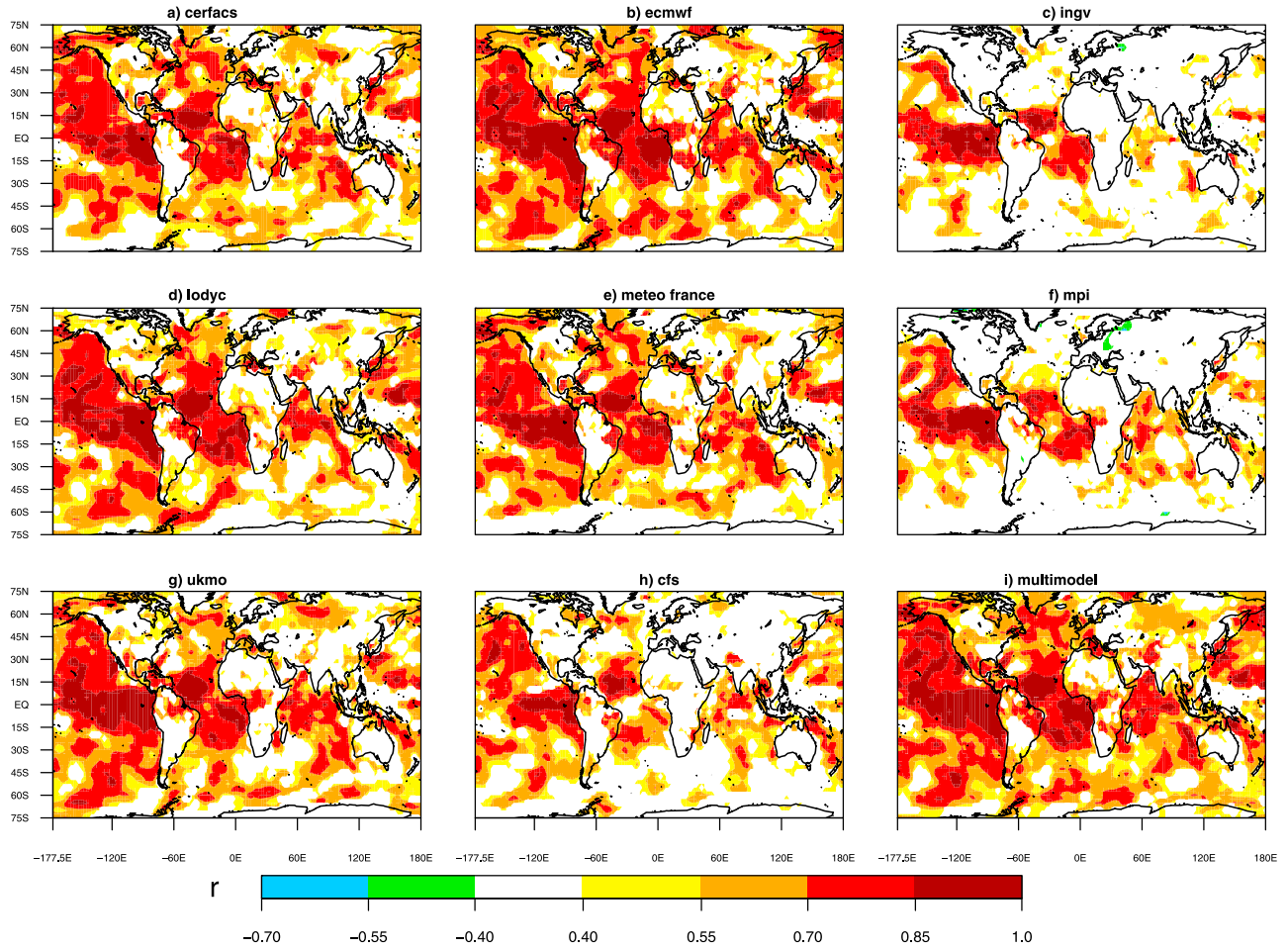

Figure 1. Actual predictive skill for each model grid point for 1981–2001 for a 30 day temporal average at a 1 day lead time for May two meter temperature forecasts for (a) CERFACS, (b) ECMWF, (c) INGV, (d) LODYC, (e) METEO FRANCE, (f) MPI, (g) UKMO, (h) CFS, and (i) the MULTI-MODEL. Non-white colors represent significant correlation r at the $p < 0.05$ level.

August and 1st November to assess the seasonal dependence of the hindcasts, and integrated for 180 days [*Palmer et al.*, 2004]. For the period being studied, CFS has 15 nine-month hindcasts initialized during each calendar month [*Saha et al.*, 2006]. The common period for the DEMETER and CFS models is 1981–2001 (21 years).

[6] Two meter air temperature at a 2.5° × 2.5° resolution from the ERA-40 re-analysis dataset [*Uppala et al.*, 2005] and monthly observed precipitation at 1.0° × 1.0° from the Global Precipitation Climatology Centre (GPCC) [*Rudolf et al.*, 2005] are used as the reference datasets. Precipitation was regridded to 2.5° resolution to match the model hindcasts' resolution.

[7] The joint probability distribution is computed between the model ensemble mean and observations using the operational hindcasts and observed climate outcomes. This joint distribution can be represented by a bivariate-Normal distribution [*Wilks*, 2006]. The conditional mean, $m(o|f)$, and variance, $\sigma^2(o|f)$, of $p(o|f)$ is $m(o|f) = m(o) + r\sigma(o)\frac{[f-m(f)]}{\sigma(f)}$ and $\sigma^2(o|f) = \sigma^2(o)(1 - r^2)$, where $m(o)$, $m(f)$, $\sigma(o)$ and $\sigma(f)$ are the means and standard deviations of the marginal distributions of $p(o)$ and $p(f)$ respectively, $\sigma^2(o)$ is the variance of the marginal distribution of $p(o)$ and r is the correlation between the forecast and resulting observation. Note that the conditional explained variance due to the forecast is reduced from the unconditional variance (climatology) in the climate variable by $r^2\sigma^2(o)$, which provides a measure of the information content from the seasonal forecast. A variety of skill scores could be used [*Wilks*, 2006], but we apply the product-moment correlation coefficient r between the observed climate and forecast ensemble mean series at a particular lead time and temporal average as it is central in determining the usefulness of seasonal forecasts for applications. Correlation represents a traditional summary measure between the forecasts and observations [*Murphy et al.*, 1989], and has been widely used in previous research [*Colman and Davey*, 1999; *Davies et al.*, 1997; *Folland et al.*, 2001; *Peng et al.*, 2000; *Van Oldenborgh et al.*, 2005; *Wu et al.*, 2009]. The methodology is applied for each model separately and for an equally-weighted (averaged) multi-model using all members from the eight models [*Hagedorn et al.*, 2005].

[8] The idealized predictability of a forecasting model is thought to be the upper limit of its predictive capability, where the forecast model and "climate" system have the same physics; that of the forecast model [*Koster et al.*, 2004]. This model estimate considers the spread (variance) of the ensemble members, which can be thought of as

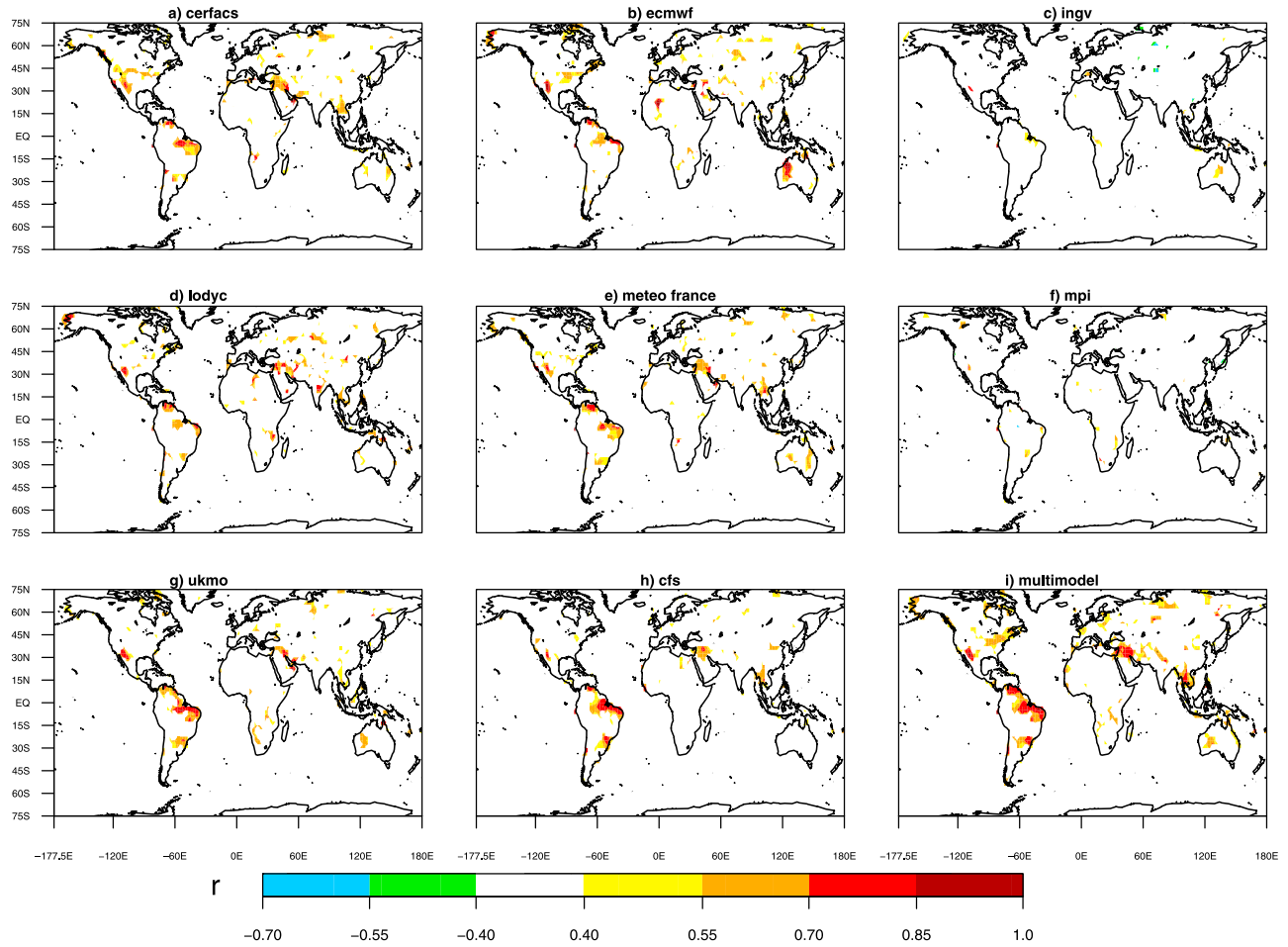

Figure 2. Actual predictive skill for each model grid point for 1981–2001 for a 30 day temporal average at a 1 day lead time for May precipitation forecasts for (a) CERFACS, (b) ECMWF, (c) INGV, (d) LODYC, (e) METEO FRANCE, (f) MPI, (g) UKMO, (h) CFS, and (i) the MULTIMODEL. (Key same as in Figure 1)

indicative of the predictive skill. If an ensemble has small (large) spread, then the forecast is likely to be insensitive (sensitive) to initial condition uncertainty, resulting in high (low) predictive skill [*Koster et al.*, 2004; *Tang et al.*, 2008]. The methodology used is done for each DEMETER and CFS model, and assumes that one member of the ensemble is the "truth" and that the remaining ensemble average is the "predictor". As before, r measures the linear association between the observed and predictor series. For DEMETER (CFS) this procedure is repeated nine (fifteen) times with each ensemble member in turn being considered as the truth. The nine (fifteen) values of r are averaged, which forms the final estimate of the system in predicting itself [*Koster et al.*, 2004].

3. Results

[9] The global actual predictive skill of temperature and precipitation for the eight models was calculated at the model grid scale; note that precipitation was only evaluated over the land masses. Figure 1 shows the realizable predictive skill of temperature for the eight models and multi-model for the first 30 day period (i.e., a 30 day temporal average at a 1 day lead time, or month-1) from 1st May. High predictive skill of $r > 0.70$ is generally confined to the oceans, especially over the equatorial Pacific and subtropical Atlantic. Skilful predictions over the equatorial Pacific in the ECMWF (Figure 1b) and UKMO (Figure 1g) models appear to be largely behind the multi-model skill (Figure 1i) in that region. Few models have noteworthy skill over land regions for the first 30 day forecast period.

[10] Figure 2 shows the realizable predictive skill of precipitation for month-1 from 1st May. Strikingly, there are very few grids with $r > 0.40$ (non-white areas), and there are fewer significant correlations over the land masses compared with temperature. Six out of eight models have significant skill over the Amazon basin, and all models have skill in the North American monsoon region. These two areas are also seen in the multi-model prediction. Figure 3 shows multi-model predictive skill of temperature and precipitation for month-1 and month-2 (second 30 day period) of May hindcasts. For month-1, high predictive skill of temperature over land ($r > 0.70$) is found over the Amazon basin, Congo basin, south-central Asia, central Europe and north–western and south–western North America. As lead time increases to 31 days (month-2 forecast), it is apparent that skilful temperature predictions reduce back to the tropics (Figure 3b) and little skill exists for precipitation (Figure 3d). The multi-model tends to improve the predictive skill over the individual models. In general the land masses have negligible skill at a 31 day

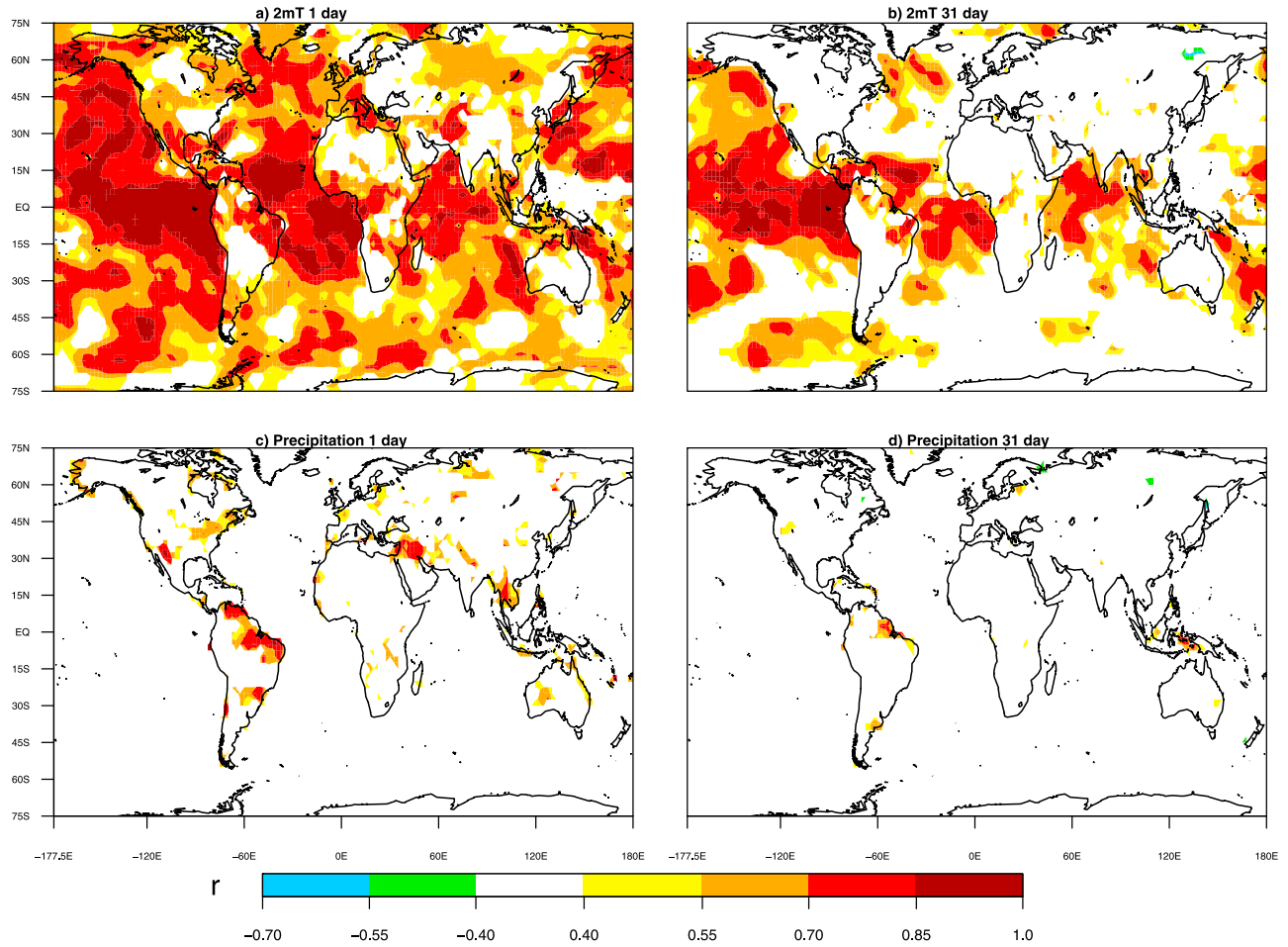

Figure 3. Multi-model forecasts for 1981–2001 for a 30 day temporal average at a 1 day lead time for May (a) two meter temperature and (c) precipitation. Multi-model forecasts for a 30 day temporal average at a 31 day lead time for May (b) two meter temperature and (d) precipitation. (Key same as in Figure 1.)

lead time, which is a relatively short lead time in terms of seasonal climate prediction.

[11] Figure 4 shows the global grid scale idealized predictive skill for May temperature hindcasts for the first 30 day period. Idealized predictive skill in the DEMETER models is higher than that seen for the real climate system. This is true for the land masses and oceans and is particularly noticeable for the models shown in the left panels of Figure 4. Low idealized skill in the extratropical regions in the CFS model could be due to the ensemble initialization, which produces members staggered throughout the month leading to members of varying ages. However, even with an ensemble of differing "initial" values, the members seem to forecast a similar climate state in the equatorial Pacific, which corroborates previous research by *Shukla* [1998]. Idealized predictive skill of May precipitation for month-1 (not shown) exhibits much less idealized predictive skill than for temperature. As the lead time increases only a narrow region of the equatorial Pacific has idealized skill (not shown). This significant decrease in idealized predictive skill, more so for precipitation than temperature, demonstrates yet again the chaotic nature of climate [*Lorenz*, 1963] and the possible futility of long-lead seasonal climate forecasting.

[12] It appears that the high idealized predictive skill evident during month-1 is attributable to the skill present in the first two weeks of the forecast when the spread of ensemble members is small. This is confirmed by calculating the idealized skill on the first and second 15 day averages, which shows a large drop off in predictive skill in the second of these 15 day periods (not shown).

4. Discussion

[13] This work has shown that limited realizable predictive skill of temperature and precipitation is found in the DEMETER and CFS seasonal climate forecasting models. Globally for 30 day temporal averages the skill deteriorates with lead time becoming primarily located over the equatorial regions, in particular the eastern Pacific. In other words, these results suggest that the equatorial regions are predominately where a change can be detected in the conditional distribution of the observations given a seasonal forecast. Generally, only during the first month of the forecasts can a change in conditional distribution of the observations be seen over the land masses. Previous research concurs with the findings here showing higher predictive skill in the tropics [*Kumar et al.*, 2007; *Peng et al.*, 2000; *Phelps et al.*, 2004; *Weigel et al.*, 2008]. Results also highlight that predictive skill in the idealized world is higher than in the real world, especially for the first month but

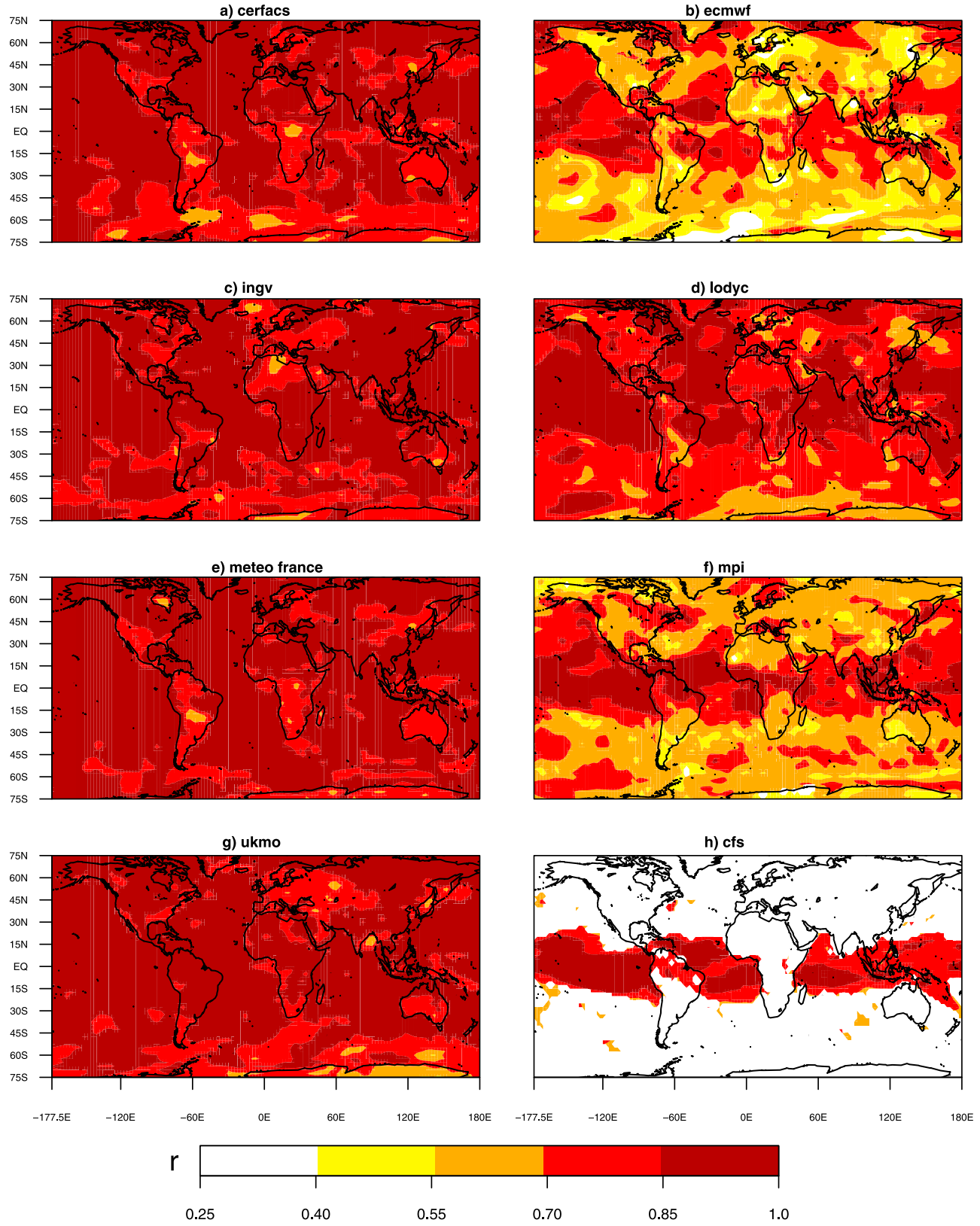

Figure 4. Idealized predictive skill for each model grid point for 1981–2001 for a 30 day temporal average at a 1 day lead time for May two meter temperature forecasts for (a) CERFACS, (b) ECMWF, (c) INGV, (d) LODYC, (e) METEO FRANCE, (f) MPI, (g) UKMO, and (h) CFS. (Key same as in Figure 1.)

degrades significantly after about 30 days. The idealized predictability estimates vary between the models (Figure 4) and depending on the noise inherent in the climate model system, the potential improvement in realizable seasonal climate predictability will also vary. However, if areas with higher idealized predictability (compared to realizable predictability) of temperature in the first month (compare Figures 1 and 4) could be translated to the real climate system, then improved month-1 climate forecasts could be attained. This realization of predictive skill would have benefits for decision making based on these forecasts.

[14] Attempts are being made by the Global Land-Atmosphere Coupling Experiment (GLACE2) (R. D. Koster et al., The contribution of soil moisture initialization to subseasonal forecast skill: First results from the GLACE-2 project, manuscript in preparation, 2009) to assess whether sub-seasonal predictive skill can be improved by having a more accurately initialized land surface. The Global Energy and Water Cycle Experiment (GEWEX) [*Sorooshian et al.*, 2005] and the Hydrologic Ensemble Prediction Experiment (HEPEX) [*Schaake et al.*, 2007] also aim to improve seasonal prediction practices. There is potential in using a multi-model approach [*Krishnamurti et al.*, 2006], but the ideal way to combine the models is unresolved [*Kirtman and Pirani*, 2009]. Given the actual skill demonstrated by operational seasonal climate forecasting models, it appears that only through significant model improvements can useful long-lead forecasts be provided that would be useful for decision makers – a quest that may prove to be elusive.

[15] **Acknowledgments.** David Lavers is a PhD student funded by the United Kingdom Natural Environment Research Council (NER/S/A/2005/13646A) and is currently based at Princeton University, USA. We acknowledge NOAA grants NA06OAR4310051 and NA17RJ2612 that also made the work possible. The DEMETER and ERA-40 datasets were retrieved from the ECMWF data server.

References

Cantelaube, P., and J. M. Terres (2005), Seasonal weather forecasts for crop yield modelling in Europe, *Tellus, Ser. A*, *57*(3), 476–487, doi:10.1111/j.1600-0870.2005.00125.x.

Carson, D. J. (1998), Seasonal forecasting, *Q. J. R. Meteorol. Soc.*, *124*(545), 1–26, doi:10.1002/qj.49712454502.

Challinor, A. J., J. M. Slingo, T. R. Wheeler, and F. J. Doblas-Reyes (2005), Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles, *Tellus, Ser. A*, *57*(3), 498–512, doi:10.1111/j.1600-0870.2005.00126.x.

Colman, A., and M. Davey (1999), Prediction of summer temperature, rainfall and pressure in Europe from preceding winter North Atlantic Ocean temperature, *Int. J. Climatol.*, *19*(5), 513–536, doi:10.1002/(SICI)1097-0088(199904)19:5<513::AID-JOC370>3.0.CO;2-D.

Davies, J. R., D. P. Rowell, and C. K. Folland (1997), North Atlantic and European seasonal predictability using an ensemble of multidecadal atmospheric GCM simulations, *Int. J. Climatol.*, *17*(12), 1263–1284, doi:10.1002/(SICI)1097-0088(199710)17:12<1263::AID-JOC191>3.0.CO;2-1.

Folland, C. K., A. Colman, D. P. Rowell, and M. Davey (2001), Predictability of northeast Brazil rainfall and real-time forecast skill, 1987–98, *J. Clim.*, *14*(9), 1937–1958, doi:10.1175/1520-0442(2001)014<1937:PONBRA>2.0.CO;2.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept, *Tellus, Ser. A*, *57*(3), 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Kirtman, B. P., and A. Pirani (2009), The state of the art of seasonal prediction: Outcomes and recommendations from the First World Climate Research Program Workshop on Seasonal Prediction, *Bull. Am. Meteorol. Soc.*, *90*(4), 455–458, doi:10.1175/2008BAMS2707.1.

Koster, R. D., et al. (2004), Realistic initialization of land surface states: Impacts on subseasonal forecast skill, *J. Hydrometeorol.*, *5*(6), 1049–1063, doi:10.1175/JHM-387.1.

Krishnamurti, T. N., A. Chakraborty, R. Krishnamurti, W. K. Dewar, and C. A. Clayson (2006), Seasonal prediction of sea surface temperature anomalies using a suite of 13 coupled atmosphere-ocean models, *J. Clim.*, *19*(23), 6069–6088, doi:10.1175/JCLI3938.1.

Kumar, A., B. Jha, Q. Zhang, and L. Bounoua (2007), A new methodology for estimating the unpredictable component of seasonal atmospheric variability, *J. Clim.*, *20*(15), 3888–3901, doi:10.1175/JCLI4216.1.

Lorenz, E. N. (1963), Deterministic nonperiodic flow, *J. Atmos. Sci.*, *20*(2), 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Luo, L., E. F. Wood, and M. Pan (2007), Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions, *J. Geophys. Res.*, *112*, D10102, doi:10.1029/2006JD007655.

Murphy, A. H. (1993), What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather Forecast.*, *8*(2), 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Murphy, A. H., and R. L. Winkler (1987), A general framework for forecast verification, *Mon. Weather Rev.*, *115*(7), 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Murphy, A. H., B. G. Brown, and Y.-S. Chen (1989), Diagnostic verification of temperature forecasts, *Weather Forecast.*, *4*(4), 485–501, doi:10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2.

Palmer, T. N., and D. L. T. Anderson (1994), The prospects for seasonal forecasting—A review paper, *Q. J. R. Meteorol. Soc.*, *120*(518), 755–793.

Palmer, T. N., et al. (2004), Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, *85*(6), 853–872, doi:10.1175/BAMS-85-6-853.

Peng, P., A. Kumar, A. G. Barnston, and L. Goddard (2000), Simulation skills of the SST-forced global climate variability of the NCEP–MRF9 and the Scripps–MPI ECHAM3 models, *J. Clim.*, *13*(20), 3657–3679, doi:10.1175/1520-0442(2000)013<3657:SSOTSF>2.0.CO;2.

Phelps, M. W., A. Kumar, and J. J. O'Brien (2004), Potential predictability in the NCEP CPC dynamical seasonal forecast system, *J. Clim.*, *17*(19), 3775–3785, doi:10.1175/1520-0442(2004)017<3775:PPITNC>2.0.CO;2.

Rudolf, B., C. Beck, J. Grieser, and U. Schneider (2005), *Global Precipitation Analysis Products of the GPCC*, Dtsch. Wetterdienst, Offenbach, Germany.

Saha, S., et al. (2006), The NCEP climate forecast system, *J. Clim.*, *19*(15), 3483–3517, doi:10.1175/JCLI3812.1.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark (2007), HEPEX: The Hydrological Ensemble Prediction Experiment, *Bull. Am. Meteorol. Soc.*, *88*(10), 1541–1547, doi:10.1175/BAMS-88-10-1541.

Shukla, J. (1998), Predictability in the midst of chaos: A scientific basis for climate forecasting, *Science*, *282*(5389), 728–731, doi:10.1126/science.282.5389.728.

Sorooshian, S., et al. (2005), Water and energy cycles: Investigating the links, *WMO Bull.*, *54*(2), 58–64.

Stockdale, T. N., M. A. Balmaseda, and A. Vidard (2006), Tropical Atlantic SST prediction with coupled ocean-atmosphere GCMs, *J. Clim.*, *19*(23), 6047–6061, doi:10.1175/JCLI3947.1.

Tang, Y., H. Lin, and A. M. Moore (2008), Measuring the potential predictability of ensemble climate predictions, *J. Geophys. Res.*, *113*, D04108, doi:10.1029/2007JD008804.

Thomson, M. C., et al. (2006), Malaria early warnings based on seasonal climate forecasts from multi-model ensembles, *Nature*, *439*(7076), 576–579, doi:10.1038/nature04503.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*(612), 2961–3012, doi:10.1256/qj.04.176.

Van Oldenborgh, G. J., M. A. Balmaseda, L. Ferranti, T. N. Stockdale, and D. L. T. Anderson (2005), Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-year period, *J. Clim.*, *18*(16), 3250–3269, doi:10.1175/JCLI3421.1.

Weigel, A. P., D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller (2008), Probabilistic verification of monthly temperature forecasts, *Mon. Weather Rev.*, *136*(12), 5162–5182, doi:10.1175/2008MWR2551.1.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed., Academic, Burlington, Mass.

Wu, R., B. P. Kirtman, and H. M. Van den Dool (2009), An analysis of ENSO prediction skill in the CFS retrospective forecasts, *J. Clim.*, *22*(7), 1801–1818, doi:10.1175/2008JCLI2565.1.

D. Lavers and E. F. Wood, Environmental Engineering and Water Resources, Department of Civil and Environmental Engineering, Princeton University, Engineering Quadrangle, Princeton, NJ 08544, USA. (davver@ceh.ac.uk)

L. Luo, Department of Geography, Michigan State University, 122 Geography Bldg., East Lansing, MI 48824, USA.