Clim Dyn
DOI 10.1007/s00382-017-3657-2

Temperature trends and prediction skill in NMME seasonal forecasts

Nir Y. Krakauer[1]

Received: 6 November 2016 / Accepted: 16 March 2017

Abstract The North American Multi-Model Ensemble (NMME) provides hindcasts and real-time predictions for monthly mean climate fields at lead times of up to a year. These global climate model outputs can be useful in constructing improved seasonal forecasts. Here, several simple methods are developed and evaluated for forecasting monthly temperatures up to a year in advance based on either unweighted or weighted NMME outputs, and compared to previously developed statistical forecast methods that use only time series of past observations. It is found that the NMME-based methods produce forecast temperature probability distributions that are appropriately shifted toward the warm end of past experience and also show skill at representing interannual variability. NMME-based methods clearly outperformed purely statistical methods for forecasting temperatures over ocean, though over land this improvement is less clear over the evaluation period tested. The NMME seasonal forecasts may be particularly useful for giving early warning of heat waves, given their societal significance and higher conditional skill under those conditions.

Keywords NMME · Seasonal forecasting · Temperature · Berkeley Earth · Heat waves

1 Introduction

Phase 1 of the North American Multi-Model Ensemble (NMME) provides hindcasts and real-time predictions for monthly mean climate fields at lead times of up to a year (Kirtman et al. 2014). These global climate model (GCM) outputs can be useful in constructing improved seasonal forecasts. The goal of this paper is to apply and evaluate methods of constructing fully probabilistic monthly mean temperature forecasts using NMME predicted fields as an input.

Much of the skill of operational seasonal temperature forecasts is due simply to accounting for the anthropogenic warming trend compared to a past climatology period (Krakauer et al. 2013). Skillful forecasts can in fact be based on only a time series of past temperatures (which allows estimation of the warming magnitude), without any use of predictions from elaborate numerical climate models, such as those in NMME (Krakauer 2012; Krakauer and Devineni 2015). On the other hand, misestimation of trends can degrade the quality of seasonal forecasts (Krakauer et al. 2013; Jia and Lin 2013). Initial analysis found that NMME models tended to underestimate the frequency of warm departures from climatology and overestimate the frequency of cold departures, suggesting that they may not fully account for warming over recent decades (Kirtman et al. 2014). Regression-based forecast methods can combine information from

This paper is a contribution to the special collection on the North American Multi-Model Ensemble (NMME) seasonal prediction experiment. The special collection focuses on documenting the use of the NMME system database for research ranging from predictability studies, to multi-model prediction evaluation and diagnostics, to emerging applications of climate predictability for subseasonal to seasonal predictions.This special issue is coordinated by Annarita Martiotti (NOAA), Heather Archambault (NOAA), Jin Huang (NOAA), Ben Kirtman (University of Miami) and Gabriele Villarini (University of Iowa).

✉ Nir Y. Krakauer
mail@nirkrakauer.net

[1] Department of Civil Engineering, The City College of New York, New York, NY, USA

Published online: 09 April 2017

numerical climate model predictions and empirical trend estimates based on past observation time series, and potentially outperform forecast methods that only use one or the other (Krakauer and Fekete 2014; Aizenman et al. 2016). Where predictions from multiple climate models are available, as for NMME, the simplest approach is to take the multi-model average, which generally is found to outperform most of the contributing models taken individually (Kirtman et al. 2014; Infanti and Kirtman 2014; Becker et al. 2014; Louise Slater et al. 2016). If different models exhibit consistently different levels of skill, a weighted average may outperform a simple average (Johnson and Swinbank 2009; Knutti et al. 2010; Du and Zhou 2011; DelSole et al. 2013; Ma et al. 2016; Wanders and Wood 2016).

While overall measures of skill at temperature prediction are useful in evaluating forecasts, it may be of particular interest to assess the performance of forecasts under cold or hot extremes, as well as where temperatures were near-normal (Barnston and Mason 2011; Becker et al. 2013). In recent decades, the frequency of record-cold conditions has decreased globally, while the frequency of record heat has increased sharply (Tebaldi et al. 2006; Rahmstorf and Coumou 2011; Tingley and Huybers 2013; Donat et al. 2013; Coumou and Robinson 2013; Matthes et al. 2015; Krakauer and Devineni 2015; Mishra et al. 2015). Extended heat waves are a major hazard to health (Fouillet et al. 2006; Sherwood and Huber 2010; Peng et al. 2011) and agriculture (Valtorta 2002; Zaitchik et al. 2006; Velde et al. 2010), so better prediction of them may enable large benefits. Ideally, a model ensemble-based seasonal forecast would capture this systematic change in the frequency of cold versus hot extremes, as well as some of the dynamic variability which modulates monthly temperatures in any particular year.

Given these considerations, the remainder of this paper is structured as follows. First, the model outputs and observations used, analysis methods, and software implementation are briefly described. Next, the magnitude of linear warming trends in an observation-based dataset is compared to that seen in NMME outputs, and NMME prediction correlations with observations are calculated with and without detrending. Then, the performance of several forecast methods based on combinations of historical observations and either mean or differentially weighted NMME predictions is compared for forecasting monthly temperatures over recent years. The ability of the forecast methods to predict different ranges of temperatures (near-normal levels vs. cold and hot extremes) is then examined. The concluding sections summarize the results and suggest directions for follow-up work.

2 Methods

2.1 Temperature data

Monthly mean surface air temperatures for 1957–2015 were taken from the Berkeley Earth Land + Ocean dataset (http://berkeleyearth.org/data/). This uses several times more station records compared to other gridded temperature data sets. The station records undergo automated tests for homogeneity and station-specific change points, and are weighted using geostatistics methods to produce spatial fields for each month (Rohde et al. 2013). The station data are used to estimate the temperature field over land and sea ice, while the temperature field over the ocean is based on the Hadley Centre sea surface temperature dataset (Rayner et al. 2003). Despite differences in station data and methodology, large-scale temperature trends (global warming and decadal and interannual variability) in Berkeley Earth are broadly similar to those of other gridded compilations (Rohde et al. 2013; Muller et al. 2013). Temperatures are given as anomalies relative to the 1951–1980 average on a global 1° grid. These were here regridded to the different 1° grid used by NMME.

2.2 Predictions from NMME models

NMME seasonal predictions (Kirtman et al. 2014) have been produced since 2011, with participating climate modeling groups initializing their simulations at the beginning of each calendar month and the predicted fields publicly available online at the IRI Data Library (Blumenthal et al. 2014) on the 8th of the month. These forecasts are for monthly mean surface air temperature (among other variables) on a 1° latitude-longitude grid. Lag-0.5 forecasts are for the month at whose beginning they were initialized, lag-1.5 forecasts are for the following month, and so forth. Additionally, the participating climate models have made available hindcasts initialized at the beginning of past months, going back about 30 years, to provide a longer period for evaluating forecast skill.

For the current work, hindcasts and predictions of temperatures for each month in 1982–2015 were used. Probabilistic forecasts were constructed and their skill evaluated only for the end of this period, 2012–2015, approximately beginning with the transition from hindcasts to real-time predictions in NMME.

Lags of 0.5–11.5 months and all GCMs with NMME hindcasts/predictions for most months since 1982 and continuing up to the present were considered. This criterion resulted in 9 GCMs selected: CMC1-CanCM3, CMC2-CanCM4, COLA-RSMAS-CCSM3, COLA-RSMAS-CCSM4, GFDL-CM2p1-aer04, GFDL-CM2p5-FLOR-A06, GFDL-CM2p5-FLOR-B01, NASA-GMAO-062012,

NCEP-CFSv2. All ensemble members for each GCM were averaged to get a mean prediction for that GCM. Where a GCM prediction was missing for a particular month, its mean was imputed based on the means of the available GCMs offset by an amount corresponding to the mean discrepancy between them and the target GCM in past years. Such a strategy for handling occasional missing values is needed in the operational setting, when predictions from any one GCM may fail to be uploaded in time to be usable in constructing a forecast. NASA-GMAO-062012 only predicted for lags up to 8.5 months and NCEP-CFSv2 up to 9.5 months, so analyses at the longest lags exclude these models.

2.3 Evaluation of warming trends

The global warming trend was calculated for global-mean BEST monthly 1982–2015 temperature anomalies using simple linear regression. The warming trend was similarly calculated for NMME models for their 1982–2015 hindcasts/predictions at each of the 12 lags. Correlation coefficients between the NMME prediction and Berkeley Earth temperature time series for each calendar month and grid point before and after removing a linear trend were also computed, in order to assess how much of each NMME model's skill at reproducing observed temperatures (at various forecast lags) is due to the long-term warming trend, versus successful simulation of interannual variability.

2.4 Forecast methods

We choose a set of relatively simple forecast methods which include ones based only on previous observations, previously intercompared for hindcasts of station monthly temperatures (Krakauer and Devineni 2015), as well as ones that use NMME outputs in a linear regression, with or without differential model weights or trend adjustment.

Each method assumes that the yearly time series of temperature at a given calendar month $T(t)$ can be represented as the sum of an explainable component $\bar{T}(t)$ and a zero-mean, normally distributed unexplainable component $\epsilon(t)$. The difference between the fitted $\bar{T}(t)$ and the observed value over past years is used to estimate the variance of $\epsilon(t)$ (Krakauer and Fekete 2014).

$$T(t) = \bar{T}(t) + \epsilon(t). \tag{1}$$

Using this framework, our goal was to estimate a probability distribution for T at a given year t_f given observations from previous years or NMME predictions. Each method generates a probability distribution for the temperature at each forecast month and grid point which has the form of a t distribution (Krakauer and Devineni 2015).

The observation-only statistical forecast methods are (see Krakauer and Devineni (2015) for more details):

- C: Climatology. Forecast probability distributions are based on the mean and standard deviation of observed temperature over a fixed past period, here taken to be 1981–2010.
- MA: Moving average. Forecast probability distributions are based on the mean and standard deviation of observed temperature over the previous 30 years. For the evaluation period here, this period will only differ from 1981–2010 by a few years.
- EW: Exponentially weighted moving average. Forecast probability distributions are based on the mean and standard deviation of observed temperature over recent decades (in this case, since 1957) but with more recent observations given greater weight (with an e-folding weighting timescale τ of 15 years). These forecasts still have a cool bias, but this tends to be less than seen in C or MA forecasts because the forecast expectation is closer to the more recently observed values.
- EW-a: Adjusted exponentially weighted moving average. This applies an additive offset to the EW forecast expectation based on a smoothing spline fit to the global-mean observation time series, in an attempt to better capture the warming of recent years.

The forecast methods which use NMME predictions employ either the multi-GCM mean $\bar{M}$ or the individual GCM means M_i for the target forecast month as well as for previous years (since 1982).

- M: NMME-mean. Only a constant offset o from the multi-GCM NMME mean $\bar{M}$ is estimated for each grid point, so the linear regression model is

$$\bar{T}(t) = o + \bar{M}(t). \tag{2}$$

- MT: NMME-mean with trend. The regression model includes a linear trend (determined for each forecast month globally by linear regression over all grid cells) to allow for NMME models not reproducing the observed trend:

$$\bar{T}(t) = o + \bar{M}(t) + \alpha t. \tag{3}$$

- MS: Scaled NMME-mean. The NMME mean is multiplied by a scale factor β (determined for each forecast month globally by linear regression over all grid cells) to allow for NMME tending to over- or under-predict the magnitude of temperature anomalies.

$$\bar{T}(t) = o + \beta\bar{M}(t). \tag{4}$$

- MST: Scaled NMME-mean with trend. Both scaling and a trend are applied.

$$\bar{T}(t) = o + \beta \bar{M}(t) + \alpha t. \quad (5)$$

– MM: NMME-multimodel. A linear combination of the individual NMME GCM means is used to allow for differential GCM skill. The weights β_i are determined for each forecast month globally, by linear regression over all grid cells.

$$\bar{T}(t) = o + \sum_i \beta_i M_i(t). \quad (6)$$

– MMT: NMME-multimodel with trend. Same as MM, but with a globally determined linear trend added.

$$\bar{T}(t) = o + \sum_i \beta_i M_i(t) + \alpha t. \quad (7)$$

Preliminary testing also considered methods in which the weights β or trend α were determined separately for each grid cell instead of globally, but, given the relatively short period available for parameter estimation, these tended to perform worse.

Each of the forecast-based methods is evaluated using NMME fields at each lag from 0.5 to 11.5 months. Since the 0.5-month lag predictions include the short-range weather forecast period, at which dynamic predictability is well known to be high, the analysis below focuses more on the relative performances of the various methods at the 1.5 month and longer lags.

2.5 Forecast skill metrics

Given a temperature forecast for given month and location as a probability distribution $p(y)$ with expectation y^* and corresponding verifying observation y, a common deterministic forecast metric is mean square error (MSE, $(y^* - y)^2$ averaged across a set of forecasts with available verifying observations), or its square root (RMSE). MSE can be decomposed into bias and variance components (Krakauer and Devineni 2015). These deterministic metrics however are only sensitive to the forecast expectation, not the full probability distribution.

A probabilistic skill score is given by the mean negative log likelihood:

$$\mathrm{NLL} = \langle -\log p(y) \rangle, \quad (8)$$

where p is the forecast probability distribution, y is the observation, and $\langle \cdot \rangle$ denotes averaging across a set of forecast-observation pairs. The NLL values for different forecast methods have units of information (e.g. bits or nats, depending on the base of the logarithm taken – natural logarithms are used here) and can be related to the methods' ability to reduce uncertainty in a decision-making framework. The difference between the NLL of a given forecast and than of some baseline forecast for the same set of observations gives the information gain of the forecast relative to the baseline:

$$\mathrm{IG} = \mathrm{NLL}_0 - \mathrm{NLL}, \quad (9)$$

where NLL_0 is for the baseline forecast. If the forecast is more skillful than the baseline, IG is expected to be positive.

Another probabilistic skill score is the continuous ranked probability score (CRPS), the mean square difference between the observed and the forecast cumulative distribution functions (Matheson and Winkler 1976; Bradley and Schwartz 2011). Unlike NLL, this is non-local, in that its value is affected by the probabilities given to all possible outcomes, not just the observed outcome (Tödter et al. 2011; Smith et al. 2015), and it is less sensitive to departures from forecast model assumptions such as normal distribution of the prediction error (Pieroth 2014). CRPS was computed using the analytic expression for its value given a forecast t distribution (Jordan et al. 2016).

Here, skill metrics are averaged across months and grid cells (weighted by grid cell area) to produce global skill measures. We primarily compare averages over land, since skillful prediction of over-land temperatures is of greater interest for many applications, and since temperature prediction for the land surface, with its shorter thermal memory, is intrinsically more challenging than for the sea surface. We also show some results averaged over the ocean for comparison. The same metrics could also be averaged for different spatiotemporal subsets in order to study, for example, whether the ranking of methods is consistent across regions, seasons, or special conditions such as El Niño events.

We can ask whether the finding of one forecast model having a better mean skill value (e.g. lower NLL) than another model is likely to be robust, or whether the difference is small enough that it does not constitute strong evidence for one forecast performing consistently better for other forecast times. As a measure of significance, the t test can be applied to the time series of mean monthly difference in the skill metric between two models of interest, with the null hypothesis being that the expected value of the difference is equal to zero. To reduce false positives, the number of degrees of freedom in the t test is adjusted for temporal correlation using a formula based on the empirical lag-1 autocorrelation of the time series (Krakauer et al. 2013; Aizenman et al. 2016). It is found that for this evaluation period (2012–2015) and data, global land mean differences in NLL between forecast methods typically needed to be at least 0.03 nats, and differences in RMSE at least 0.03 K, to be significantly different from zero at the 95% level, though these thresholds varied somewhat depending on which methods were being compared.

2.6 Forecast skill by category

In addition to measures that average across all forecasts in a given time period, we can consider how well particular outcome categories, such as extremely cold or hot conditions, are forecast. To do this, we derive from the temperature climatology for each location and calendar month 100 equal-probability categories (percentiles) based on the mean and standard deviation of 1981–2010 observations (and assuming a normal distribution). Then we can compute the mean forecast probability for each category, both unconditional (averaging across all months) and conditional on the observation turning out to fall in that category. A climatology-based forecast for 2012–2015 will give approximately equal probability to each category, even though, due to the warming trend, hot months (relative to climatology) were in fact much more common than cold months. The observation-based forecast methods that include an adjustment for the warming trend (such as EW-a) will forecast higher probabilities for the hotter categories, largely independent of the category that was observed. The NMME-based forecast methods may forecast higher conditional probabilities than the unconditional average, assuming that they can capture interannually-varying factors that increase the likelihood of, for example, hot and cold extremes.

2.7 Software implementation

The NMME models and observations were read and processed and forecasts generated and evaluated using version 0.0.6 of the SeFo package for the free programming environment Octave (Eaton 2012). SeFo includes modules downloading and reading past observations and model ensemble output; producing forecast probability distributions; and diagnosing the performance of specified methods over any desired available verification period. The software architecture is described elsewhere (Krakauer 2016). The commands used to generate all analyses and plots presented in this paper are included in this version of SeFo as the script file `sefo_script_2016a.m`.

3 Results

3.1 Warming trends and correlation with observations

The over-land linear warming trend in Berkeley Earth over 1982-2015 was 0.29 K/decade (regression nominal 95% confidence interval: 0.26–0.32). The warming trend for the NMME multi-GCM mean over the same period at 0.5 months lead was lower, 0.22 (0.20–0.24) K/decade (Table 1). In fact, the NMME mean warming trend remained very similar at all lead times, with a range of only 0.213–0.226 K/decade (Fig. 1). This consistency however seems to be fortuitous, as different NMME models' simulated warming rates changed with lead time, for example decreasing from 0.27 to 0.17 K/decade in CMC2-CanCM4 as the lead time increased from 0.5 to 11.5 months, while increasing from 0.22 to 0.34 K/decade in all 3 GFDL models. The models for the most part reproduced, for example, the cooling seen after the Pinatubo eruption in 1991 and warming during strong El Niño events in 1998 and 2015 (Fig. 2). NCEP-CFSv2 had the largest warming rate (close to observations) at short lags, with the GFDL models surpassing it at longer lags (Fig. 1). Over the oceans, the observed warming was slower, 0.15 (0.14–0.16) K/decade, and the NMME mean trend was again slightly less than observed, 0.117–0.134 K/decade depending on the lag.

Table 1 Warming trends in surface air temperature over land for NMME forecasts, 1982–2015, at two different forecast lead times

Model	0.5 month	1.5 month
CMC1-CanCM3	0.20 (0.17–0.23)	0.17 (0.15–0.19)
CMC2-CanCM4	0.27 (0.24–0.30)	0.25 (0.23–0.27)
COLA-RSMAS-CCSM3	0.03 (0.01–0.05)	−0.02 (−0.05 to +0.02)
COLA-RSMAS-CCSM4	0.28 (0.25–0.31)	0.28 (0.26–0.31)
GFDL-CM2p1-aer04	0.22 (0.19–0.25)	0.24 (0.22–0.26)
GFDL-CM2p5-FLOR-A06	0.22 (0.20–0.24)	0.24 (0.22–0.26)
GFDL-CM2p5-FLOR-B01	0.22 (0.20–0.24)	0.23 (0.21–0.25)
NASA-GMAO-062012	0.24 (0.21–0.26)	0.24 (0.23–0.26)
NCEP-CFSv2	0.29 (0.27–0.32)	0.30 (0.28–0.31)
Mean	0.22 (0.20–0.24)	0.21 (0.20–0.23)

Units are K/decade, ranges in parentheses are regression 95% confidence intervals

The warming rate based on observations (BEST) was 0.29 K/decade (0.26–0.32)

COLA-RSMAS-CCSM3 was unique in having little warming trend at any forecast lead time. One contributor to this lack of warming was an unusual cold excursion of several K in a number of COLA-RSMAS-CCSM3 runs initialized toward the end of the study period (late 2014 and early 2015), though even with these outlier months excluded COLA-RSMAS-CCSM3 showed very little warming compared to observations (Fig. 2).

All the NMME climate models showed positive mean associations between their temperature prediction time series and the Berkeley Earth observations over 1982–2015. At all lags, the multimodel mean had a higher mean correlation with observations than any of the individual models. Correlation strength averaged over land grid cells decreased with prediction lead time, from 0.58 for the multimodel mean (with a range of 0.14–0.56 for individual

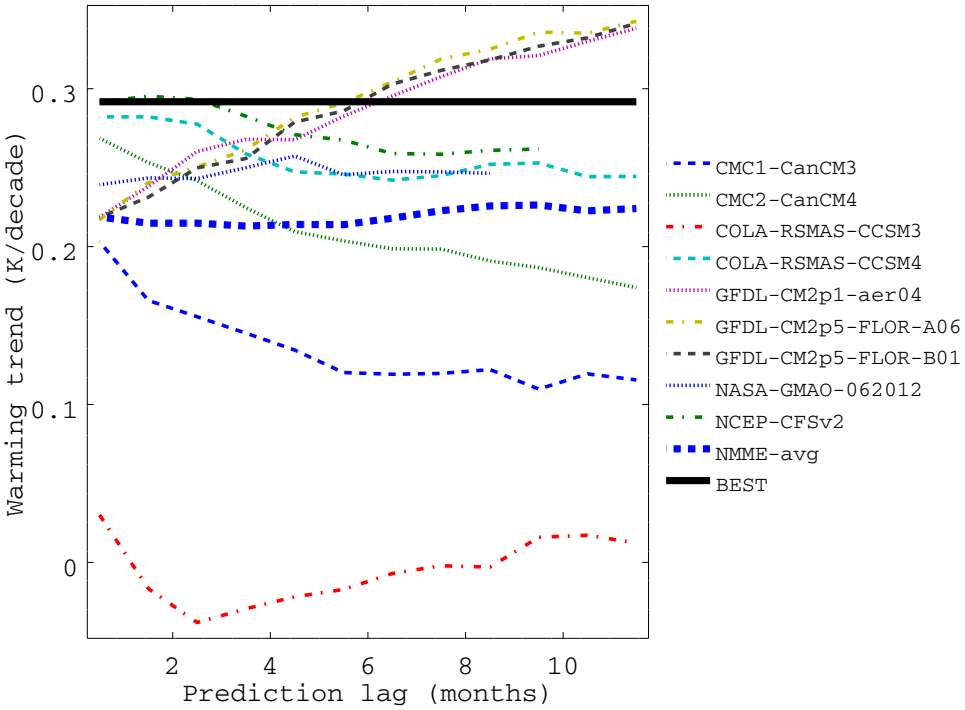

Fig. 1 Global (surface air temperature over land) warming rate of NMME model and multimodel mean temperature forecasts for 1982–2015, as a function of forecast lag. The warming rate based on observations (BEST) is also shown for reference

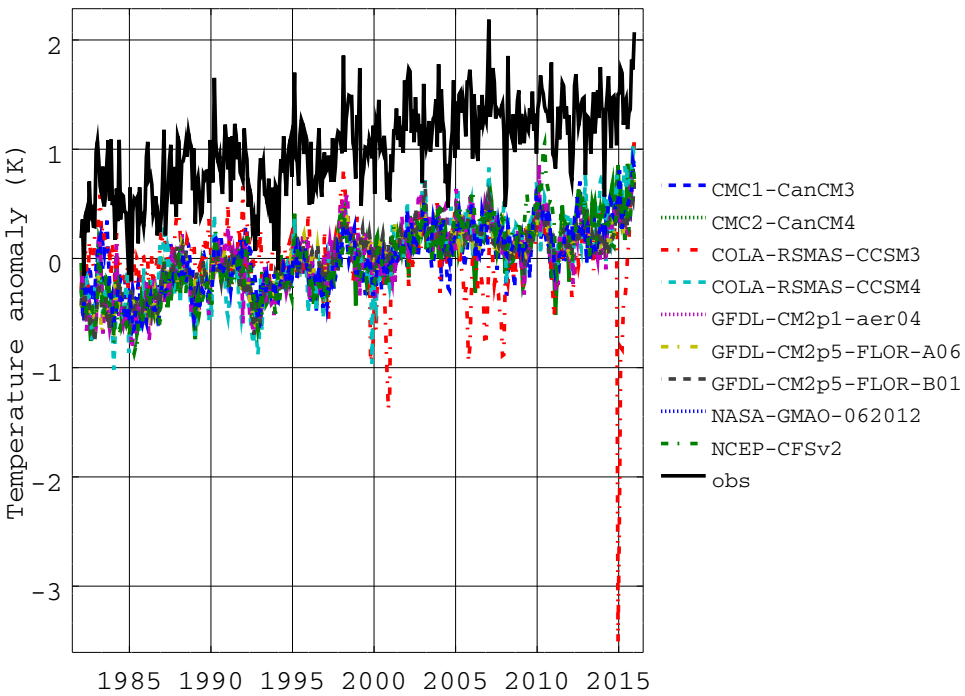

Fig. 2 Global mean surface air temperatures (over land) in NMME model predictions (at lag 1.5 months) and observations (BEST) for 1982–2015, as a function of forecast lag. The 1982–2015 mean seasonal cycle has been subtracted from each series, and the observation time series is offset 1 K so that it can be compared with the model series more easily

models) at lag 0.5 to 0.37 (0.09–0.31) at lag 1.5 and 0.26 (0.04–0.23) at lag 11.5 (Fig. 3a). Correlations remained positive after detrending and again were better for the multimodel mean than for the individual models, going from 0.54 (0.13–0.52) at lag 0.5 and 0.28 (0.10–0.22) at lag 1.5 to 0.21 (0.02–0.11) at lag 11.5 (Fig. 3b). Out of the individual NMME models, COLA-RSMAS-CCSM3 showed the worst performance (lowest mean correlation) at all lags. The best model (highest mean correlation) was NCEP-CFSv2 for lags 0.5, 1.5 and 2.5 and varied between CMC2-CanCM4, GFDL-CM2p5-FLOR-A06, GFDL-CM2p5-FLOR-B01, NASA-GMAO-062012 for different longer lags. Random permutation of the forecast years showed that mean correlations of 0.05 and above can be considered

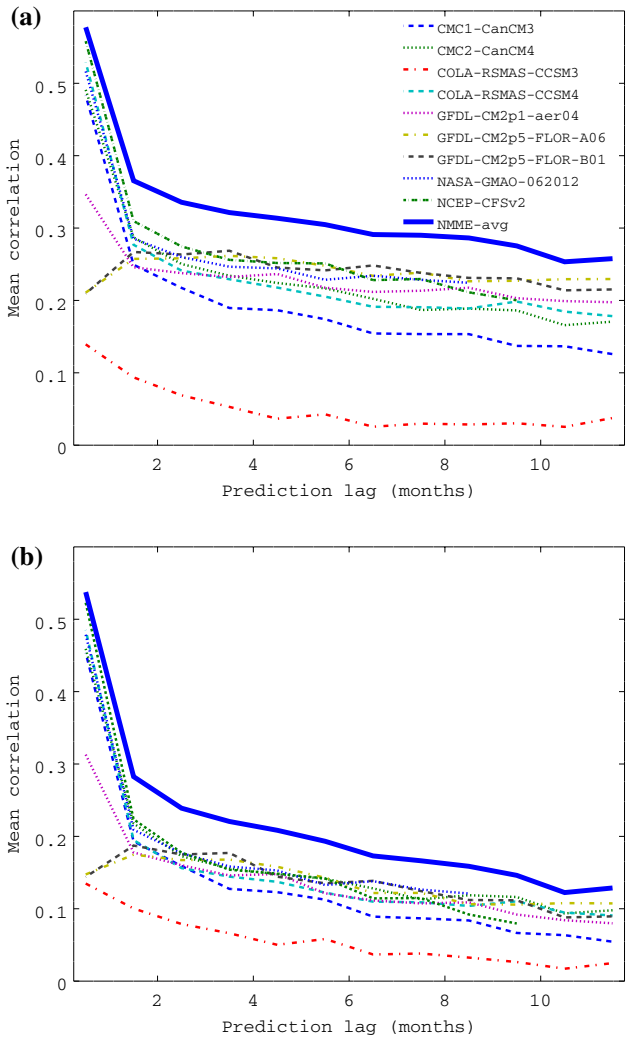

Fig. 3 **a** Mean correlations with observations of NMME model and multimodel mean temperature forecasts over land grid cells for 1982–2015. **b** After detrending

to be significantly greater than 0. Over-ocean correlations showed a similar decrease with increasing forecast lag but were higher than land correlations, for example 0.57 for the multimodel mean (0.27–0.53 for individual models) for lag 1.5 before detrending and 0.52 (0.28–0.48) after detrending. Unlike over land, the CMC models were the ones that had the highest correlations with observations in the first few lags.

3.2 Mean performance of forecast methods

Of the observation-only forecast methods, EW-a performed the best for 2012–2015 (lowest NLL and RMSE) and C the worst. C, MA, EW forecasts had a pronounced cool bias of 0.3–0.4 K, as these methods do not fully account for the warming seen in recent decades. This cool bias was reduced to under 0.1 K in EW-a, indicating the global mean warming trend was successfully estimated in the bias adjustment step (Table 2).

Depending on the lag, the best-performing NMME-based forecast method (lowest NLL) was usually MM or MMT, i.e. applying optimal linear combination of the NMME models (with globally-constant weights), either with or without bias adjustment. The difference between these two methods in NLL was not significant at any lag. These also had lowest RMSE. The trend term in MT, MST, MMT resulted in smaller cool biases compared to the models without this term, consistent with the underestimation of the observed warming trend by the NMME ensemble average that was found above. Even without the trend term, however, NMME-based forecasts had a much smaller bias than for example the climatology forecast C, reflecting the ability of the NMME ensemble to capture the majority of the observed warming amount. As expected, for each NMME-based method, forecast performance degraded (NLL and RMSE rose) as the NMME prediction lead time increased (longer lags; Fig. 4).

Comparing the observation-based and NMME-based forecast performance, at lag 1.5 all the NMME-based methods performed better than the best observation-based method EW-a in terms of land RMSE, NLL, and CRPS (Table 2), though the differences in NLL and RMSE were not statistically significant. As longer forecast lags, the performance of the NMME-based methods deteriorated, though the MST and MMT methods gave lower mean NLL than EW-a up to lag 6.5 (Fig. 4). Only at the shortest lag, 0.5 months, did the NMME-based methods significantly outperform EW-a. On the other hand, over ocean, NMME-based methods outperformed observation-based methods more convincingly, with all NMME-based methods having significantly lower NLL than EW-a at lags 1.5 (Table 3) and 2.5, and MMT having lower NLL than EW-a even at lag 11.5 (Fig. 5).

It is possible to map mean performance measures in order to offer additional insight, though given the relatively short verification period, too much should not be read into the detailed spatial patterns. As an example, Fig. 6 shows the distribution of information gain for MMT (at lag 1.5) relative to EW-a (i.e. $NLL_{EW\text{-}a} - NLL_{MMT}$). Positive information gain from the NMME-based forecast is most pronounced near the Equator, particularly in ocean areas such as the eastern Pacific and parts of the Atlantic. Given the geographic origin of the NMME models, it is encouraging that the information gain is positive in much of North America, with the noteworthy exception of the southeast United States, as well as Central America. For many land areas the information gain is near zero, suggesting little added skill from the NMME-based forecast. The information gain is negative over much of Antarctica (an area

Table 2 Skill measures for temperature forecast methods, 2012–2015, averaged over land, with lag-1.5 month NMME predictions used as inputs

Method	NLL	RMSE	Bias	CRPS
C	1.614	1.682	−0.413	0.8088
MA	1.584	1.667	−0.363	0.7950
EW	1.558	1.655	−0.318	0.7849
EW-a	1.497	1.627	−0.085	0.7597
M	1.479	1.587	−0.057	0.7409
MT	1.477	1.587	−0.011	0.7405
MS	1.475	1.585	−0.094	0.7397
MST	1.468	1.583	−0.026	0.7373
MM	1.471	1.585	−0.058	0.7386
MMT	1.466	1.584	−0.005	0.7370

NLL mean negative log likelihood of observation (nats), *RMSE* root mean square error of forecast expectation (K), *bias* mean bias of forecast expectation (K); *CRPS* continuous ranked probability score (K)

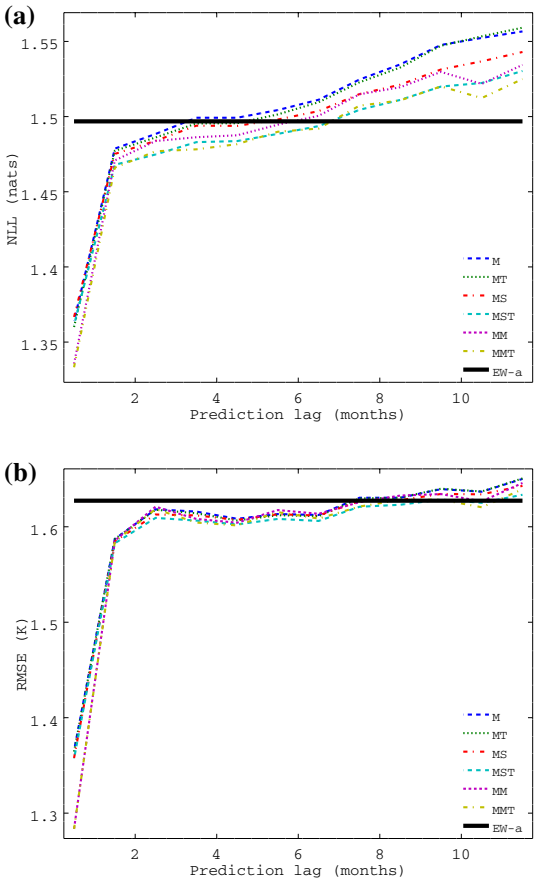

Fig. 4 **a** Negative log likelihood for forecasts of surface air temperatures over land, 2012–2015, using NMME predictions at different lags. The performance of the EW-a method, which does not use any NMME predictions, is shown as a horizontal line. **b** Root mean square error

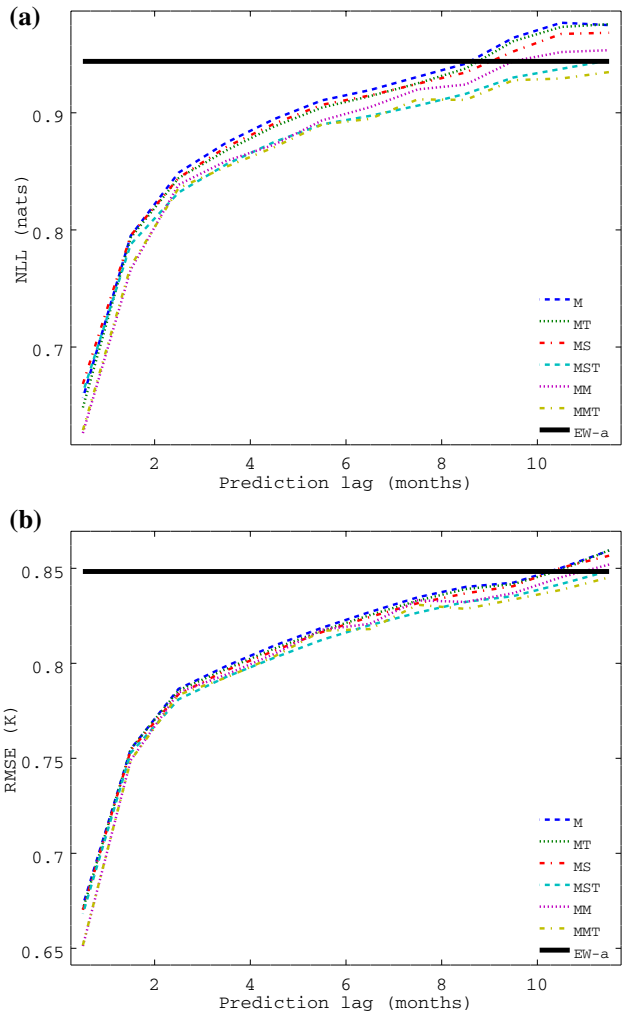

Fig. 5 Same as Fig. 4, but averaged over ocean

Table 3 Skill measures as in Table 2, but averaged over ocean

Method	NLL	RMSE	Bias	CRPS
C	1.057	0.915	−0.253	0.4251
MA	1.017	0.893	−0.225	0.4134
EW	0.998	0.872	−0.204	0.4050
EW-a	0.944	0.848	+0.029	0.3902
M	0.795	0.755	−0.050	0.3374
MT	0.793	0.755	−0.003	0.3373
MS	0.795	0.754	−0.071	0.3371
MST	0.788	0.753	+0.001	0.3360
MM	0.766	0.749	−0.036	0.3321
MMT	0.769	0.749	+0.017	0.3330

where both observations and climate model simulations may be weak), which hurts the land average performance for MMT and the other NMME-based methods.

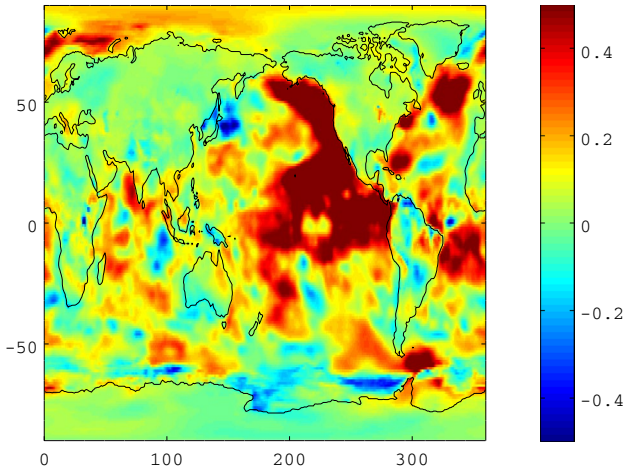

Fig. 6 Mean information gain (nats) of MMT over EW-a temperature forecast

3.3 Forecast performance by temperature departure category

Due to warming, observations for 2012–2015 showed low frequencies for the cold categories defined by the 1981–2010 climatology and high frequencies for the hot categories, whereas in a stationary climate the observations would be expected to be roughly equally distributed across categories. In fact, over land, observations in the hottest percentile were over 6 times as common as observations in the coldest percentile (3.49 versus 0.57%). The EW-a forecast is able to reproduce the observed asymmetry, whereas C, MA, EW forecast too-high probabilities for cold categories and too-low probabilities for hot categories (Fig. 7a). The NMME-based methods all more or less reproduced the asymmetry between categories (Fig. 7b).

Considering now the mean forecast probabilities of the category actually observed, even the observation-based methods were able to increase these relative to the unconditional probabilities, presumably by identifying spatial and seasonal differences in the warming trend that made warm extremes more likely, but the NMME methods (at least at reasonably short lags, such as 1.5 months) did better, showing that the NMME models' dynamic skill is useful for predicting hot and cold extremes, particularly the former, at seasonal lead times. For example, at lag 1.5 the MMT mean probability of the hottest percentile (3.49% of observations) went from 3.64% for all forecasts to 16.86% when this category was actually observed, while the mean probability of the coldest category (0.57% of observations) went from 0.62% for all forecasts to 1.40% when this category was actually observed (Fig. 8a). By contrast, for EW-a the increases were from 2.86 to 4.91 and 0.64 to 1.57%, so showing less skill at forecasting extremely hot conditions

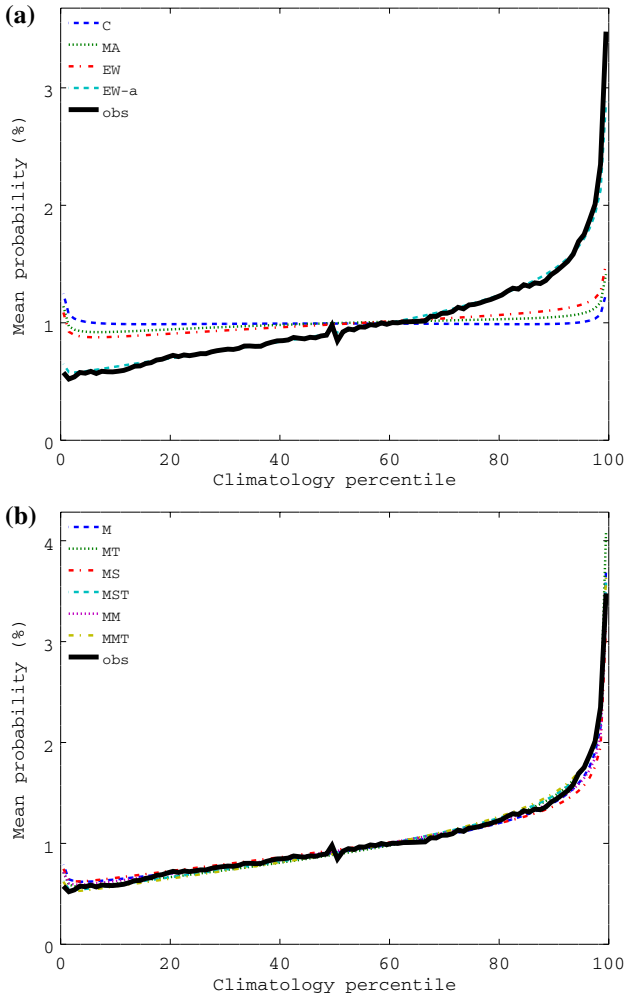

Fig. 7 **a** Mean probabilities of each climatology percentile of surface air temperature being forecast for the various observation-only (no NMME output) forecast methods. The observed frequency of each percentile is also shown. **b** Same, for the various forecast methods including NMME output (predictions at 1.5 month lag)

(Table 4). Considering the mean logarithm of the forecast probability (as used in the NLL skill score), which is more sensitive to forecasts of small probabilities, shows enhanced dynamic skill (conditional minus unconditional score) in MMT for both hot and cold extremes, whereas when conditions are closer to the climatology period median, dynamic skill is small (Fig. 8b). When MMT forecasts are categorized by the climatological absolute temperature, we see that forecast skill under all percentile outcomes, including the hot extremes, is greater in hot regions and seasons compared to cold ones (Fig. 8c), consistent with positive information gain for MMT over EW-a being seen more in the tropics compared to high latitudes (Fig. 6). Nevertheless, the pattern of greater dynamic skill for extreme outcomes compared to near-median ones is seen across the range of climatological temperatures (Fig. 8c).

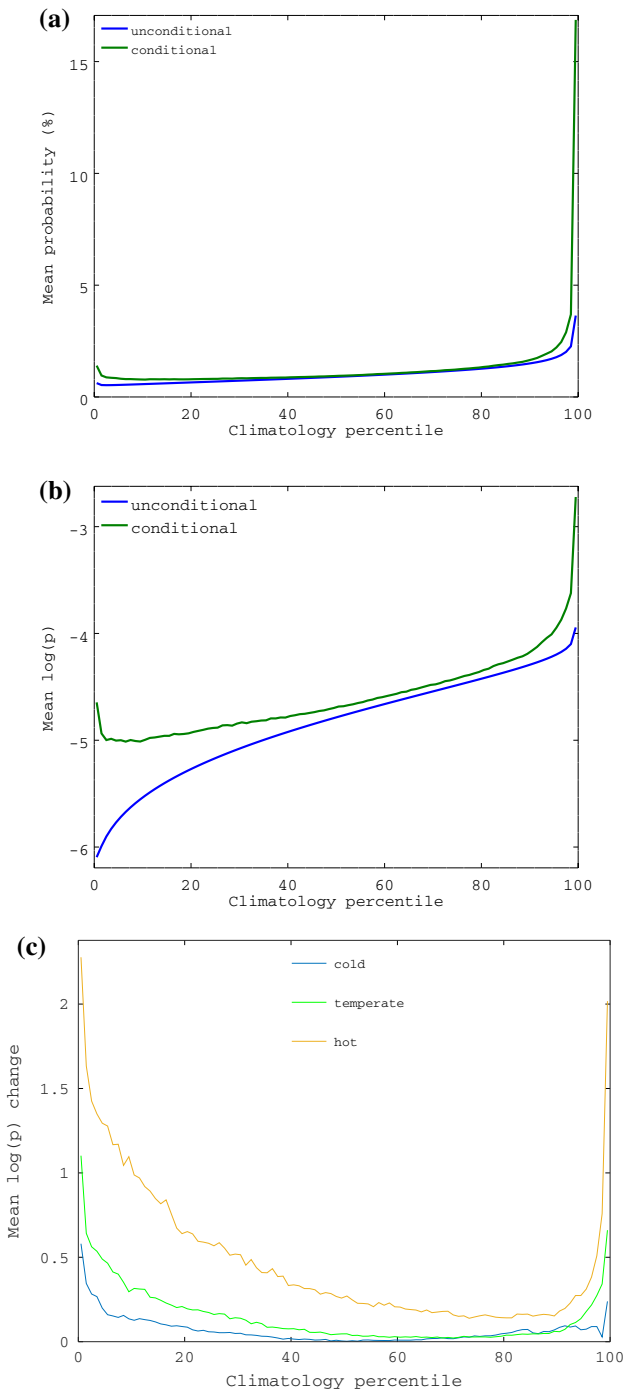

Fig. 8 For the MMT forecast method (using NMME predictions at 1.5 month lag): **a** unconditional and conditional mean probabilities of each climatology percentile of surface air temperature being forecast. Forecast skill results in higher conditional compared to unconditional probabilities. **b** Same, but for the natural logarithms of the probabilities. **c** Mean conditional minus unconditional log probability as a function of climatology percentile, for three categories of BEST climatological mean temperature: cold (<0°C, 50% of land grid cell and month combinations), hot (>25°C, 13%), and temperate (the remaining 37%)

4 Discussion

Similar to findings for purely statistical prediction of monthly station temperatures (Krakauer and Devineni 2015), it was found here for gridded temperatures that global adjustment of climatology for the warming trend, as implemented in method EW-a, gives good calibration (low bias) and reduced variance (lower RMSE) compared to methods that do not fully factor in the ongoing warming, like MA and EW. Here, this work is extended to compare such statistical forecasts with forecasts that use the NMME dynamic climate model outputs at various lead times. This results in clear improvements in performance at forecasting temperatures over ocean. However, over land, the simple NMME-based methods tried only result in clearly improved forecasts for the shortest lag, 0.5 months, which includes the short-range weather forecast period.

The hindcast/prediction series of most of the NMME models shows a warming trend, but the models' warming is mostly weaker than observed. Empirical adjustment for this underestimation of warming, as in the MT, MST, MMT, generally reduces bias but does not always improve forecast performance (as measured by NLL or RMSE), presumably because this underestimation is not strong enough or consistent enough to be accurately extrapolated given the number of available hindcast/prediction years. The NMME models' ability to capture temperature variability is not solely due to the warming trend, but also reflects skill at representing interannual climate variability such as that due to El Niño events. Comparing the NMME models' global temperature series to observations can serve to point out needs for improvement in model structure or initialization, as illustrated by the lack of warming and cold excursions seen for COLA-RSMAS-CCSM3.

Forecasts with differential climate model weighting (MM, MMT) tended to perform slightly better than a simple average of NMME models. Note that the weights were chosen using regression over all past hindcast and forecast years and hence could not respond quickly to large excursions seen in COLA-RSMAS-CCSM3 in part of the 2012-2015 evaluation period, which would have affected NMME-based forecast performance. Automated quality control checks might be useful for flagging outlier climate model outputs in an operational seasonal forecast system, with criteria perhaps based on the amount of deviation from previous years' or other climate models' simulations.

Averaging the forecast probability distributions from different methods (Ariely et al. 2000; Kim and Swanson 2014; Moral-Benito 2015) could be considered as a way of potentially combining the strengths of each method, as already tried in the seasonal weather forecasting context (Krishnamurti et al. 1999; Casanova and Ahrens 2009; Wang et al. 2012; Hawthorne et al. 2013; Dutton

Table 4 Mean probabilities (%) of occurrence of over-land temperatures in 2012–2015 in the hottest and coldest percentile of the 1982–2011 climatology for observations and different forecast methods, either unconditional (for all outcomes) or conditional (on the temperature being reaching the hottest/coldest percentile)

	Hot		Cold	
Method	Uncond	Cond	Uncond	Cond
Obs	3.49	100	0.57	100
C	1.25	1.25	1.25	1.25
MA	1.41	1.70	1.14	1.34
EW	1.50	2.15	1.08	2.17
EW-a	2.86	4.91	0.64	1.57
M	3.72	18.21	0.79	1.67
MT	4.08	19.36	0.71	1.55
MS	3.16	15.80	0.75	1.53
MST	3.58	17.03	0.63	1.33
MM	3.28	15.86	0.72	1.58
MMT	3.64	16.86	0.62	1.40

The forecasts use 1.5 month lag NMME predictions as inputs

et al. 2013; Ma et al. 2016). Given the statistically similar skill scores of the best-performing methods, simple averaging may lead to better results than Bayesian model averaging, where forecast weights must be estimated from scarce previous data (Claeskens et al. 2016). The presence of statistically significant correlation between observed and NMME-modeled temperatures over both land and ocean at even the longest lag suggests that there may be scope to use more sophisticated methods than those evaluated here to produce more skillful seasonal probabilistic forecasts.

The forecasts show the most conditional skill for temperatures that are extreme relative to the 1981–2010 climatology, with NMME-based methods being particularly good at forecasting hot extremes (Fig. 8). For thin-tailed distributions like the normal and *t*, it is expected that a shift in the mean or variance due to information from climate models would proportionally have the most impact on probabilities of extreme events. It is also possible that hot extremes are more likely to involve predictable dynamics, such as surface albedo and soil moisture feedbacks (Zaitchik et al. 2007; Fischer et al. 2007; Zampieri et al. 2009; Weisheimer et al. 2011; Hirschi et al. 2011; Vogel et al. 2017). Given the growing importance to human health and agriculture of, particularly, hot extremes (Valtorta 2002; Tebaldi et al. 2006; Sherwood and Huber 2010; Jentsch et al. 2011; Mishra et al. 2015), this property may make NMME-based temperature forecasts even more useful than average skill scores like NLL and RMSE imply.

5 Conclusions

Several simple methods are developed and evaluated for forecasting monthly temperatures up to a year in advance based on either unweighted or weighted NMME outputs. It is found that these methods produce forecast temperature probability distributions that are appropriately shifted toward the warm end of past experience and also show skill at representing interannual variability. Seasonal forecasts may be particularly useful for giving early warning of heat waves, given their societal significance and the higher conditional skill for those conditions.

References

Aizenman H, Grossberg MD, Krakauer NY, Gladkova I (2016) Ensemble forecasts: probabilistic seasonal forecasts based on a model ensemble. Climate 4(2):19

Ariely D, Au WT, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten G, Zauberman TS (2000) The effects of averaging subjective probability estimates between and within judges. J Exp Psychol Appl 6(2):130–147

Barnston AG, Mason SJ (2011) Evaluation of IRI's seasonal climate forecasts for the extreme 15% tails. Weather Forecast 26(4):545–554

Becker EJ, van den Dool H, Peña M (2013) Short-term climate extremes: prediction skill and predictability. J Clim 26(2):512–531

Becker E, van den Dool H, Zhang Q (2014) Predictability and forecast skill in NMME. J Clim 27(15):5891–5906

Blumenthal MB, Bell M, del Corral J, Cousin R, Khomyakov I (2014) IRI Data Library: enhancing accessibility of climate knowledge. Earth Perspect 1(1):1–12

Bradley AA, Schwartz SS (2011) Summary verification measures and their interpretation for ensemble forecasts. Mon Weather Rev 139(9):3075–3089

Casanova S, Ahrens B (2009) On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. Mon Weather Rev 137:3811–3822

Claeskens G, Magnus JR, Vasnev AL, Wang W (2016) The forecast combination puzzle: a simple theoretical explanation. Int J Forecast 32(3):754–762

Coumou D, Robinson A (2013) Historic and future increase in the global land area affected by monthly heat extremes. Environ Res Lett 8(3):034018

DelSole T, Jia L, Tippett MK (2013) Scale-selective ridge regression for multimodel forecasting. J Clim 26(20):7957–7965

Donat MG, Alexander LV, Yang H, Durre I, Vose R, Dunn RJH, Willett KM, Aguilar E, Brunet M, Caesar J, Hewitson B, Jack C, Klein Tank AMG, Kruger AC, Marengo J, Peterson TC, Renom M, Oria Rojas C, Rusticucci M, Salinger J, Sanhouri Elrayah A, Sekele SS, Srivastava AK, Trewin B, Villarroel C, Vincent LA, Zhai P, Zhang X, Kitching S (2013) Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. J Geophys Res 118(5):2098–2118

Dutton JA, James RP, Ross JD (2013) Calibration and combination of dynamical seasonal forecasts to enhance the value of predicted probabilities for managing risk. Clim Dyn 40(11–12):3089–3106 **(CFSv2, ECMWF)**

Eaton JW (2012) GNU Octave and reproducible research. J Process Control 22(8):1433–1438

Fischer EM, Seneviratne SI, Vidale PL, Luthi D, Schar C (2007) Soil moisture–atmosphere interactions during the 2003 European summer heat wave. J Clim 20(20):5081–5099

Fouillet A, Rey G, Laurent F, Pavillon G, Bellec S, Guihenneuc-Jouyaux C, Clavel J, Jougla E, Hémon D (2006) Excess mortality related to the August 2003 heat wave in France. Int Arch Occup Environ Health 80(1):16–24

Hawthorne S, Wang QJ, Schepen A, Robertson D (2013) Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. Water Resourc Res 49(9):5427–5436

Hirschi M, Seneviratne SI, Alexandrov V, Boberg F, Boroneant C, Christensen OB, Formayer H, Orlowsky B, Stepanek P (2011) Observational evidence for soil-moisture impact on hot extremes in southeastern Europe. Nat Geosci 4(1):17–21

Infanti JM, Kirtman BP (2014) Southeastern U.S. rainfall prediction in the North American Multi-Model Ensemble. J Hydrometeorol 15(2):529–550

Jentsch A, Kreyling J, Elmer M, Gellesch E, Glaser B, Grant K, Hein R, Lara M, Mirzae H, Nadler SE, Nagy L, Otieno D, Pritsch K, Rascher U, Schädler M, Schloter M, Singh BK, Stadler J, Walter J, Wellstein C, Wöllecke J, Beierkuhnlein C (2011) Climate extremes initiate ecosystem-regulating functions while maintaining productivity. J Ecol 99(3):689–702

Jia XJ, Lin H (2013) The possible reasons for the misrepresented long-term climate trends in the seasonal forecasts of HFP2. Mon Weather Rev 141(9):3154–3169

Johnson C, Swinbank R (2009) Medium-range multimodel ensemble combination and calibration. Q J R Meteorol Soc 135(640A):777–794

Jordan A, Krüger F, Lerch S (2016) scoringRules: scoring rules for parametric and simulated distribution forecasts. R package version 0.9.1

Jun D, Zhou B (2011) A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. Mon Weather Rev 139(10):3284–3303

Kim HH, Swanson NR (2014) Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. J Econometr 178(2):352–367

Kirtman BP, Min D, Infanti JM, Kinter JL, Paolino DA, Zhang Q, van den Dool H, Saha SMPM, Emily B, Peitao P, Patrick T, Jin H, David GD, Michael KT, Anthony GB, Shuhua L, Anthony R, Siegfried DS, Michele R, Max S, Zhao EL, Jelena M, Young-Kwon L, Joseph T, Kathleen P, William JM, Bertrand D, Eric FW (2014) The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction, phase-2 toward developing intra-seasonal prediction. Bull Am Meteorol Soc 95:585–601

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23(10):2739–2758

Krakauer NR (2012) Estimating climate trends: application to United States plant hardiness zones. Adv Meteorol 2012:404876

Krakauer NY, Devineni N (2015) Up-to-date probabilistic temperature climatologies. Environ Res Lett 10(2):024014

Krakauer NY, Fekete BM (2014) Are climate model simulations useful for forecasting precipitation trends? Hindcast and synthetic-data experiments. Environ Res Lett 9(2):024009

Krakauer NY, Grossberg MD, Gladkova I, Aizenman H (2013) Information content of seasonal forecasts in a changing climate. Adv Meteorol 2013:480210

Krakauer NY, Puma MJ, Cook BI (2013) Impacts of soil-aquifer heat and water fluxes on simulated global climate. Hydrol Earth Syst Sci 17(5):1963–1974

Krakauer NY (2016) SeFo: a package for generating probabilistic forecasts from NMME predictive ensembles. J Open Res Softw 4(1):e19

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhan ZC, Williford E, Gadgil S, Surendran S (1999) Improved weather and seasonal climate forecasts from multimodel superensemble. Science 285(5433):1548–1550

Ma F, Ye A, Deng X, Zhou Z, Liu X, Duan Q, Jing X, Miao C, Di Z, Gong W (2016) Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. Int J Climatol 36(1):132–144

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Manag Sci 22(10):1087–1096

Matthes H, Rinke A, Dethloff K (2015) Recent changes in Arctic temperature extremes: warm and cold spells during winter and summer. Environ Res Lett 10(11):114020

Mishra V, Ganguly AR, Nijssen B, Lettenmaier DP (2015) Changes in observed climate extremes in global urban areas. Environ Res Lett 10(2):024005

Moral-Benito E (2015) Model averaging in economics: an overview. J Econ Surv 29(1):46–75

Muller RA, Curry J, Groom D, Jacobsen R, Perlmutter S, Rohde R, Rosenfeld A, Wickham C, Wurtele J (2013) Decadal variations in the global atmospheric land temperatures. J Geophys Res 118D(11):5280–5286

Peng RD, Bobb JF, Tebaldi C, McDaniel L, Bell ML, Dominici F (2011) Toward a quantitative estimate of future heat wave mortality under global climate change. Environ Health Perspect 119(5):701–706

Pieroth M (2014) Non-gaussian forecast skill in ensemble prediction systems. Master's thesis, Goethe Universität

Rahmstorf S, Coumou D (2011) Increase of extreme events in a warming world. Proc Natl Acad Sci (USA) 108(44):17905–17909

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108(D14):4407

Rohde R, Muller RA, Jacobsen R, Muller E, Perlmutter S, Rosenfeld A, Wurtele J, Groom D, Wickham C (2013) A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinform Geostat Overv 1(1):1000101

Rohde R, Muller R, Jacobsen R, Perlmutter S, Rosenfeld A, Wurtele J, Curry J, Wickham C, Mosher S (2013) Berkeley Earth temperature averaging process. Geoinform Geostat Overv 1(2):1000103

Sherwood SC, Huber M (2010) An adaptability limit to climate change due to heat stress. Proc Natl Acad Sci (USA) 107(21):9552–9555

Slater JL, Villarini G, Bradley AA (2016) Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. Clim Dyn:1–16. doi:10.1007/s00382-016-3286-1

Smith LA, Suckling EB, Thompson EL, Maynard T, Du H (2015) Towards improving the framework for probabilistic forecast evaluation. Clim Change 132(1):31–45

Tebaldi C, Hayhoe K, Arblaster JM, Meehl GA (2006) Going to the extremes. Clim Change 79(3–4):185–211

Tödter J (2011) New aspects of information theory in probabilistic forecast verification. Master's thesis, Goethe University

Tingley MP, Huybers P (2013) Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. Nature 496(7444):201–205

Valtorta SE (2002) Animal production in a changing climate: impacts and mitigation. In: 15th Conf. on Biometeorology/Aerobiology and 16th International Congress of Biometeorology

van der Velde M, Wriedt G, Bouraoui F (2010) Estimating irrigation use and effects on maize yield during the 2003 heatwave in France. Agric Ecosyst Environ 135:90–97

Vogel MM, Orth R, Cheruy F, Hagemann S, Lorenz R, van den Hurk BJJM, Seneviratne SI (2017) Regional amplification of projected changes in extreme temperatures strongly controlled by soil moisture-temperature feedbacks. Geophys Res Lett 44(3):1511–1519

Wanders N, Wood EF (2016) Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environ Res Lett 11(9):094007

Wang QJ, Schepen A, Robertson DE (2012) Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. J Clim 25(12):5524–5537

Weisheimer A, Doblas-Reyes FJ, Jung T, Palmer TN (2011) On the predictability of the extreme summer 2003 over Europe. Geophys Res Lett 38:L05704

Zaitchik BF, Evans JP, Geerken RA, Smith RB (2007) Climate and vegetation in the Middle East: interannual variability and drought feedbacks. J Clim 20:3924–3941

Zaitchik BF, Macalady AK, Bonneau LR, Smith RB (2006) Europe's 2003 heat wave: a satellite view of impacts and landatmosphere feedbacks. Int J Climatol 26(6):743–769

Zampieri M, D'Andrea F, Vautard R, Ciais P, de Noblet-Ducoudré N, Yiou P (2009) Hot European summers and the role of soil moisture in the propagation of Mediterranean drought. J Clim 22:4747–4758