Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations

2016 Environ. Res. Lett. 11 094007

(http://iopscience.iop.org/1748-9326/11/9/094007)

Environmental Research Letters

LETTER

Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations

Niko Wanders and Eric F Wood

Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544, USA

E-mail: nwanders@princeton.edu

Keywords: sub-seasonal forecasting, NMME phase 2, extreme events, weighted ensemble mean, global

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
5 February 2016

REVISED
12 July 2016

ACCEPTED FOR PUBLICATION
5 August 2016

PUBLISHED
31 August 2016

Abstract

Sub-seasonal to seasonal weather and hydrological forecasts have the potential to provide vital information for a variety of water-related decision makers. Here, we investigate the skill of four sub-seasonal forecast models from phase-2 of the North American Multi-Model Ensemble using reforecasts for the period 1982–2012. Two weighted multi-model ensemble means from the models have been developed for predictions of both sub-seasonal precipitation and temperature. By combining models through optimal weights, the multi-model forecast skill is significantly improved compared to a 'standard' equally weighted multi-model forecast mean. We show that optimal model weights are robust and the forecast skill is maintained for increased length of time and regions with a low initial forecast skill show significant skill after optimal weighting of the individual model forecast. The sub-seasonal model forecasts models show high skill over the tropics, approximating their skill at monthly resolution. Using the weighted approach, a significant increase is found in the forecast skill for dry, wet, cold and warm extreme events. The weighted mean approach brings significant advances to sub-seasonal forecasting due to its reduced uncertainty in the forecasts with a gain in forecast skill. This significantly improves their value for end-user applications and our ability to use them to prepare for upcoming extreme conditions, like floods and droughts.

Introduction

Flood and drought events occur in all regions of the world with large societal impact (Kundzewicz and Kaczmarek 2000). Early-warning decision support systems can help to reduce the societal vulnerability to these hydrological extreme events. Such systems rely on high-quality real-time hydrological forecasts, which are provided by a combination of meteorological forecasts and hydrological modelling. Because the hydrological forecasts rely heavily on meteorological input data, low skill and high uncertainty in the meteorological forecast results in a decrease in hydrological forecast skill and an inability to offer meaningful early warnings for anticipated extreme events (Wanders and Wada 2015). For short-range forecasts (up to 14 days), high resolution (both in space and time) skilful weather model forecasts are available from a number of centres (Fan and Van den Dool 2011, Magnusson and Källén 2013). However, to increase preparedness and reduce vulnerability to hydrological extremes it is important to extend the forecast range beyond the two week period. Applications that will benefit from this extended forecast information include amongst others, crop modelling (Ray *et al* 2015), flood (Wanders *et al* 2014) and drought forecasting (Sheffield *et al* 2014) and planning of reservoir operation (Demargne *et al* 2014). Seasonal forecast models, ranging from 14 days to one year, bridge the gap between climate and weather models, and are available at a coarser resolution and lower temporal resolution (typically monthly timescale and 1° spatial resolution). The needs of water managers and other end users are to have forecasts of extreme hydrological conditions beyond the first two weeks at high temporal resolution, which has stimulated increased interest from the hydrological community

for high quality meteorological sub-seasonal forecasts at the daily scale (Kirtman *et al* 2014).

This is now available from the North American Multi-Model Ensemble phase 2 (NMME-2) project that has provided sub-seasonal forecasts for a 31 year period at a daily temporal resolution (Kirtman *et al* 2014). This project is the follow-up from NMME phase 1 that provided multi-model seasonal forecasts at a monthly temporal resolution. The sub-seasonal climate forecasts, due to their increased temporal resolution (from monthly to daily), can significantly improve seasonal hydrologic forecasts. However, to assess the added value of NMME-2 for hydrology it is important to understand the skill of this product in forecasting meteorological conditions. The uncertainty in the meteorological forecasts will be imposed upon the hydrological simulations and will impact the skill of the hydrological forecasts.

The objective of this study is to assess the sub-seasonal forecast skill of the NMME-2 ensemble for daily precipitation and temperature over 22 global regions (Giorgi and Francisco 2000). We use the output from four available NMME-2 models to analyse sub-seasonal forecast skill in daily precipitation and 2 m air temperature for these 22 regions for the period 1982–2012 (Kirtman *et al* 2014). The models used in the analysis are: Canadian Coupled Climate Model version 3 and 4 (CanCM3, CanCM4, Merryfield *et al* 2013), the Forecast-oriented Low Ocean Resolution (FLOR-B01, Vecchi *et al* 2014) and the Community Climate System Model (CCSM4, Hurrell *et al* 2013). In addition, we evaluate the sub-seasonal forecast skill of a multi-model ensemble mean forecast constructed using three different approaches: equally weighted individual mean model forecasts, optimally weighted mean model forecasts with the model weights constrained to $\geqslant 0$, and optimally weighted mean model forecasts with unconstrained model weights.

Methods

Seasonal forecast models

In this study four available sub-seasonal forecast models were used for the period 1982–2012 (table S1). These models are part of the NMME-2 forecast ensemble and provide hindcast initialised every month with a daily temporal resolution (Kirtman *et al* 2014). All models provide daily precipitation and daily mean 2 m air temperature at a 1° spatial resolution. More details on the individual models can be found in their corresponding documentation (Hurrell *et al* 2013, Merryfield *et al* 2013, Vecchi *et al* 2014). In total 31 years of forecast data have been evaluated, where one forecast is issued every month, leading to a total of 372 re-forecasts per model (table S1). The archive for CCSM4 covered only 78% of the period, leading to a reduced 290 available forecasts. A bi-weekly temporal aggregation was used to estimate the sub-seasonal forecast skill, while a monthly temporal aggregation is used for the seasonal forecast skill.

Reference dataset

To validate the (sub)-seasonal hindcasts made in NMME-2 we used an independent observation based reference dataset, the Princeton Global Forcing (Sheffield *et al* 2006). This dataset covers the NMME-2 hindcast period 1982–2012 and is available with a daily temporal and 1° spatial resolution globally. The monthly average values are derived from observations (*in situ* and satellite) statistically downscaled (temporally) by combining high resolution observations (e.g. satellite precipitation) with NCEP-NCAR reanalysis that is not part of the multi-model seasonal forecast system. To the extent possible, this ensures that the reference dataset is as fully independent as feasible to validate the seasonal predictions. The Princeton Global Forcing dataset has proven to be a reliable and widely used dataset.

Weighted multi-model ensemble mean

The first approach assigns equal weights to each model ensemble mean forecasts and so ignores prior knowledge of their forecasts skill. This is one of the most commonly used procedures in seasonal forecasting (Krishnamurti *et al* 1999), where it is assumed that the equally weighted forecast will provide the best forecast for future conditions, since each model is equally likely to represent the truth. In the second and third approaches a multi-model forecast is developed based on weighing each model forecast according to its skill to forecast observations for a given initialisation month and forecast lead time. In the second approach, the weights are constrained to be $\geqslant 0$ while in the third approach this constraint is removed. The implication of zero-weights is that models are removed from the ensemble due to lack of skill, while negative weights indicate that they show a consistent negative skill in the hindcast period. Using step-wise regression, the models that explain the largest portion of the observed variance are favoured over models that can only explain minor or identical parts of the observed variance. This approach has been tested in a modified approach on low resolution seasonal forecasts (DelSole 2007, DelSole *et al* 2013), synthetic experiments (Weigel *et al* 2008), sea surface temperature (Peña and van den Dool 2008), weather forecasts (Kharin and Zwiers 2002, Casanova and Ahrens 2009) and climate reanalysis simulations (Haughton *et al* 2015) with mixed results, but hasn't been applied for high-resolution, operational sub-seasonal ensemble forecasting models.

In this study, we applied the standard cross-validation procedure often referred to as 'leave one out cross-validation' to each analyzed hindcast (see Wilks 2006 section 6.4.4) to assess model skill. Here,

one year of the hindcast is held back (the target year) and the weights computed on the remaining 30 years. The target year is then forecast. The target year is shifted, year by year, with the weights recomputed from the remaining years each time until all 31 years are covered. This provides 31 target-year forecasts, each not included in the computation of the optimal model weights, which are used to assess the forecast skill of the various multi-model prediction systems.

To generate the weighted multi-model ensemble mean for a region, multivariate linear regression is used to estimate the optimal weights. The ensemble weighted mean is given by:

$$Y(m, l, t) = \sum_{i=1}^{I} \alpha_{i,m,l} * \overline{X_{i,m,l}(t)},$$

where, $Y(m, l, t)$ is the weighted ensemble mean for that region at time/forecast year t that is calculated by multiplying the ensemble mean of model, $\overline{X_{i,m,l}(t)}$, by the weight obtained from the multivariate regression ($\alpha_{i,m,l}$), for model i, forecast initialisation month m and a lead time l. The weights are determined for every forecast initialisation month and lead time separately due to the varying skill of the models over the seasons. While the anomalies in temperature have no constraints (due to the continuous distribution of temperature), the range of precipitation anomalies is limited, by conditioning that precipitation should exceed or equal zero precipitation, to ensure valid forecasts.

The uncertainty in all generated weighted ensemble forecasts is obtained from the variance within the model ensemble members and the covariance among the models ensemble means, following the variance calculation for a multi-variate linear regression. Including the model variances ensures a realistic representation of the ensemble uncertainty and, in general, prevents over confident forecasts. By taking into account the covariance among models, models with a shared heritage (e.g., CanCM3 and CanCM4) will not dominate the ensemble mean when they are over represented in the total ensemble. The forecast uncertainty is given by:

$$\begin{aligned} \mathrm{var}(Y)(m, l) = & \sum_{i=1}^{I} \alpha_{i,m,l}{}^{2} * \mathrm{var}(X_{i,m,l}) \\ & + \sum_{j=1}^{J} 2 * \alpha_{i,m,l} * \alpha_{j,m,l} \ \mathrm{cov} \\ & \times (\overline{X_{i,m,l}}, \ \overline{X_{j,m,l}}), \end{aligned}$$

where the variance of each model at a specific initialisation month m at lead time l is determined from the spread between the individual ensemble members and the covariance is determined from the covariance between the individual model ensemble means. By using the provided individual variances and covariance between the models, a more realistic estimate can be made of the ensemble spread at given leads and given initialisation months.

Bootstrapping procedure

To analyse the robustness of the obtained weighted means a bootstrapping experiment was designed. From the 31 year period covered by NMME-2, x random years (ranging from 6 to 30) are selected to generate the weights (α) that are then used to generate the optimally weighted ensemble means (Y). Y is validated against the remaining years of observations derived from the Princeton Global Forcing dataset, that were not used to derive to coefficients of the constrained and unconstrained weights. This will prevent over-fitting of the multi-variate linear model and produce an independent validation dataset. The procedure is repeated 100 times for each x random years for all 22 regions and the anomaly correlations derived to quantify the predictive skill. The range of anomaly correlations is compared to the equally weighted ensemble mean and the individual models, to access the forecast skill. For the equally weighted ensemble, the skill does not change with increasing the number of random years, because the weights remain constant. The bootstrapping procedure is performed separately for precipitation and temperature.

Model forecast skill evaluation

The skilfulness of sub-seasonal forecast models is evaluated using the Brier score (BS) (Brier 1950), where the BS is given by:

$$\mathrm{BS}(\mathrm{lim}, \ l) = \frac{1}{T} \sum_{t=1}^{T} (P(X_{\mathrm{lim},l}(t)) - \mathrm{sgn}(\mathrm{obs}_{\mathrm{lim}}))^{2},$$

where, $\mathrm{sgn}(\mathrm{obs}_{\mathrm{lim}})$ indicates a binary value, indicating whether the observation exceeds the event threshold (lim, fraction ranging from 0 to 1), $P(X_{\mathrm{lim},l}(t))$ provides the probability values for every model forecast at lag (l) and time t to exceed the limit. As a benchmark BS for comparison a climatological forecast (i.e. a forecast that reflect the probability of an event happening at any given moment) as the reference BS. From the climatology the reference BS can be derived by:

$$\mathrm{BS}_{\mathrm{ref}}(\mathrm{lim}) = \mathrm{lim}(1 - \mathrm{lim})$$

which is identical to the uncertainty in the decomposed BS. When the reference BS is larger than the uncertainty the forecast is skilful and shows a higher skill than the chance forecast reference. The Brier skill score (BSS) is then given by:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}},$$

where BSS can range from $-\infty$ to 1, where 1 is a perfect forecast and all negative values indicate an unskilled forecast.

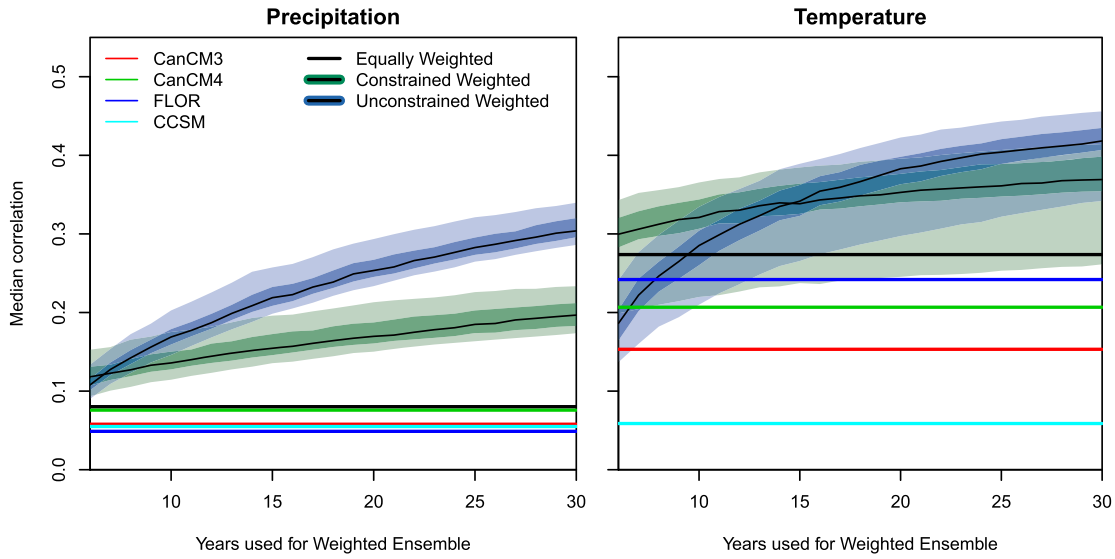

Figure 1. Anomaly correlation over all regions in a bootstrapping experiment to determine the robustness of the constrained and unconstrained weighted ensemble mean. Horizontal lines indicate the forecast skill of separate models and the 'standard' equal weight ensemble mean. The correlations are plotted as a function of the numbers of years that were used to calculate the weights, where the validation of these weights is always performed on the remaining years of forecasts. Shaded areas (dark, 50%, light 95% interval) indicate the range of obtained correlations with the weighted means and the lines indicate the median correlation.

Results

Weighted ensemble mean

To ensure that the obtained weights for the constrained ensemble means are stable in time and do not dependent on specific event or years, a sensitivity analysis is performed to evaluate the robustness of the weights. Using a bootstrapping analysis, a number of years (ranging between 6 and 30) is selected to compute the weights, which is then validated on the remaining years and compared against the equally weighted mean and individual model skills. For both precipitation and temperature, we find that both optimally weighted ensemble means to outperform the individual models as well as the 'standard' equally weighted mean in terms of anomaly correlation (figure 1). For a superior sub-seasonal forecast using optimal weights, 6 years of precipitation hindcast are required, while mean temperature forecast require a minimum of 11 years of hindcasts data due to the higher initial forecast skill for temperature anomalies. Using more than 25 years of forecast data has little impact on the skill of the constrained weighted means, since the weights change little as does the forecast skill. This indicates that the use of optimal weights will always be superior to the assumption that models should be equally weighted; a finding that has a direct impact on the way multi-model forecasts should be constructed. The constrained weighted multi-model mean has a higher initial forecast skill compared to the unconstrained counterpart when limited sampling is used, which would correspond to having a small hindcast data set. This arises due to the increased degrees of freedom in the fitting of the unconstrained weights with the limited number of sampling years that result in the estimated weights being overconfident and thus lead to a poorer performance.

To ensure that the assigned weights are robust, we calculated the difference between the parameters obtained after parameter fitting with the full data record and subsets of the hindcast data ranging from 6 to 30 years. Weights were found to be robust and show a decreasing normalised mean difference with increasing number of sampling years (figure S1). These results ensured that applied methodology results in robust parameters and parameters are not overfitted to match the observations. The finding that the forecast skill from the optimally weighted multi-model mean forecast is higher than that of any of the individual models or the equally weighted multi-model mean forecast carries over to individual, regional forecasts (figure S2).

Seasonal forecast skill

Forecast skill for seasonal precipitation is not equally distributed over the globe (figures 2(a)–(c) and S3) with high seasonal forecast skill (monthly temporal aggregation) over the tropics (e.g. Amazon, Indonesia). In general, the CanCM and CCSM models have higher skill than FLOR for short lead times, thereafter, FLOR gains in forecasting skill. This is probably due to the lack of initialisation of land states in FLOR but rather uses its land surface climatology from an AMIP simulation when initialising its forecasts (Gates *et al* 1998, Jia *et al* 2015). For seasonal temperature forecasts, the model skill is higher and the difference amongst models is reduced (figures 2(d)–(f) and S4). The tropics show up as areas of high forecast skill

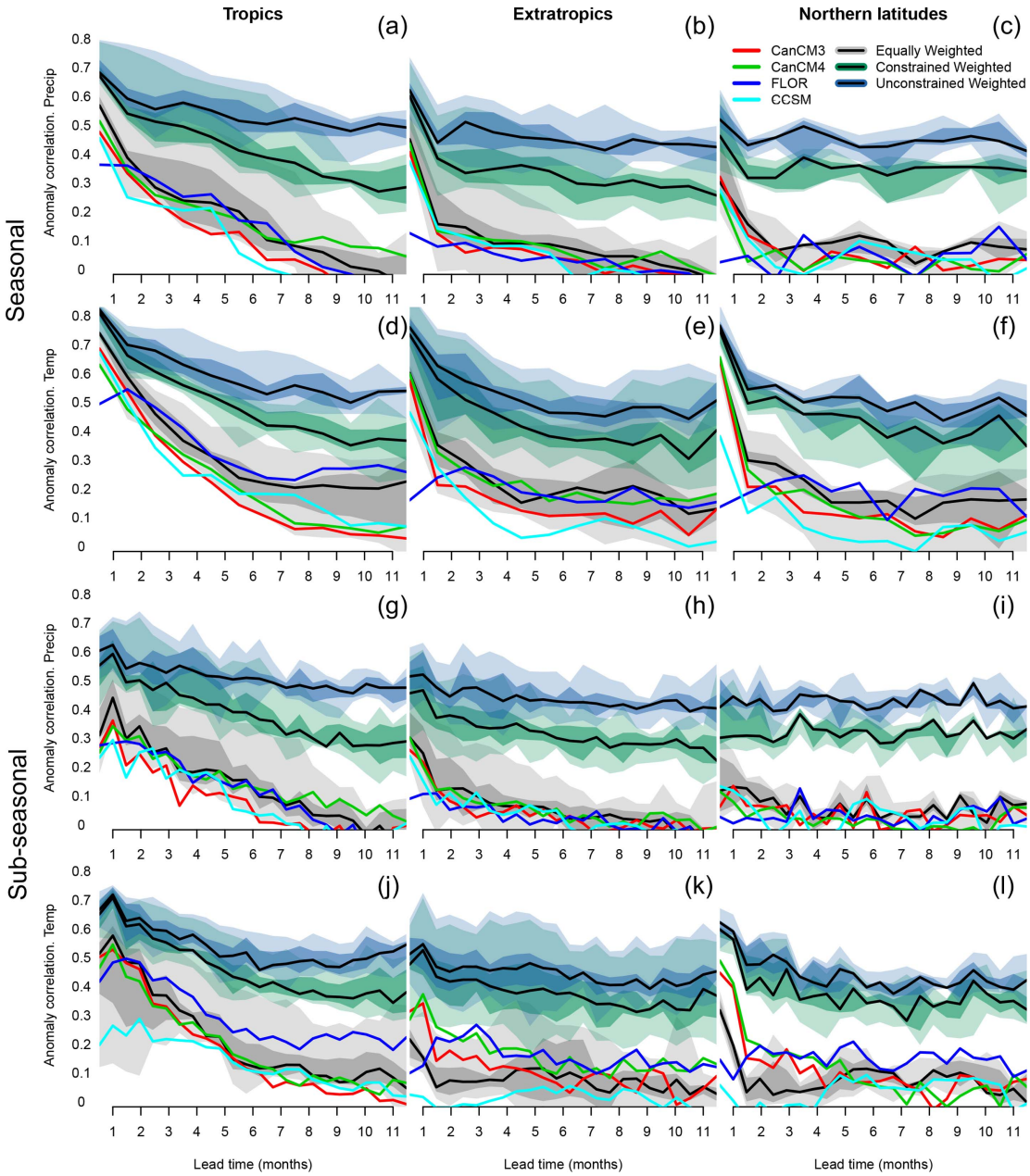

Figure 2. Anomaly correlation between seasonal forecasted and observed precipitation (a)–(c) and temperature (d)–(f) and sub-seasonal precipitation (g)–(i) and temperature (j)–(l) anomalies. Correlations are derived over the tropics (left), extratropics (middle) and northern latitudes (right, areas above 60N) and grouped for each lead time. Uncertainty band are given for the ensemble means, where the inner (darker) band indicates the middle 50% and the outer band indicates the 95% interval of the obtained anomaly correlations.

while, in contrast to the precipitation forecasts, relatively high skill is found in the northern latitudes (locations above 60°N). Both optimally weighted multi-model ensemble mean forecasts significantly outperform the individual model forecast skill (one-sided t-test, $\alpha = 0.05$), with the equally weighted ensemble mean forecast being similar to the best individual model's skill. For long lead times this superiority of the optimally weighted unconstrained multi-model mean forecast is even more skilful relative to the individual models (or their equally weighted forecast). This is due to using information on the (negative) skill of models to make more skilful forecasts. An example is provided for the 6 month lead precipitation forecast over South Africa, where the skill of CanCM3 is negative (figure S3). In this case the weighted ensemble uses a large negative weight for the CanCM3 forecast to obtain positive skill at a 6 month lead in that region.

Sub-seasonal forecast skill

The sub-seasonal (bi-weekly temporal aggregation) forecast skill is found to be similar to the seasonal forecast skill, especially at leads below 6 months (figures 2(g)–(l)). There is a large agreement amongst the models on the regional forecast skill and the rate in

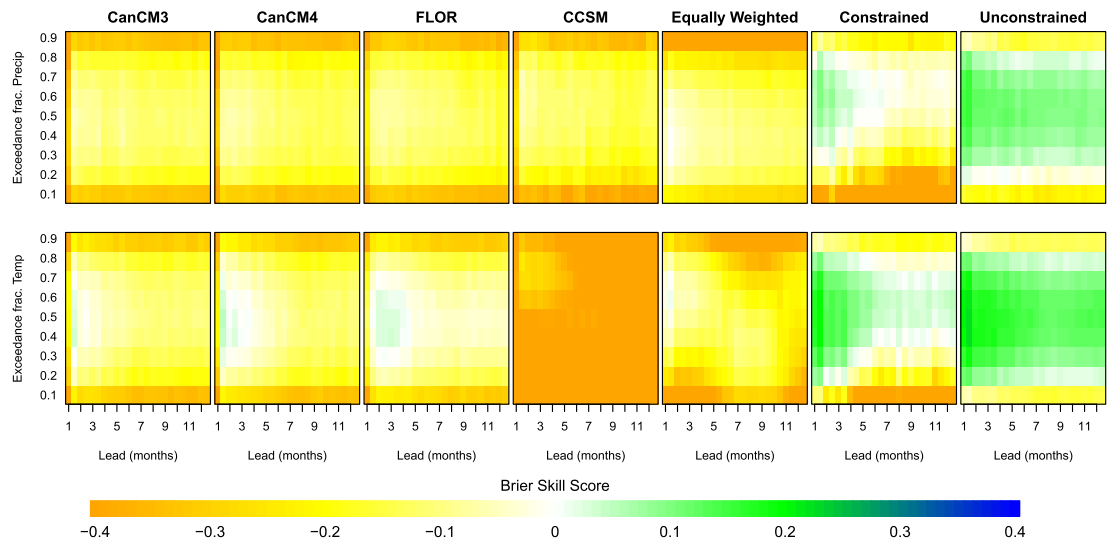

Figure 3. Brier skill score (BSS) of four sub-seasonal forecast models and the ensemble means for different thresholds and forecast lags combined for all regions used in this study. The skill is determined by calculating the Brier score (BS) for each forecast and compare that to the climatological predictability of an event (BS_{ref}, see methods for additional information). Negative values of the BSS indicate forecasts worse than the climatological predictability and 1 indicates a perfect forecast.

which the skill is declining at longer lead times. The forecast over the tropical regions again show high skill, while the skill in the northern latitudes is quickly lost or already absent after the first weeks. Again, a significant gain in skill is found by using the optimally weighted multi-model ensemble mean forecast. In some cases, the forecasts skill in the equally weighted mean deteriorates due to the poor skill of individual models (also shown in figures S5 and S6), leading to insignificant positive correlations (two-sided t-test, $\alpha = 0.05$). The high skill levels found from the sub-seasonal optimally weighted multi-model forecasts are promising for hydrological forecasting and related applications that depend on making high temporal resolution forecasts, like the predictions of heat waves, drought events, pluvial periods or forecasting of crop water requirements.

Forecasting extremes

Before these sub-seasonal forecasts are used for applications, the forecast skill based on anomaly correlations and the anomaly forecasts need to be transformed into useful information for end users such as water managers. Here, we focus on the BS for specific events and exceedance thresholds for both precipitation and temperature when compared to the climatological predictability (BS_{ref}) using the BSS. As expected, the forecast models show the highest skill in the first months after initialisation and for intermediate thresholds (figure 3). For the extreme thresholds (e.g. Q10/low or Q90/high precipitation totals), the skill is lower and decreases quickly. The skill is strongly affected by the large ensemble spread of some of the models (e.g. figure S2), indicating that the precision is low, which strongly impact the BSS in a negative way. The equally weighted multi-model mean forecasts shows low skill when forecasting temperature extremes due to the low skill of one of the models; while this has little impact on the optimally weighted multi-model forecasts.

The sub-seasonal forecast skill for accumulated precipitation is provided for the 2011 Horn of Africa drought (figure 4). This drought was significant for the fact that it was poorly forecasted, which in turn resulted in a substantial number of fatalities and economic damage. The hindcasts clearly reflect the low sub-seasonal model forecast skill for the drought. In fact, some model forecasts show strong negative correlations and predict pluvial conditions instead of drought. As a result, the equally weighted, ensemble mean forecast shows poor sub-seasonal forecast skill, while the optimally weighted ensemble mean forecast developed in this paper shows a strong forecast skill. For this example, the observations from this particular year have been excluded from the multivariate linear regression to prevent overfitting of the data. The constrained ensemble mean forecast has a high weight for the FLOR forecasts, which are in general skilful for this region. In fact, the FLOR model outperform the constrained ensemble mean forecast in this scenario. The unconstrained weighted mean takes advantage of its flexibility and uses the skill of FLOR and the consistent negative skill of CanCM4 to make a more accurate forecast. The uncertainty in the weighted mean forecast is low (e.g. compared to FLOR), yet the forecasts are not over-confident and the observed anomalies are within the forecast uncertainty range.

Another example is provided for the Brazil floods in January 2011, where all models show a dry forecast whereas the unconstrained weighted ensemble multi-

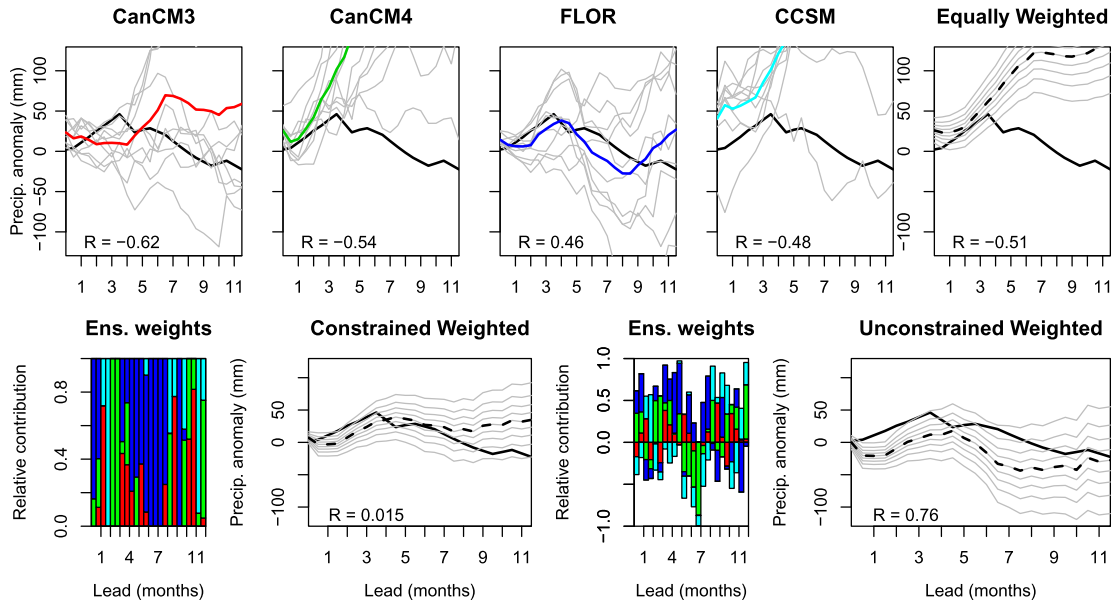

Figure 4. Accumulated precipitation anomalies for Eastern Africa for March 2011, for the 2011 horn of Africa drought. Top rows indicate the individual sub-seasonal model forecasts (ensemble mean, bold colour lines), individual ensemble members (grey lines) and the observed anomalies (bold black line). Bottom row indicates optimal weights applied to the individual models and the resulting multi-model forecast (dashed line) and observations (solid line). Grey lines indicate confidence intervals at 10% increments. The total of the ensemble weights equals 1, and negative weights indicate that the original forecast anomaly is multiplied by a negative weight in the forecasted weighted ensemble. For all forecasts, anomaly correlation between the forecasts and observed anomalies are provided in the bottom left of each panel. The target year (in this case 2011) is not used in estimating the model weights.

model forecast predicts potential flooding (figure S7). For this forecast the constrained weighted mean forecast is clearly limited by the low forecast skill of the individual model forecast, however, this forecast still outperformance any of the individual model forecasts or the equally weighted multi-model forecast.

Both examples show that the uncertainty in the cumulative forecast is reduced by the use of the optimally weighted multi-model ensemble mean forecast while the correlations show an increase between forecasts and observations. This higher skill and lower forecast uncertainty provides higher confidence for decision makers to act when faced with an extreme event forecast.

Discussion and conclusions

From our analysis, the individual NMME phase 2 sub-seasonal forecasts and an equally weighted multi-model mean forecasts show strong skill over the tropics, while the forecasts for the extratropics and northern latitudes only show skill at shorter leads. The optimally weighted multi-model mean forecasts show a higher skill in general and in particular higher skill is found for the longer lead times in the extratropics, with the optimally weighted unconstrained multi-model forecasts being most skilful overall. These skilful sub-seasonal forecasts provide new opportunities for end users that rely on sub-seasonal forecasts and forecasts with a daily temporal resolution, such as water managers and reservoir operators. The results obtained in this study clearly show that some models are more skilful over specific regions and for either precipitation or temperature, and thus should be used accordingly in a multi-model forecast system. This information is exploited in the creation of the optimally weighted, multi-model ensemble mean forecasts, where skilful forecast models are given a higher weight in the ensemble mean. Apart from utilising the information on the positive skill of the models, negative skill of the models can be used in the unconstrained weighted multi-model forecasts. Even though individual model forecast skill can be better than an equally weighted ensemble mean, the results from this study suggest that the newly applied constrained and unconstrained optimally weighted multi-model mean forecasts are overall superior to individual models and the equally weighted ensemble mean forecast.

The positive impact of using unconstrained weights in this study illustrates that for some scenarios or combinations of lead time and region, a sub-seasonal forecast model can show consistent negative predictive skill. This is not a desired scenario for real-life applications, but it does provide scientists with insights that could help during model improvements (e.g. understanding of important teleconnections that may lead to improved sub-seasonal skill for specific regions). Negative anomaly correlations were found in earlier studies on seasonal predictions skills (e.g. Jia *et al* 2015), and it was suggested that the consistent negative skill could be related to imperfect initial conditions (Wang *et al* 2010). Although indicating that a

model has a consistent negative anomaly correlation (and thus assigning it a negative or zero-weight) is often not seen as a positive outcome, we want to take a positive view and use that information to our advantage to improve ensemble seasonal forecasts. In scientific discussions we had with peers and experts in the field, it was clear that the use of negative weights is not a desirable (long-term) solution for some scientists. However, most agree that it can provide an a (short-term) solution to provide added value to seasonal forecast applications and help to identify areas where skill is currently lacking.

In this study, we found that a weighted ensemble approach outperformed an equally weighted ensemble in contrast to an earlier study by Kharin and Zwiers (2002). They show that no additional skill was found in a weighted ensemble approach for their super-ensemble compared to a standard equally weighted ensemble. One of the main reasons for this difference could be the extended period that is covered by the NMME phase 2 hindcast dataset compared to the 10 year period in Kharin and Zwiers (2002). Following the results from figure 1, a 10 year period is too short for computing stable and significantly improved results for a weighted ensemble, which confirms their findings. Additionally, temperature and precipitation anomalies were studied here, compared to 500 hPa geopotential height in Kharin and Zwiers (2002). DelSole *et al* (2013) showed that in most of the world no significant gain can be found when a weighted multi-model approach is implemented for 2 m air temperature and precipitation at a finer resolution of 2.5° grid. Although, the study of DelSole *et al* (2013) is more in line with this work (a 46 year hindcast period and identical variables), the grid by grid approach does not provide a spatially consistent pattern in the predictive skill assessment of the models and the model covariance estimates. In this study, we used 22 global regions, instead of a grid by grid comparison, to reduce the noise in the estimates of model predictive skill and covariance. This leads to stable coefficients that results in a more consistent performance of the weighted ensemble approach, which could be the cause for the difference between the two studies.

Given the results from this study and the comparison with earlier work, we argue that to successfully implement the optimally weighted ensemble multi-model forecast system, two requirements need to be met. The observations and the independent models must show some degree of correlation (either negative or positive), and a sufficient number of historic forecasts (minimum of 10 years) is required to accurately assign stable weights. We argue that a leave-one-out cross validation is the best strategy to determine if the weights are stable and a consistent performance of the weighted ensemble approach is found. Finally, we recommend that the forecasts are spatially upscaled to a spatial resolution at which the predictive skill of the models is spatially consistent.

We have shown that in the case of the NMME-2 forecasts, sufficient data is present to generate a stable performance in the weighted mean (figure 1) and that the individual models are correlated to the observations (figures 2 and S3–S6). Another advantage of the optimal weighting is that the cross-correlations between models can be used to prevent the generation of a biased ensemble mean (e.g. over-representation by models with a similar heritage). Finally, it is shown that the occurrence of extreme events can be forecasted with a higher accuracy (figures 3, 4 and S7) than previously obtained, while the uncertainty in the forecast is reduced, hence improving their usability for operational systems. For specific purposes, one could also optimise the weighting of the models to the applications that they will be used for. When end-users are very interested in forecasting drought, the weights can be optimised to favour models that show a higher forecast skill in forecasting such extreme events.

The next step forward will be to implement the output of these sub-seasonal forecasts and the newly created ensemble mean into decision support systems to assess their quality for end-user applications. This could significantly improve the usability of sub-seasonal forecasts and their impact on decision making and hazard prevention measures. The weighted ensemble approach could also be implemented for all other ensemble forecasting systems as it shows a high potential for operational forecasting systems and could help to advance seasonal forecasting in general.

Acknowledgments

We would like to acknowledge four anonymous reviewers that helped to improve the manuscript. NW was supported by a NWO Rubicon Fellowship 825.15.003 (Forecasting to Reduce Socio-Economic Effects of Droughts) and EFW was supported by the NOAA Climate Program Office under grant NA15OAR4310075 (Assessing NMME Phase-2 Forecasts for Improved Predictions of Drought and Water Management). The forecast data from the North American Multi-Model Ensemble phase 2 are freely available at http://earthsystemgrid.org/.

References

Brier G W 1950 Verification of forecasts expressed in terms of probability *Mon. Weather Rev.* **78** 1–3

Casanova S and Ahrens B 2009 On the weighting of multimodel ensembles in seasonal and short-range weather forecasting *Mon. Weather Rev.* **137** 3811–22

DelSole T 2007 A bayesian framework for multimodel regression *J. Clim.* **20** 2810–26

DelSole T, Yang X and Tippett M K 2013 Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Q. J. R. Meteorol. Soc.* **139** 176–83

Demargne J *et al* 2014 The science of NOAA's operational hydrologic ensemble forecast service *Bull. Am. Meteorol. Soc.* **95** 79–98

Fan Y and Van den Dool H 2011 Bias correction and forecast skill of NCEP GFS ensemble week-1 and week-2 precipitation, 2 m surface air temperature, and soil moisture forecasts *Weather Forecast.* **26** 355–70

Gates W L *et al* 1998 An overview of the results of the atmospheric model intercomparison project (AMIP I) *Bull. Am. Meterol. Soc.* **73** 1962–70

Giorgi F and Francisco R 2000 Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM *Clim. Dyn.* **16** 169–82

Haughton N, Abramowitz G, Pitman A and Phipps S J 2015 Weighting climate model ensembles for mean and variance estimates *Clim. Dyn.* **45** 3169–81

Hurrell J W *et al* 2013 The community Earth system model: a framework for collaborative research *Bull. Am. Meteorol. Soc.* **94** 1339–60

Jia L *et al* 2015 Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model *J. Clim.* **28** 2044–62

Kharin V V and Zwiers F W 2002 Climate predictions with multimodel ensembles *J. Clim.* **15** 793–9

Kirtman B *et al* 2014 The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction *Bull. Am. Meteorol. Soc.* **95** 585–601

Krishnamurti T N, Kishtawal C M, LaRow T E, Bachiochi D R, Zhang Z, Williford C E, Gadgil S and Surendran S 1999 Improved weather and seasonal climate forecasts from multimodel superensemble *Science* **285** 1548–50

Kundzewicz Z W and Kaczmarek Z 2000 Coping with hydrological extremes *Water Int.* **25** 66–75

Magnusson L and Källén E 2013 Factors influencing skill improvements in the ECMWF forecasting system *Mon. Weather Rev.* **141** 3142–53

Merryfield W J, Lee W-S, Boer G J, Kharin V V, Scinocca J F, Flato G M, Ajayamohan R S, Fyfe J C, Tang Y and Polavarapu S 2013 The Canadian seasonal to interannual prediction system: I. Models and initialization *Mon. Weather Rev.* **141** 2910–45

Peña M and van den Dool H 2008 Consolidation of multimodel forecasts by ridge regression: application to Pacific sea surface temperature *J. Clim.* **21** 6521–38

Ray D K, Gerber J S, MacDonald G K and West P C 2015 Climate variation explains a third of global crop yield variability *Nat. Commun.* **6** 5989

Sheffield J, Goteti G and Wood E F 2006 Development of a 50-yr high-resolution global dataset of meteorological forcings for land surface modeling *J. Clim.* **19** 3088–111

Sheffield J *et al* 2014 A drought monitoring and forecasting system for Sub-Sahara African water resources and food security *Bull. Am. Meteorol. Soc.* **95** 861–82

Vecchi G A *et al* 2014 On the seasonal forecasting of regional tropical cyclone activity *J. Clim.* **27** 7994–8016

Wanders N, Karssenberg D, de Roo A, de Jong S M and Bierkens M F P 2014 The suitability of remotely sensed soil moisture for improving operational flood forecasting *Hydrol. Earth Syst. Sci.* **18** 2343–57

Wanders N and Wada Y 2015 Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century *Geophys. Res. Lett.* **42** 10689–95

Wang W, Chen M and Kumar A 2010 An assessment of the CFS real-time seasonal forecasts *Weather Forecast.* **25** 950–69

Weigel A P, Liniger M A and Appenzeller C 2008 Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. R. Meteorol. Soc.* **134** 241–60

Wilks D S 2006 *Statistical Methods in the Atmospheric Sciences* 2nd edn (New York: Academic) p 627