Response to Reviewer 1

The manuscript aims at developing a statistically based seasonal precipitation forecast model for Western Ethiopia. The target area is separated into homogeneous regions by means of a k-means cluster analysis of summer precipitation amounts. Eight regions with similar precipitation variability are defined. For each of them, a linear regression based forecast model is calibrated. Results are compared with a general forecast for the entire region and are found to be superior. In a final step the forecast is downscaled to a high resolution grid, again by means of a liner regression approach. The target of the study is timely, since local precipitation predictions are often required for water management and planning, and the manuscript is well structured and easy to follow. However I have some serious concerns about the calibration and particularly the evaluation of the statistical model. Further I would recommend to give some detailed information on the climate characteristics of the cluster regions and the major large scale influences.

1) Introduction, clustering and different predictor variables: An introduction into the climate of the target region is missing. Further, a detailed analysis of the precipitation characteristics of each cluster would be a basis for the interpretation of the modeling results. Some of the precipitation time series in Fig. 5 look highly correlated. Are simple statistical techniques really able to forecast those slight differences? And if, which predictor variables are responsible for the spatial variations of precipitation in Western Ethiopia? An analysis of the predictor-predictant relationships for each cluster would not only give some insights into the model structure and the large scale climate mechanisms of the target area, but also help to support (or scrutinize) the results of the modeling exercise.

We thank the reviewer for the comment. To address it, firstly, we provide additional description of the climate in the study region. The texts are inserted into Section 2 “Application to western Ethiopia and objectives of the study” (Page 3 Line 2):

“Precipitation in western Ethiopia peaks in the summer with approximately 70% of annual total precipitation falling during the main raining season - also known as the *Kiremt* season spanning from June to September (JJAS). On average, the seasonal total precipitation in the study region is approximately 760 mm; however in the northwest, precipitation can exceed 1200 mm (Fig. 1a). Along with the high spatial variability in this mountainous region, the temporal variability is also significant with spatial-average seasonal total precipitation ranging from 650 mm in dry years up to 900 mm in wet years (Fig. 1b). These highly variable spatial and temporal precipitation patterns have made skillful seasonal predictions challenging, particularly at local scales (e.g. Gissila et al., 2004, Block and Rajagopalan, 2007).”

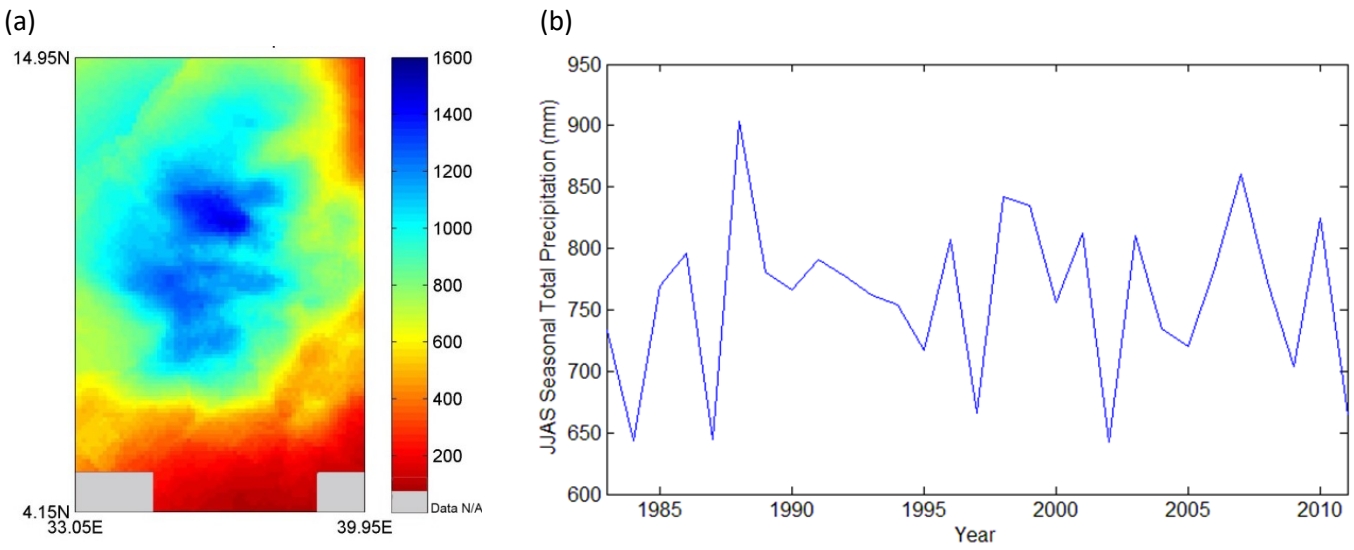


Figure 1: Spatial and temporal variability of June-September seasonal total precipitation in western Ethiopia: (a) spatial pattern of temporal-average, and (b) spatial-average time series.

Secondly, the precipitation characteristics of each cluster is described in a previous publication on cluster analysis (Zhang et al., 2016), however we agree that a brief summary should be provided here to help set the content and make the work more integrated. Therefore, we have added the text below to the original manuscript at the end of Section 3.1 "Cluster analysis" (Page 4 Line 26):

"The mean time series of each cluster illustrates high intra-correlation within the cluster and low inter-correlation between any two clusters, indicating strong coherency of the clustering results. For a detailed analysis including a complete correlation table and unique patterns for each cluster-level time series associated with large climate variables, readers are referred to Zhang et al. (2016)."

As we can see from the analysis in Zhang et al. (2016), some of the clusters contain stronger tele-connections to equatorial Pacific SSTs (an indicator of El Niño or La Niña), while other clusters are more affected by regional/local climate variables such as the pressure systems surrounding the African continent. This motivates us to use cluster analysis as a precursor to find proper predictors for each homogeneous region, which can capture the differences between the targeted predictand in each cluster.

Additional discussion on which predictor variables may be responsible for the spatial variations of precipitation in western Ethiopia is provided based on a combination of the previous concurrent connection to large-climate variable analysis (Zhang et al., 2016) and the prediction results in the discussion section (Page 15 Line 24):

"A previous study (Zhang et al., 2016) has shown that the influence of ENSO on JJAS precipitation in western Ethiopia decreases generally from north to south, and is likely one of the reasons why skills are relatively low in southwestern Ethiopia. Cluster 5 was also identified with the strongest connection to equatorial Pacific SST (Zhang et al., 2016), which is consistent with the highest skill found in this study. Other regions with low prediction skill show relatively strong connections to SST and pressure systems in neighboring oceanic regions. However, connections with those climate patterns appear to be less robust than with ENSO, making the predictions in their associated regions less skillful."

While additional analysis into the predictor-predictand relationship is clearly possible, we believe associating precipitation patterns with concurrent climate variables provides a solid inference into this relationship and strong support for explaining climate mechanism.

2) Calibration of the statistical model and overfitting: Correlations between crossvalidated modeling results and observations in the order of 0.7-0.85 are very high (in fact they exceed the skills of well known forecast models) and are questionable. I believe, that those results are due to overfitting (particularly due to the predictor selection). The predictors for each of the clusters are selected based on all years, the cross-validation is only performed for the calibration of the linear model. In order to fully evaluate the model skill, the predictor selection must be included in the cross-validation (i.e. chose predictors at each step of the cross validation, e.g. based on a correlation threshold). Most likely the model skill will significantly drop, I could imagine that a step wise selection of predictors might slightly improve the results.

We thank the reviewer for the insightful comment. We admit that the model is overfit due to our methodological procedure in the predictor selection process as the reviewer notes. Therefore, we re-performed the entire process including predictor selection and regression based on cross-validation; that is, we dropped the target year when creating the global correlation map to search for predictors. As a result, there are total of 1044 (29 x 9 x 4) global correlation maps given 29-year time-series, 8 clusters plus 1 non-cluster scenario, and 4 climate variables. Hence, we developed a program to help automatically select highly correlated and justifiable regions as predictors. A description of the method is added to Section 3.2 "Statistical modeling approach" (Page 6 Line 11) and also included here with new figures:

"To avoid overfitting, the entire process including predictor selection and statistical modeling is processed using cross-validation. To start, drop-one-year precipitation observations for JJAS averaged across the region and each cluster are spatially correlated independently with each global climate variable. As a result, there are total of 1044 global correlation maps given the 29-year time-series, eight clusters plus one non-cluster, and four climate variables. Hence, a program to automatically select highly correlated and justifiable regions as predictors is developed. The following steps describe the subsequent statistical modeling process (Fig. 3):

(1) Grid-cells within each justifiable region (e.g. equatorial Pacific; Fig. 4) with correlation above the 99% significance level are identified (Fig. 5).

(2) The top 10% of the identified grid-cells with the highest correlation in each region is then selected, in order to boost the potential model skill.

(3) For each region, data of the selected grid-cells within the region are spatially averaged (defined as "pre-predictors").

(4) Pre-predictors are combined and transformed (for each cluster or non-cluster, and each dropped year analysis separately) through principal component analysis (PCA; Jolliffe, 2002).

(5a) The top principal components (PCs) from the PCA with a total of 95% variance explained are used as predictors – the direct inputs into the MLR model, otherwise known as the principal component regression (PCR). For the *direct* case, PCR is used to directly predict the grid-level precipitation; for the *indirect* case, PCR is used to predict the intermediate cluster-level precipitation.

(5b) For the *indirect* case only, cluster-level predictions are regressed to the grid-level. Note that the downscaling of cluster-level predictions to grid-level predictions is also cross-validated to avoid overfitting."

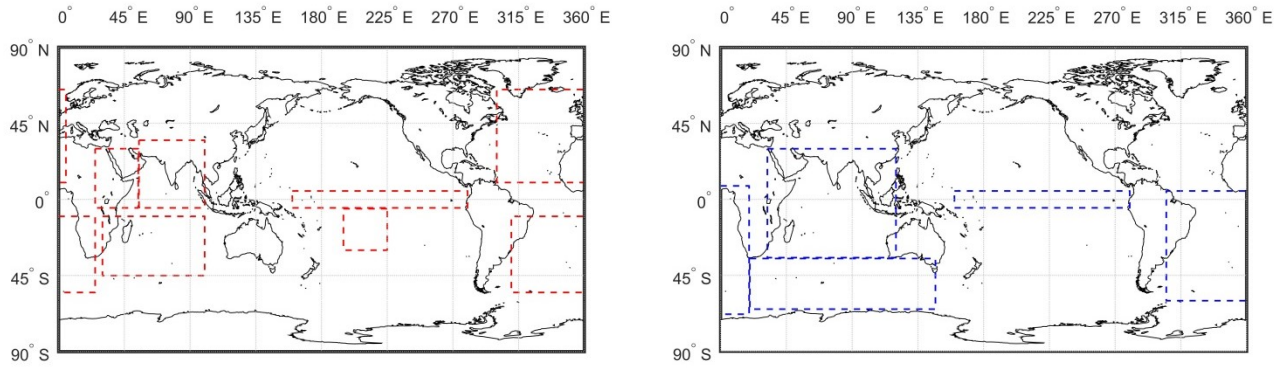


Figure 4: Justifiable climate regions globally for selecting predictors: (a) For SLP and GH at 500 mb with regions including EP, ES, LO, AH, SH, MH, and AM. For SAT, only LO is included. (b) For SST with regions including EP, NI, SI, and AT. *Note: EP - equatorial Pacific region, ES – Tahiti island for ENSO measurement, LO - local region, AH - Azores High, SH - St Helena High, MH - Mascarene High, AM - SW Asian Monsoon, NI - North Indian Ocean, SI - South Indian Ocean, AT - Equatorial/South Atlantic Ocean.*

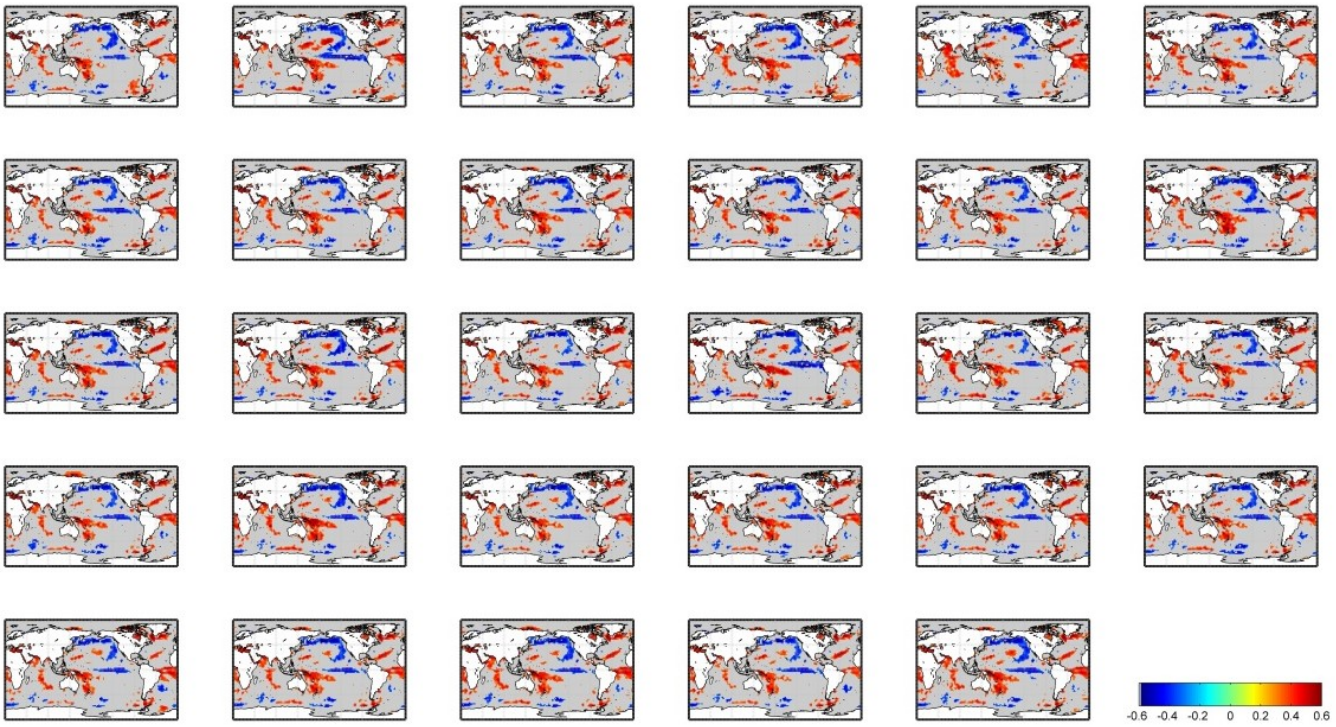


Figure 5: Correlation map between mean JJAS seasonal precipitation time series in Cluster 5 and global SST under cross-validation, with correlations lower than 99% significance level masked out (one-tail test).

The results are also updated accordingly. As the reviewer expected, the model skill decreases significantly. Under the same model of PCR, the correlations of cluster-level prediction and observation now range from -0.180 to 0.504 (compared to 0.683 - 0.838 originally), with Cluster 5 having the highest correlation while Cluster 6 showing the lowest. Similarly for RPSS, 5 out of 8 clusters show skillful prediction compared to climatology (Table 2). However, we do see improvement over non-cluster scenarios for some of the clusters.

Table 2: Correlation coefficients (Corr.) and RPSS for predictions (drop-one-year cross-validated) at cluster level compared to observations under C-I and NC-I scenario. (PCR model)

Cluster	C1	C2	C3	C4	C5	C6	C7	C8	Non-cluster
Corr.	0.163	-0.010	0.179	0.188	0.504	-0.180	0.351	-0.122	0.297
RPSS(%)	33.41	-21.66	43.01	12.46	27.40	-37.79	20.63	-55.96	13.25

We have also tried stepwise regression with forward selection and backward elimination algorithm (von Storch and Zwiers, 1999). The skill does increase a little for some cluster, but other clusters had a deterioration of prediction skills. One table showing the results from stepwise regression with forward

selection probability of 0.05 and backward elimination probability of 0.05 (can also be understood as significance levels) is presented below.

Table 2b: Correlation coefficients (Corr.) and RPSS for predictions (drop-one-year cross-validated) at cluster level compared to observations under C-I scenario. (Stepwise model)

Cluster	1	2	3	4	5	6	7	8
Corr.	0.154	-0.094	0.175	0.248	0.508	-0.197	0.305	-0.044
RPSS	-3.31	-44.17	40.60	-49.89	33.24	-77.10	25.00	-24.87

Comparing to PCR, stepwise regression overfits some specific predictors, such as the SST in the equatorial Pacific region or SLP for the St. Helena High, whereas PCR extracts the main signal first and then fits with a regression model. As a result, less noise is left in the predictors (PCs) using PCR than those in the stepwise regression. Hence, PCR is more reliable, regardless of the prescribed rule for selecting a certain number of predictors, while stepwise regress is extremely sensitive to different prescriptions of significance levels of selecting and eliminating probability. Therefore, we consider PCR a more reliable method in this case and keep the results from PCR for this work with added discussion on stepwise at the end. Table 2 above is included in the revised manuscript together with other updated tables, figures and result analysis, which are also included here:

(Page 10 Line 5) “Correlations between cluster-level model predictions and observations range from -0.18 to 0.50, with Cluster 5 having the highest correlation and Cluster 6 the lowest (Table 2). In approximately 1/5 of the 29 years, the observation falls outside the prediction envelope (Fig. 7), indicating model overfitting and an inability of the predictors to capture precipitation variability. For RPSS, 5 out of 8 clusters indicate superior prediction skill over climatology (Table 2). Improvement in terms of RPSS over the non-cluster scenario is evident for Cluster 1, 3, 5 and 7. Among all clusters, Cluster 5, in agriculturally rich central-northwestern Ethiopia (Fig. 2), performs the best, with correlation and RPSS values of 0.5 and 27%, respectively. Cluster 2, 4, 6 and 8, however, show deteriorated RPSS compared to non-cluster scenario, although those clusters are mainly regions outside Ethiopia and southern Ethiopia (Fig. 2) where water resources and agricultural activities are considerably less (Fig. 1).”

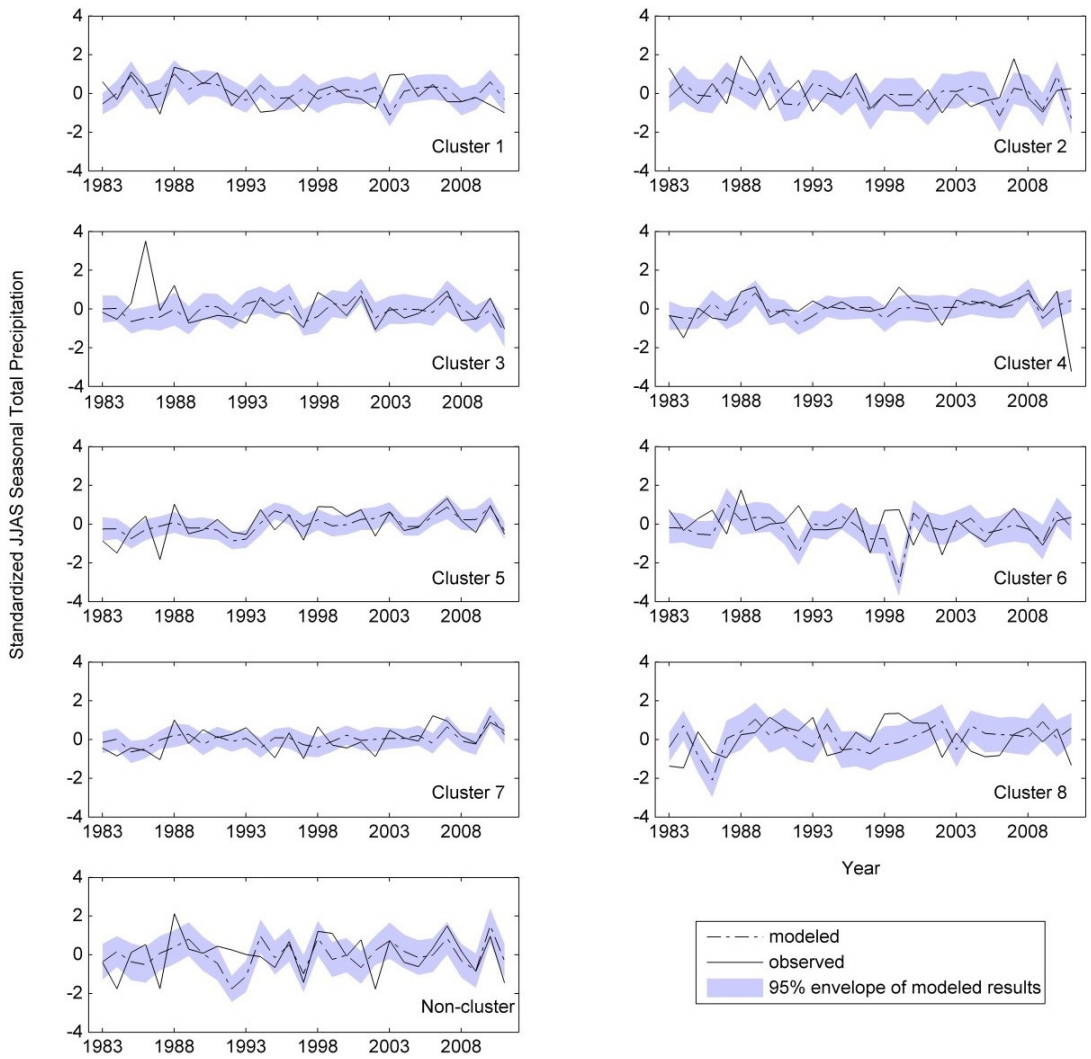


Figure 7: Cluster-level predictions and observations under C-I and NC-I scenario, with drop-one-year cross-validation. The 95% envelope shows the 95% confidence interval constructed using model errors.

(Page 11 Line 5) "At the grid scale, depending on the case (*direct* or *indirect)*, and for different clusters, correlations between predictions and observations can favor the *clustered* case or the *non-clustered* case (Fig. 8). In general, the *indirect* model provides a smoother pattern of correlations, with grid-cells showing a negative correlation in the *direct* case now improved to near or above zero (Fig. 8). For example, Cluster 5 under the *indirect* case illustrates a more consistent positive correlation within the cluster. Some parts of the region reach a correlation over 0.6, such as central-northwestern Ethiopia (Cluster 5), which is consistent with the region of high cluster-level prediction skill. The percentage of grid-cells with correlations passing the 95% significance test is the highest for the NC-D case (Table 3); however, some locations demonstrate the lowest skills among all four scenarios."

"

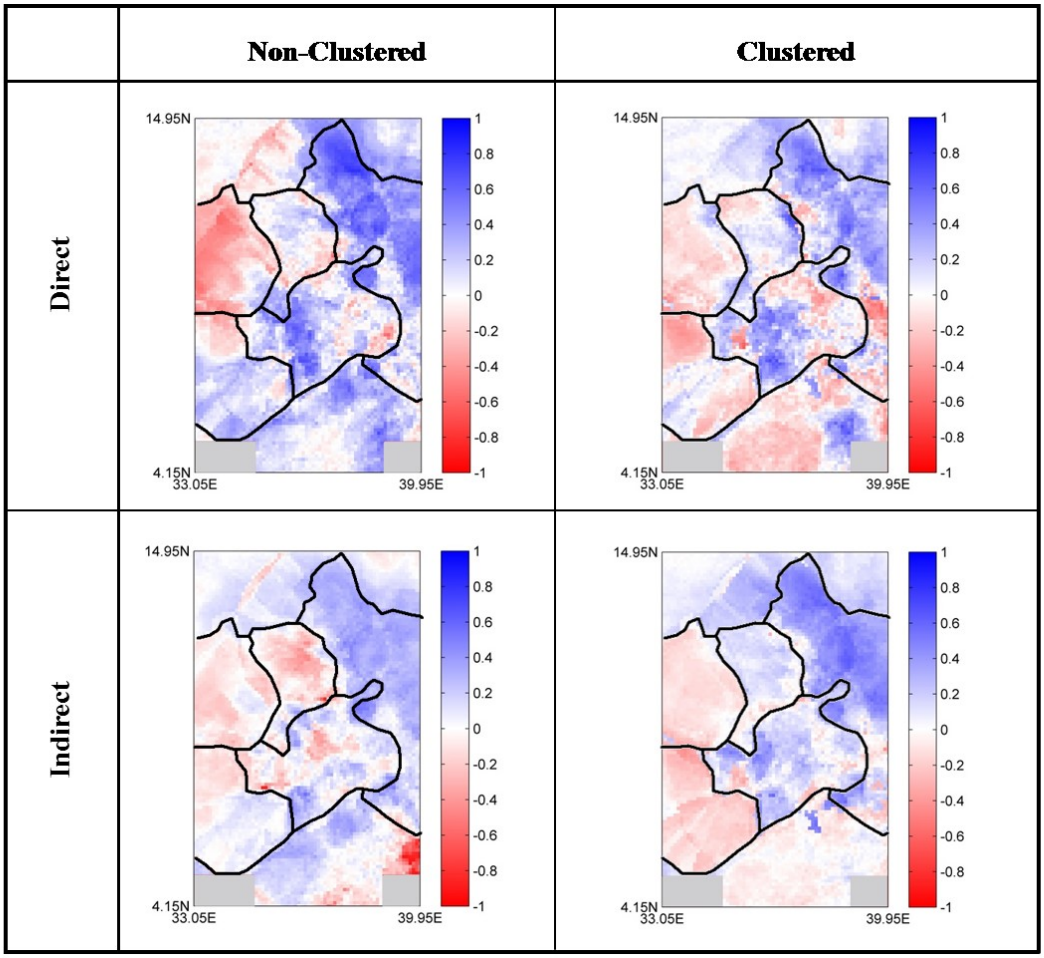


Figure 8: Pearson correlations between grid-level observations and predictions under four scenarios, with the clustering boundary delineated roughly in black.

(Page 12 Line 3) "Similar findings are evident by evaluating the RPSS except for Cluster 8; instead of improving with increased RPSS in the *indirect* case, the grid-scale predictions deteriorate given poor cluster-level prediction (for the C-I case). However, the percentage of grid-cells with positive RPSS values overall is still the highest for the C-I case (Table 3), indicating the *clustered indirect* case is superior in terms of the number of grid-cells with improved skill compared to using climatology, particularly for grid-cells associated with skillful intermediate cluster-level predictions. The predictions are most skillful for the same region of central-northwestern Ethiopia (Cluster 5; Fig. 9) with 87% of its grid-cells showing positive RPSS and a spatial average RPSS value of 15% under the C-I scenario (Table 4). "

Table 3: Grid-level Pearson correlation and RPSS statistics

Statistical Model	Grid-level correlations			Grid-level RPSS		
	mean	stdev	significant corr %	mean (%)	stdev (%)	positive RPSS %
NC-D	0.128	0.258	19.3%	-5.21	27.0	42.8%
NC-I	0.063	0.186	3.13%	-2.26	14.6	43.9%
C-D	0.055	0.230	10.6%	-14.0	31.0	33.9%
C-I	0.081	0.206	12.4%	-9.60	29.4	44.4%
Dynamical Model						
(1)	-0.105	0.209	0.51%	-31.4	25.4	5.70%
(2)	0.133	0.171	6.26%	-14.2	24.6	27.0%
(3)	0.086	0.130	2.08%	-14.9	25.2	26.2%
(4)	0.027	0.156	0.38%	-14.4	19.3	22.6%
(5)	0.067	0.170	1.64%	-9.66	17.0	28.4%
(6)	0.139	0.165	6.53%	-5.66	16.7	38.1%
(7)	0.102	0.130	1.67%	-8.64	17.6	31.7%
(8)	0.009	0.185	0.90%	-10.3	14.8	26.7%
(9)	0.244	0.149	23.1%	-2.33	21.8	46.0%
(10)	0.244	0.149	21.2%	-1.09	16.8	48.9%

Table 4: Grid-level Pearson correlation and RPSS statistics for grid-cells within Cluster 5

Statistical Model	Grid-level correlations			Grid-level RPSS		
	mean	stdev	significant corr %	mean (%)	stdev (%)	positive RPSS %
NC-D	0.378	0.211	60.7%	19.1	22.9	80.3%
NC-I	0.265	0.111	12.8%	8.33	14.8	70.3%
C-D	0.229	0.244	30.5%	6.91	24.1	62.3%
C-I	0.345	0.165	55.7%	14.7	13.3	87.1%
Dynamical Model						
(9)	0.353	0.110	46.8%	8.21	18.2	65.7%
(10)	0.248	0.130	18.4%	3.92	16.2	59.5%

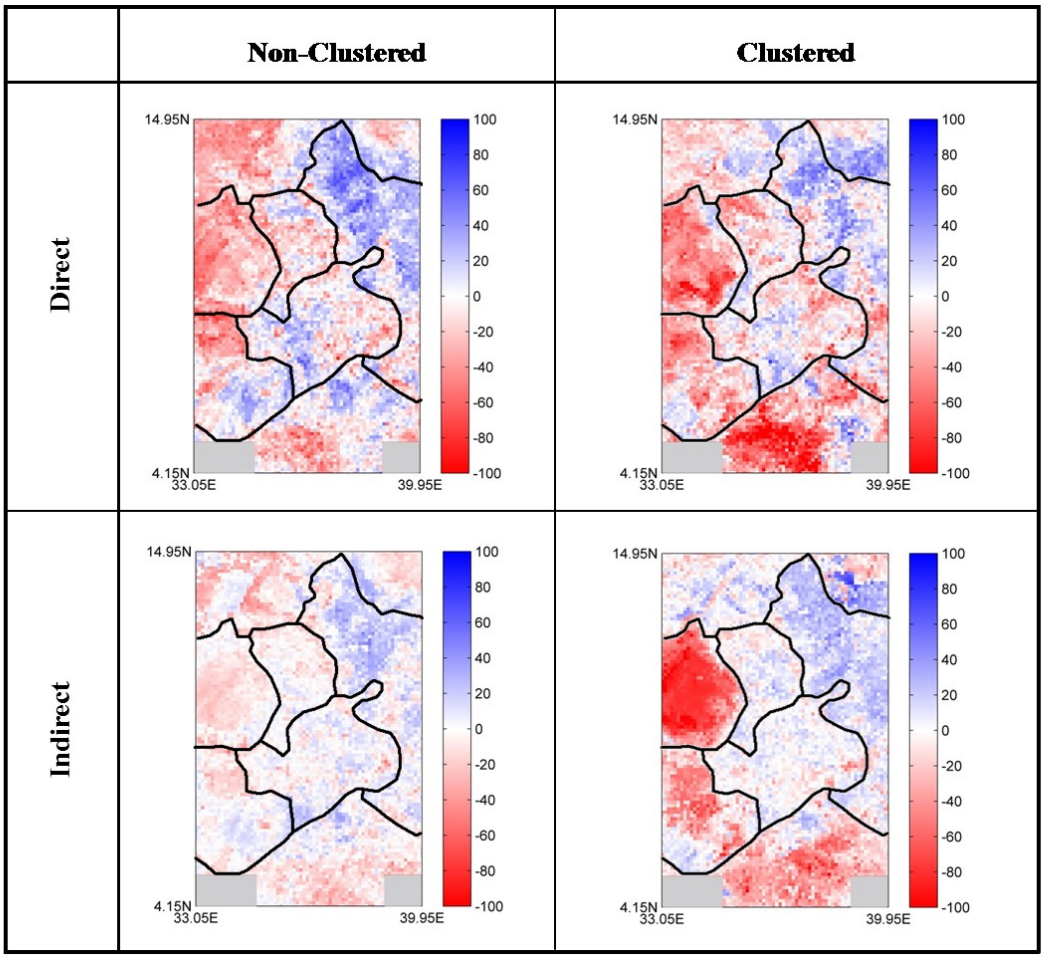


Figure 9: Grid-level RPSS (%) under four scenarios using climate variables as predictors, with the clustering boundary delineated roughly in black.

3) Evaluation of the Downscaling approach: As the predictor selection, the downscaling procedure is not included in the cross-validation. I recommend to conduct the crossevaluation for the entire modeling chain. That means, the predictor selection, cluster forecast and downscaling approach need to be calibrated based on (n-1) years, in order to forecast gridded precipitation for the remaining year.

In general, one should be aware, that there is a linear dependence of the cluster based and the gridded forecast. Thus, the downscaling approach will better reproduce the local climate, however the variability (drought and moist years) will be equal to the culstered result. The term “gridded” forecast is somehow misleading – I would prefer “downscaling of regional forecast”

The downscaling procedure is included in the cross validation. We apologize if this is not clear and added one sentence to the statistical model steps (Page 6 Line 29):

"(5b) For the *indirect* case only, cluster-level predictions are regressed to the grid-level. Note that the downscaling of cluster-level predictions to grid-level predictions is also cross-validated to avoid overfitting."

We also revised "gridded forecast" to "prediction at grid level" and "downscaling of cluster-level to grid-level prediction" for more appropriate descriptions.

Detailed remarks: 1) The abstract is very short and could certainly be more informative (e.g. by including some results)

We thank the reviewer for the comment and we have revised the abstract with highlights on results as the reviewer suggested. The added texts are also included here:

"… makes clear advances in prediction skill and resolution, as compared with previous studies. The statistical model improves versus the non-clustered case or dynamical models for a number of specific clusters in northwestern Ethiopia, with some cluster having regional average correlation and RPSS values of approximately 0.5 and 27%, respectively. The general skill of the two best performing dynamical models over the entire study region is superior to that of the statistical models, although dynamical models issue predictions at a lower resolution."

2) The discussion of state of the art forecasting models in the introduction is very short. Particularly during recent years, several studies investigated the skill of statistical models for regional scale precipitation forecasts (some of them are even based on clustering or PCA). I would recommend to better discuss the literature and the advantage of your approach in the introduction. See for example:

Hertig, E. and Jacobeit, J.: Predictability of Mediterranean climate variables from oceanic variability. Part II: Statistical models for monthly precipitation and temperature in the Mediterranean area, Clim. Dynam., 36, 825–843, doi:10.1007/s00382-010- 0821-3, 2010.

Suárez-Moreno, R. and Rodríguez-Fonseca, B.: S4CAST v2.0: sea surface temperature based statistical seasonal forecast model, Geosci. Model Dev., 8, 3639–3658, doi:10.5194/gmd-8-3639-2015, 2015.

Gerlitz, L., Vorogushyn, S., Apel, H., Gafurov, A., Unger-Shayesteh, K. & Merz, B.: A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south Asia, HESS 20, 4605–4623 , doi:10.5194/hess-20-4605-2016 , 2016.

We thank the reviewer for the comment. The recommended literatures are closely related to our methodology, although they are not based on the same study region. We have added them to the current literature review collection under Section 3.2 "statistical modeling approach" (Page 5 Line 6) with the following texts inserted:

"Many studies have investigated statistical models for seasonal climate prediction. These studies vary by pre-classification of predictor or predictand regime, predictor selection process, and statistical methods. For example, Hertig and Jacobeit (2011) investigate sea surface temperature (SST) regimes as potential predictors for subsequent precipitation and temperature in the Mediterranean region. Through techniques including multiple applications of PCA, 17 stationary SST regimes were identified. Gerlitz et al. (2016) apply a k-means cluster analysis to grid-cells identified with significant correlations in the predictor field in order to facilitate predictor selection. Suárez-Moreno and Rodríguez-Fonseca (2015) investigate stationarity based on a long time series using a 21-year moving correlation window. The statistical prediction models are then applied to each stationary period respectively and the entire period for comparison. Despite diverse methods in seasonal prediction, multiple linear regression (MLR) is favored by many as a statistical modeling approach given its well-developed theory, simple model structure, efficient processing, and often skillful outcomes (e.g. Omondi et al., 2013, Camberlin and Philippon, 2002, Diro et al., 2008). As mentioned, only a few studies have focused on seasonal precipitation prediction in Ethiopia (Gissila et al., 2004, Block and Rajagopalan, 2007, Korecha and Barnston, 2007, Diro et al., 2008, Diro et al., 2011, Segele et al., 2015), and almost all of them include the applications of MLR. This study also applies MLR to predict seasonal precipitation, yet differentiates from other studies by applying predictions to pre-defined homogeneous regions and further translating to local-level predictions."

3) The predictor selection is based on correlation maps, and regions with potential forecast skill are identified (see Tab.1). Please map the regions and show the correlation maps for some clusters.

We have added one figure showing regions for selecting predictors and one correlation map for a cluster as an example. Please refer to Figure 4 and 5.

4) P7,l10: PCAs are cross-validated. This is somehow unclear to me. PCA is usually used for dimension reduction. Is the cross-validation done for the loadings of the pca in order to investigate how these change based on different input data?

PCR is cross-validated as indicated on Page 7 Line 17. To be specific, we performed PCA on the time series with the target-year data dropped, and then use the resultant PCs as predictors for the regression model. After the regression coefficient estimates are obtained, the principal components for the dropped year are *reconstructed*, and then multiplied with the coefficient estimates respectively in order to obtain the final predicted value for the dropped year. Please see more details in Wilks (2011).

5) Dynamical Models: The section on dynamical models is poorly integrated. Please give some more information on the models in general. If (as expected) the skill of the statistical model drops as a consequence of the cross-evaluation, a more detailed comparison of skills might be interesting.

We apologize for our poorly integrated section on dynamical model and have improved it with an additional introduction of the NMME models (Page 8 Line 22):

"The North American Multi-Model Ensemble (NMME; Kirtman et al., 2014) is an experimental multi-model system consisting of coupled dynamical models from various modeling centers in North America. To our knowledge, it is also the most extensive multi-model seasonal prediction archive. The NMME provides gridded climate predictions that cover regions globally and with different lead times. The hindcasts of monthly mean precipitations are easily accessible through the International Research Institute for Climate and Society (IRI) website (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/), and can be easily aggregated to seasonal totals for comparison with the statistical model results in this study."

We agree that after the model is revised, the skills of statistical model and dynamical model are comparative now. We have updated with Table 3 and 4 (see above), and provided additional result analysis with comparison between them. Texts are also copied here (Page 13 Line 7):

"The RPSS values based on the prediction ensembles of each dynamical model improve significantly after bias correction. The median RPSS values over all the grid-cells are now close to zero (Fig. 10) with two models, NASA-GMAO and NCEP-CFSv2, showing the highest RPSS value (-2.3% and -1.1%, respectively; Table 3). These two dynamical models also exhibit generally higher grid-level correlations over the study region (averaging 0.24 for both models; Table 3 and Fig. 11), as compared with other NMME models. The two best performing dynamical models after bias correction show advantage over statistical models, as assessed by correlation and RPSS metrics; however, all other dynamical models are inferior to the statistical models under NC-D and C-I scenarios, particularly given the percent of grid-cells with significant correlation and positive RPSS metrics (Table 3).

Within a certain cluster, statistical models may perform better than all dynamical models. For example, for Cluster 5, all statistical models show higher average RPSS values than that of all dynamical models (Table 4). The percentage of grid-cells with significant correlation reaches 61% for the statistical model under NC-D scenario, compared to the highest value of 47% among all the dynamical models. Similarly, the percentage with positive RPSS achieves 87% under C-I scenario as opposed to 66% for dynamical models. Note that the dynamical models also produce predictions in a lower spatial resolution (1°×1°) than the statistical models (0.1°×0.1°)."

6) P9,l15: Please give some more information on the performance measures (BIC, AIC, GCV).

We have eliminated the performance measures BIC AIC GCV, as the revised predictor selection rule produces a dynamic number of predictors depending on which years is dropped and for different clusters (for the *indirect* case) or grid-cells (for the *direct* case).

7) P10: How exactly is the envelope (uncertainty interval) calculated? Is this based on the assumption that cross-validated residuals of the regression are normal distributed?

Yes. Q-Q plots are evaluated to verify normally distributed residuals (results not included) as indicated on Page 8 Line 3. To make it clear, we have added the following text to the same line:

"A 95% confidence interval of the cross-validated predictions is also constructed conditioned on model errors. Q-Q plots are evaluated to verify normally distributed residuals (results not included)."

References

BLOCK, P. J. & RAJAGOPALAN, B. 2007. Interannual Variability and Ensemble Forecast of Upper Blue Nile Basin Kiremt Season Precipitation. *J. Hydrometeor,* 8**,** 327-343.

CAMBERLIN, P. & PHILIPPON, N. 2002. The East African March–May Rainy Season: Associated Atmospheric Dynamics and Predictability over the 1968–97 Period. *Journal of Climate,* 15**,** 1002-1019.

DIRO, G. T., BLACK, E. & GRIMES, D. I. F. 2008. Seasonal forecasting of Ethiopian spring rains. *Meteorological Applications,* 15**,** 73-83.

DIRO, G. T., GRIMES, D. I. F. & BLACK, E. 2011. Teleconnections between Ethiopian summer rainfall and sea surface temperature: part II. Seasonal forecasting. *Climate Dynamics,* 37**,** 121-131.

GERLITZ, L., VOROGUSHYN, S., APEL, H., GAFUROV, A., UNGER-SHAYESTEH, K. & MERZ, B. 2016. A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south Asia. *Hydrol. Earth Syst. Sci.,* 20**,** 4605-4623.

GISSILA, T., BLACK, E., GRIMES, D. I. F. & SLINGO, J. M. 2004. Seasonal forecasting of the Ethiopian summer rains. *International Journal of Climatology,* 24**,** 1345-1358.

HERTIG, E. & JACOBEIT, J. 2011. Predictability of Mediterranean climate variables from oceanic variability. Part II: Statistical models for monthly precipitation and temperature in the Mediterranean area. *Climate Dynamics,* 36**,** 825-843.

JOLLIFFE, I. 2002. *Principal component analysis*, Wiley Online Library.

KIRTMAN, B. P., MIN, D., INFANTI, J. M., KINTER, J. L., PAOLINO, D. A., ZHANG, Q., VAN DEN DOOL, H., SAHA, S., MENDEZ, M. P., BECKER, E., PENG, P., TRIPP, P., HUANG, J., DEWITT, D. G., TIPPETT, M. K., BARNSTON, A. G., LI, S., ROSATI, A., SCHUBERT, S. D., RIENECKER, M., SUAREZ, M., LI, Z. E., MARSHAK, J., LIM, Y.-K., TRIBBIA, J., PEGION, K., MERRYFIELD, W. J., DENIS, B. & WOOD, E. F. 2014. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society,* 95**,** 585-601.

KORECHA, D. & BARNSTON, A. G. 2007. Predictability of June–September Rainfall in Ethiopia. *Monthly Weather Review,* 135**,** 628-650.

OMONDI, P., OGALLO, L. A., ANYAH, R., MUTHAMA, J. M. & ININDA, J. 2013. Linkages between global sea surface temperatures and decadal rainfall variability over Eastern Africa region. *International Journal of Climatology,* 33**,** 2082-2104.

SEGELE, Z. T., RICHMAN, M. B., LESLIE, L. M. & LAMB, P. J. 2015. Seasonal-to-Interannual Variability of Ethiopia/Horn of Africa Monsoon. Part II: Statistical Multi-Model Ensemble Rainfall Predictions. *Journal of Climate***,** 150129124820009.

SUÁREZ-MORENO, R. & RODRÍGUEZ-FONSECA, B. 2015. S⁴CAST v2.0: sea surface temperature based statistical seasonal forecast model. *Geosci. Model Dev.,* 8**,** 3639-3658.

VON STORCH, H. & ZWIERS, F. W. 1999. *Statistical analysis in climate research,* Cambridge, Cambridge University Press.

WILKS, D. S. 2011. *Statistical methods in the atmospheric sciences*, Academic press.

ZHANG, Y., MOGES, S. & BLOCK, P. 2016. Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial-Temporal Variability: Application to Western Ethiopia. *Journal of Climate,* 29**,** 3697-3717.

The authors proposed local level seasonal predictions of precipitation in the Western Ethiopia using the statistical approach at eight homogeneous regions clustered using k-means clustering technique. Large scale climate variables are used as potential predictors in developing the statistical model and unique set of predictors were assigned for each region. Results are compared with the dynamical prediction at regional scale and reported as the statistical approach is superior. This study is timely and helpful for the study region where rainfall is highly variable in space. The manuscript is well written and easy to understand and follow. I think this study will be a nice addition however a few moderate issues need to be resolved first. Please see my comments below.

1. The statistical approach is fully data driven approach that depends on the quality and length of the data. So how efficient is this technique in the area where there is sparse and poor quality data in the case of developing country like Ethiopia? The authors did not show how good the gridded rainfall data is through either validating with gauged data for selected weather stations or previous literatures that support the quality of this data.

We thank the reviewer for the comment. The dataset we use has been shown to reproduce station data over areas with both densely and sparsely distributed station networks. The original data is at a 10-day time interval, and we aggregate it to JJAS seasonal total precipitation, which may help offset some random errors and better represent the observations. Regarding the technique itself, as the cluster analysis and statistical model are purely data-driven, the data length and quality is essential to produce skillful results. With time, as the data length and availability improve, results are expected to become more skillful.

To justify the data quality, the following texts are added to the manuscript (Page 4 Line 7):

"…1983–2011 (29 years). This product has been verified against station data and has been deemed representative of observed precipitation in western Ethiopia (Dinku et al.,

2014). ”

and in the discussion (Page 15 Line 35):

“As observational datasets continue to grow, data-driven cluster analysis and statistical modeling approaches may be expected to improve.”

2. The author argue that this study gives perdition of seasonal precipitation at high resolution in the region. However, the classification of the homogeneous regions by NMA, Ethiopia, for the study region is almost equivalent (Koricha et al., 2007, pg 7685). I do not see any benefits of this study interms of the spatial resolutions at regional scale for Western Ethiopia.

We thank the reviewer for the comment. We agree that classifications of homogeneous regions are similar but still different. The prediction at regional scale in this study helps to verify our classification method, which is distinct from NMA’s classification (see details in Zhang et al. (2016)). The regional-scale prediction also serves as an intermediate predictand for the grid-scale prediction under the *indirect* case. However, like the reviewer indicated, the highlight of the work should be the high-resolution prediction at grid scale.

3. No effort have been made on finding out the time lag between the predictor variables and the seasonal rainfall in the study area. For example which month of sea surface temperature really affects the seasonal rainfall in the study area.

We confirm that only one lead time is investigated in this work, and mention in the discussion that longer prediction lead times and evaluation of other relevant characteristics (e.g. intra-seasonal dry spells, seasonal onset or cessation, etc.) warrant future attention. However, for the purposes of this study, we consider only one lead time and instead focus on whether cluster analysis serves as a useful precursor to seasonal precipitation prediction at the local scale. Additional lead times can also be

readily applied to the current framework and are likely to be informative as to which months in the season-ahead are most related to JJAS seasonal precipitation.

4. I don't see any comparison of the result with the current operational NMA forecast in the manuscript, however, it is reported in the conclusion section asif the result is compared at the regional scale with NMA operational forecast.

We have compared our results and NMA's results qualitatively only, and briefly mention in the discussion that at the regional scale our results are more skillful than NMA's prediction, based on Korecha and Sorteberg (2013). We have revised the sentence to avoid confusion and added more detailed comparison between the two products (Page 15 Line 3):

"At the regional scale, the approach shows promise for northwestern Ethiopia (Cluster 1, 3, 5, and 7), particularly compared to current NMA operational forecasts, which are only moderately more skillful than climatology (Korecha and Sorteberg, 2013). The regional average RPSS in this study under the clustered case ranges from 21% to 43% for northwestern Ethiopia, as opposed to values under 6% for NMA operational forecast (Korecha and Sorteberg, 2013)."

5. The abstract is too short and mainly focused on the merit of conducting this study. It would be good if it is supported with some finding. There is no introduction given about the study area interms of the rainfall patter, topography etc.

We thank the reviewer for the comment. More content is added to the original manuscript with one additional figure (also included here):

In introduction (Page 3 Line 2):

"Precipitation in western Ethiopia peaks in the summer with approximately 70% of annual total precipitation falling during the main raining season - also known as the Kiremt season spanning from June to September (JJAS). On average, the seasonal total precipitation in the study region is approximately 760 mm; however in the northwest,

precipitation can exceed 1200 mm (Fig. 1a). Along with the high spatial variability in this mountainous region, the temporal variability is also significant with spatial-average seasonal total precipitation ranging from 650 mm in dry years up to 900 mm in wet years (Fig. 1b). These highly variable spatial and temporal precipitation patterns have made skillful seasonal predictions challenging, particularly at local scales (e.g. Gissila et al., 2004, Block and Rajagopalan, 2007)."

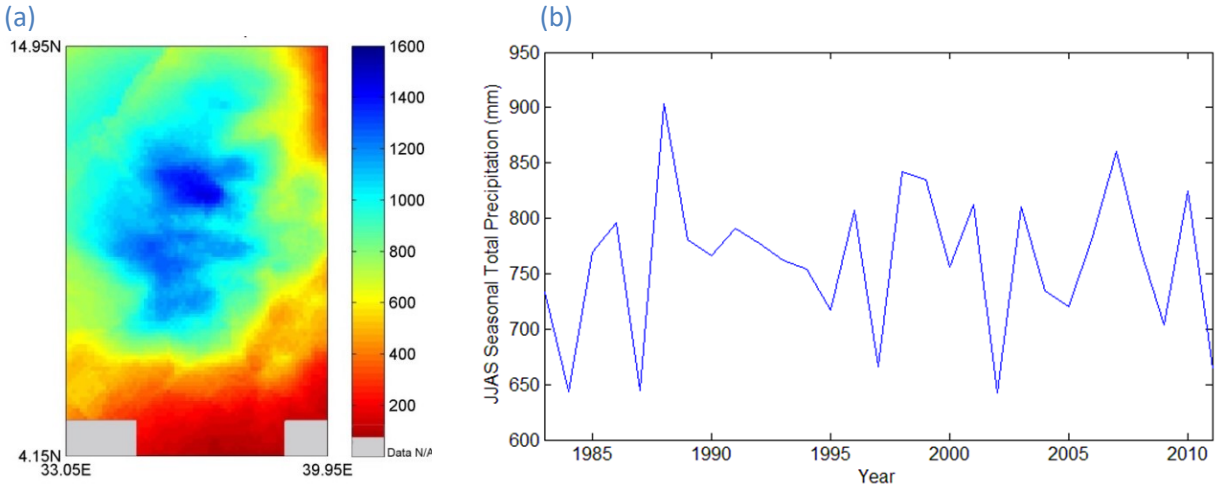


Figure 1: Spatial and temporal variability of June-September seasonal total precipitation in western Ethiopia: (a) spatial pattern of temporal-average, and (b) spatial-average time series.

In abstract:

"… makes clear advances in prediction skill and resolution, as compared with previous studies. The statistical model improves versus the non-clustered case or dynamical models for a number of specific clusters in northwestern Ethiopia, with some cluster having regional average correlation and RPSS values of approximately 0.5 and 27%, respectively. The general skill of the two best performing dynamical models over the entire study region is superior to that of the statistical models, although dynamical models issue predictions at a lower resolution."

Reference from reviewer's comments:

Korecha, D., and Sorteberg, A.: Validation of operational seasonal rainfall forecast in

Ethiopia, Water Resources Research, 49, 20 7681-7697, 10.1002/2013wr013760, 2013.

Reference from the author's responses:

BLOCK, P. J. & RAJAGOPALAN, B. 2007. Interannual Variability and Ensemble Forecast of Upper Blue Nile Basin Kiremt Season Precipitation. *J. Hydrometeor,* 8**,** 327-343.

DINKU, T., HAILEMARIAM, K., MAIDMENT, R., TARNAVSKY, E. & CONNOR, S. 2014. Combined use of satellite estimates and rain gauge observations to generate high-quality historical rainfall time series over Ethiopia. *International Journal of Climatology,* 34**,** 2489-2504.

GISSILA, T., BLACK, E., GRIMES, D. I. F. & SLINGO, J. M. 2004. Seasonal forecasting of the Ethiopian summer rains. *International Journal of Climatology,* 24**,** 1345-1358.

KORECHA, D. & SORTEBERG, A. 2013. Validation of operational seasonal rainfall forecast in Ethiopia. *Water Resources Research,* 49**,** 7681-7697.

ZHANG, Y., MOGES, S. & BLOCK, P. 2016. Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial-Temporal Variability: Application to Western Ethiopia. *Journal of Climate,* 29**,** 3697-3717.

Does objective cluster analysis serve as a useful precursor to seasonal precipitation prediction at local scale? Application to western Ethiopia

Ying Zhang[1], Semu Moges[2], Paul Block[1]

[1]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, 53706, USA
[2]School of Civil and Environmental Engineering, Addis Ababa University, Addis Ababa, 1000, Ethiopia

Correspondence to: Paul Block (paul.block@wisc.edu)

Abstract. Prediction of seasonal precipitation can provide actionable information to guide management of various sectoral activities. For instance, it is often translated into hydrological forecasts for better water resources management. However, many studies assume homogeneity in precipitation across an entire study region, which may prove ineffective for operational and local-level decisions, particularly for locations with high spatial variability. This study proposes advancing local-level seasonal precipitation predictions by first conditioning on regional-level predictions, as defined through objective cluster analysis, for western Ethiopia. To our knowledge, this is the first study predicting seasonal precipitation at high resolution in this region, where lives and livelihoods are vulnerable to precipitation variability given the high reliance on rain-fed agriculture and limited water resources infrastructure. The combination of objective cluster analysis, spatially high-resolution prediction of seasonal precipitation, and a modeling structure spanning statistical and dynamical approaches makes clear advances in prediction skill and resolution, as compared with previous studies. The statistical model improves versus the non-clustered case or dynamical models for a number of specific clusters in northwestern Ethiopia, with some cluster having regional average correlation and RPSS values of approximately 0.5 and 27%, respectively. The general skill of the two best performing dynamical models over the entire study region is superior to that of the statistical models, although dynamical models issue predictions at a lower resolution.

1 Primer on prediction models and cluster analysis

Seasonal precipitation prediction can provide potentially actionable information to guide management of various sectoral activities. For instance, precipitation prediction is often translated into a hydrological forecast, which can be used to optimize reservoir operations, provide early flood or drought warning, inform waterway navigation, etc. As a primary input to soil moisture, precipitation prediction is also essential to agricultural management – farmers can take advantage of anticipated preferable climatic conditions or avoid unnecessary costs under expected undesirable conditions. Two types of models are commonly used for seasonal precipitation prediction: statistical and dynamical. Dynamical models, such as general circulation models (GCMs), include complex physical climate processes, while statistical models are purely data-driven, relating observations and hydroclimate variables directly.

While both modeling approaches have produced skillful seasonal predictions for a variety of applications (e.g. Barrett, 1993, Hammer et al., 2000, Shukla et al., 2016), each has noteworthy drawbacks. Dynamical models often require a significant amount of time to build and parameterize, whereas statistical models require considerably fewer resources (e.g. Mutai et al., 1998, Gissila et al., 2004, Block and Rajagopalan, 2007, Diro et al., 2008, Diro et al., 2011b, Block and Goddard, 2012). Dynamical models also suffer from their high sensitivity to initial uncertain conditions, particularly given a long lead time. Consequently, a number of simulations are typically produced, each with unique initial conditions, to provide a range of possible outcomes (e.g. Roeckner et al., 1996, Anderson et al., 2007). Furthermore, the outputs from dynamical models often require additional bias correction, typically using statistical methods, to better match observations (e.g. Ines and Hansen, 2006, Block et al., 2009, Teutschbein and Seibert, 2012). Statistical models, on the other hand, are highly dependent on substantial high-quality historical data to capture hydroclimatic patterns and signals, particularly extreme conditions, which is often not available. Additionally, statistical models are often linear by construction, and may not well capture non-linear complex interactions and feedbacks. The physical nature of dynamical models, however, allows for prediction under non-stationary conditions, and also when insufficient historical data is available, whereas statistical models, by construction, typically rely on stationary relationships (Schepen et al., 2012).

~~Given these features of both model types for seasonal prediction, many studies have explored the combination of statistical and dynamical model outputs (e.g. Coelho et al., 2004, Block and Goddard, 2012, Schepen et al., 2012). In general, the combined predictions are typically superior to individual models, however this is not always the case, and is dependent on location, predicted seasons, lead time, comparable model skill, etc. (e.g. Metzger et al., 2004).~~

The spatial extent selected for statistical seasonal prediction is critical. It is not uncommon to simply assume homogeneity in precipitation across an entire study region; however, this limits addressing potential spatial variability. While this may be suitable for very broad regional planning, it is often ineffectual for operational and local-level decisions, particularly for locations with high spatial variability. This prompts the need for delineation of sub-regional scale homogeneous regions, often defined through cluster analysis. Defining these homogeneous regions, however, is a non-trivial process. There are a variety of methods to delineate homogeneous regions, including comparing annual cycles (e.g. unimodal and bimodal distributions in precipitation) between stations (or grid-cells), comparing station correlations with regional averages, applying empirical orthogonal functions (EOF), various clustering techniques, and other methods of increasing complexity (e.g. Parthasarathy et al., 1993, Mason, 1998, Landman and Mason, 1999, Gissila et al., 2004, Diro et al., 2008, Diro et al., 2011b, Singh et al., 2012). In addition, delineation of the sub-region size is also important to consider. Smaller sized homogeneous sub-regions do not necessarily lead to improved predictions, as the noise at overly small scales can dominate any real signals representing spatial coherency of precipitation. For additional

discussion regarding defining homogeneous sub-regions and cluster analysis, the reader is referred to Zhang et al. (2016) and Badr et al. (2015).

2 Application to western Ethiopia and objectives of the study

Precipitation in western Ethiopia peaks in the summer with approximately 70% of annual total precipitation falling during the main raining season - also known as the *Kiremt* season spanning from June to September (JJAS). On average, the seasonal total precipitation in the study region is approximately 760 mm; however in the northwest, precipitation can exceed 1200 mm (Fig. 1a). Along with the high spatial variability in this mountainous region, the temporal variability is also significant with spatial-average seasonal total precipitation ranging from 650 mm in dry years up to 900 mm in wet years (Fig. 1b). These highly variable spatial and temporal precipitation patterns have made skillful seasonal predictions challenging, particularly at local scales (e.g. Gissila et al., 2004, Block and Rajagopalan, 2007).

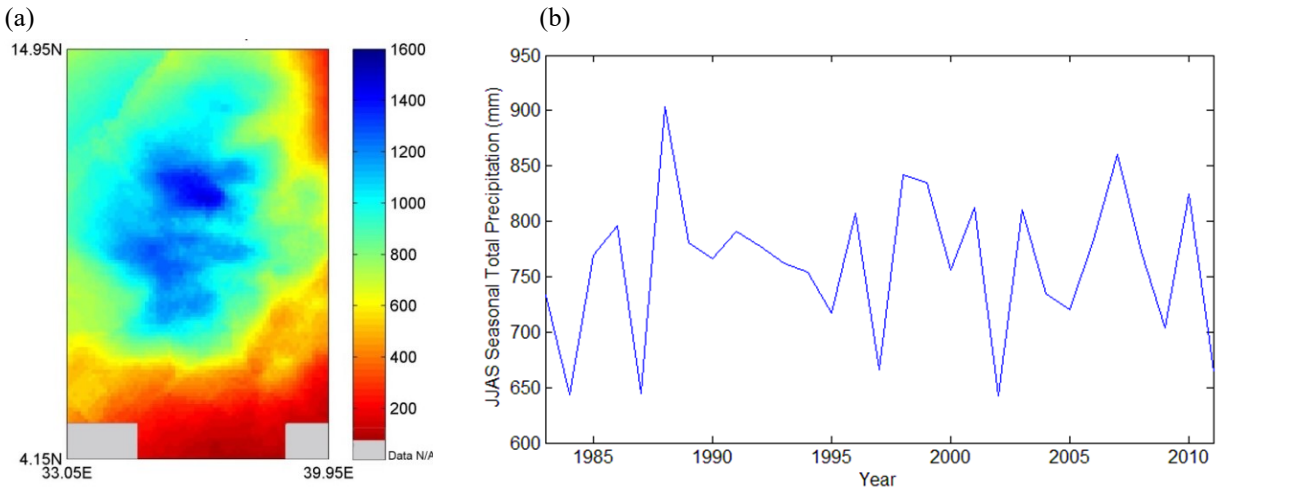


Figure 1: Spatial and temporal variability of June-September seasonal total precipitation in western Ethiopia: (a) spatial pattern of temporal-average, and (b) spatial-average time series.

Ethiopia is vulnerable to fluctuations in precipitation given its reliance on rain-fed agriculture and limited water resources infrastructure. The majority of agriculture and infrastructure are in western Ethiopia, where water resources are relatively rich compared to other parts of the country (Awulachew et al., 2007). ~~However, precipitation is highly varied temporally and spatially in the *Kiremt* season – the major rainy season spanning June through September (JJAS) – making skillful seasonal predictions challenging, particularly at local scales (e.g. Gissila et al., 2004, Block and Rajagopalan, 2007).~~ Operational precipitation predictions in Ethiopia have been issued by its National Meteorological Agency (NMA) since 1987 using an analog methodology (i.e. locating a similar climate scenario in the past – an analog – to predict future conditions), however this approach has produced only marginally skillful outcomes (Korecha and Sorteberg, 2013). For NMA's prediction, the country is divided into eight homogeneous regions, for which NMA produces independent predictions. Similarly, others have also addressed seasonal prediction in Ethiopia contingent on both temporal and spatial precipitation patterns. Gissila et al. (2004) divide Ethiopia into four regions conditioned on the seasonal cycle and interannual variability coherence prior to prediction, while Diro et al. (2009) apply a similar approach but with dynamic cluster boundaries, allowing for different delineations for each rainy season. Segele et al. (2015) consider statistical precipitation predictions across Ethiopia as a whole, as well as for northeastern Ethiopia and at two Ethiopian

cities. Block and Rajagopalan (2007) predict the average summertime (JJAS) precipitation over the upper Blue Nile basin – a region they claim is homogenous at inter-annual time scales. Korecha and Barnston (2007) select an all-Ethiopia average precipitation index to characterize predictability broadly, with minimal attention to operational-level predictions. All of these studies focus on predicting regional average precipitation based on subjective clustering methods applying a limited number of stations or coarsely gridded data; no local predictions at a finer spatial scale are explored.

This study moves forward by exploring local-level seasonal precipitation prediction through the use of regional-level predictions, based on previous cluster analyses over western Ethiopia (Zhang et al., 2016). The advantages of defining homogeneous regions for seasonal prediction at operational (small) scales will be demonstrated by comparing approaches with and without undertaking a cluster analysis *a priori*. The combination of objective cluster analysis, spatially high-resolution prediction of seasonal precipitation, and a modeling structure spanning statistical and dynamical approaches makes clear advances compared to previous studies.

3 Modeling high-resolution seasonal prediction

To evaluate high-resolution seasonal precipitation prediction comparing with versus without cluster analysis *a priori*, statistical models are developed and further compared with bias-corrected dynamical model predictions. Four scenarios are evaluated based on two criteria – (1) *clustered* vs. *non-clustered* and (2) *direct* vs. *indirect*. In the *clustered* case, predictions are produced for each homogeneous region (cluster) given a unique set of predictors. In the *non-clustered* case, the entire study region is considered as one cluster and thus only one set of predictors is utilized for predictions. For the *direct* case, precipitation is predicted directly at the local level (grid scale); for the *indirect* case, the average precipitation within each homogeneous region is predicted first (as an intermediary), and then regressed to local-level (grid scale) predictions. Combinations of the two criteria form four scenarios – *clustered direct* (C-D), *non-clustered direct* (NC-D), *clustered indirect* (C-I), and *non-clustered indirect* (NC-I) predictions.

3.1 Cluster analysis

Using a k-means clustering technique, western Ethiopia – the major agricultural region of the country – is divided into eight homogeneous regions (Fig. 2~~1~~), conditioned on the interannual variability of total precipitation in JJAS, the same variable that is to be predicted. Precipitation is based on a 0.1˚×0.1˚ gridded precipitation dataset from NMA (Dinku et al., 2014), consisting of 7320 grid-cells across 1983–2011 (29 years). This product has been verified against station data and has been deemed representative of observed precipitation in western Ethiopia (Dinku et al., 2014). Given the high-resolution gridded dataset, k-means clustering is performed for a range of predefined numbers of clusters; the optimal number of clusters is identified by comparing the within-cluster sum of square errors (WSS). During the clustering process, each grid-cell is assigned and reassigned to clusters until the WSS is minimized. This does not require any subjective delineation or manual delineation of boundaries between clustered stations or grid-cells; instead, an automated and objective delineation is performed. The mean time series of each cluster illustrates high intra-correlation within the cluster and low inter-correlation between any two clusters, indicating strong coherency of the clustering results. For a detailed analysis including a complete correlation table and unique patterns for each cluster-level time series associated with large climate variables, readers are referred to Zhang et al. (2016).

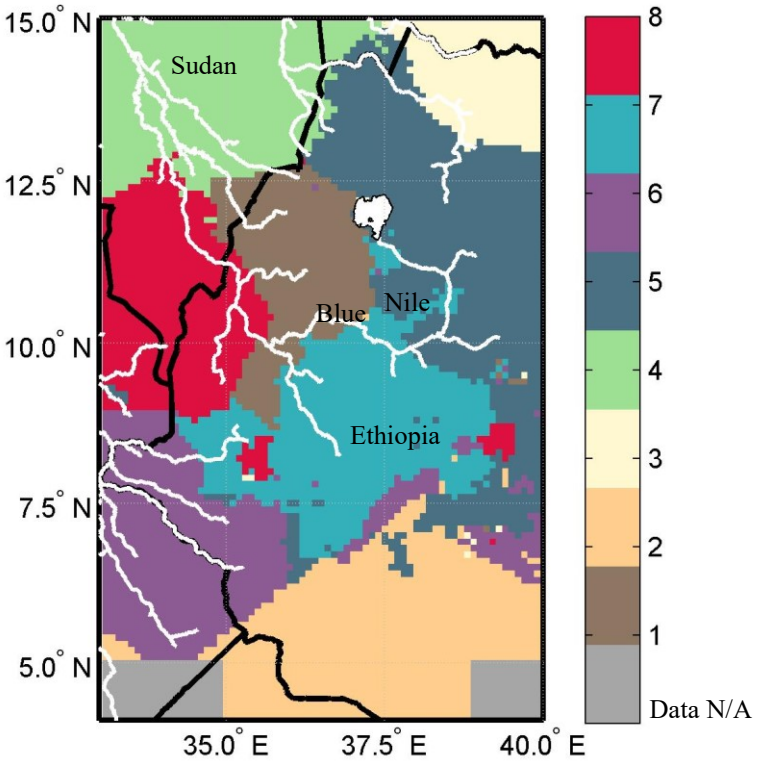


Figure ~~1~~2: Regionalization map of 8 homogeneous regions marked by different colors, with country boundary and river profile. After Zhang et al. (2016)

3.2 Statistical modeling approach

Many studies have investigated statistical models for seasonal climate prediction. These studies vary by pre-classification of predictor or predictand regime, predictor selection process, and statistical methods. For example, Hertig and Jacobeit (2011) investigate sea surface temperature (SST) regimes as potential predictors for subsequent precipitation and temperature in the Mediterranean region. Through techniques including multiple applications of PCA, 17 stationary SST regimes were identified. Gerlitz et al. (2016) apply a k-means cluster analysis to grid-cells identified with significant correlations in the predictor field in order to facilitate predictor selection. Suárez-Moreno and Rodríguez-Fonseca (2015) investigate stationarity based on a long time series using a 21-year moving correlation window. The statistical prediction models are then applied to each stationary period respectively and the entire period for comparison. Despite diverse methods in seasonal prediction, multiple linear regression (MLR) ~~Multiple linear regression (MLR)~~ is favored by many as a statistical modeling approach given its well-developed theory, simple model structure, efficient processing, and often skillful outcomes (e.g. Omondi et al., 2013, Camberlin and Philippon, 2002, Diro et al., 2008). As mentioned, only a few studies have focused on seasonal precipitation prediction in Ethiopia (Gissila et al., 2004, Block and Rajagopalan, 2007, Korecha and Barnston, 2007, Diro et al., 2008, Diro et al., 2011b, Segele et al., 2015), and almost all of them include the applications of MLR. This study also applies MLR to predict seasonal precipitation, yet differentiates from other studies by applying predictions to pre-defined homogeneous regions and further translating to local-level predictions.

Large-scale climate variables are often evaluated as potential predictors in statistical seasonal precipitation prediction models, commonly including sea surface temperatures (SST) in the equatorial Pacific Ocean representing the well-known of the El Nino-Southern Oscillation (ENSO) (Stone et al., 1996). For Ethiopia, the ENSO phenomenon is considered a significant indicator of

precipitation variability, particularly in the main JJAS rainy season (e.g. NMSA, 1996, Camberlin, 1997, Bekele, 1997, Segele and Lamb, 2005, Diro et al., 2011a, Elagib and Elhag, 2011). In addition to ENSO, the effect of Indian Ocean SST and regional atmospheric pressure systems such as the St. Helena, Azores, and Mascarene Highs also have notable influence on Ethiopia's precipitation variability (e.g. Kassahun, 1987, Tadesse, 1994, NMSA, 1996, Shanko and Camberlin, 1998, Goddard and Graham, 1999, Latif et al., 1999, Black et al., 2003, Segele and Lamb, 2005). Consequently, season-ahead (March-May) or month-ahead (May) large-scale climate variables that are physically relevant in potentially modulating moisture transport to the basin (or cluster) are selected as potential predictors. Four climate variables are selected here for further evaluation based on outcomes of the aforementioned prediction studies: SST, sea level pressure (SLP), geopotential height (GH) at 500mb, and surface air temperature (SAT). All climate variables are from the National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP/NCAR) reanalysis dataset (Kalnay et al., 1996) at a 2.5˚×2.5˚ grid scale.

To avoid overfitting, the entire process including predictor selection and statistical modeling is processed using cross-validation. To start, drop-one-year precipitation observations for JJAS averaged across the region and each cluster are spatially correlated independently with each global climate variable. As a result, there are total of 1044 global correlation maps given the 29-year time-series, eight clusters plus one non-cluster, and four climate variables. Hence, a program to automatically select highly correlated and justifiable regions as predictors is developed. The following steps describe the subsequent statistical modeling process (Fig. 3):

(1) Grid-cells within each justifiable region (e.g. equatorial Pacific; Fig. 4) with correlation above the 99% significance level are identified (Fig. 5).

(2) The top 10% of the identified grid-cells with the highest correlation in each region is then selected, in order to boost the potential model skill.

(3) For each region, data of the selected grid-cells within the region are spatially averaged (defined as "pre-predictors").

(4) Pre-predictors are combined and transformed (for each cluster or non-cluster, and each dropped year analysis separately) through principal component analysis (PCA; Jolliffe, 2002).

~~Predictor selection and statistical modeling are developed according to the following five steps – for the region as a whole (non-clustered) and for each pre-defined cluster (Fig. 2):~~

~~(1) Precipitation observations for JJAS averaged across the region and each cluster are spatially correlated independently with each global climate variable (e.g. Fig. 3).~~

~~(2) For each spatial correlation, regions with justifiable climatic associations and statistically significant correlations at the 95% level are identified and selected (Table 1).~~

~~(3) For each climate variable region selected (Table 1), data within the region are spatially averaged for 1983-2011 (defined as "pre-predictors").~~

~~(4) Pre-predictors are combined and transformed (for the region or each cluster separately) through principal component analysis (PCA; Jolliffe, 2002).~~

(5a) The top principal components (PCs) from the PCA with a total of 95% variance explained are used as predictors – the direct inputs into the MLR model, otherwise known as the principal component regression (PCR). For the *direct* case, PCR is used to directly predict the grid-level precipitation; for the *indirect* case, PCR is used to predict the intermediate cluster-level precipitation.

(5b) For the *indirect* case only, cluster-level predictions are regressed to the grid-level. Note that the downscaling of cluster-level predictions to grid-level predictions is also cross-validated to avoid overfitting.

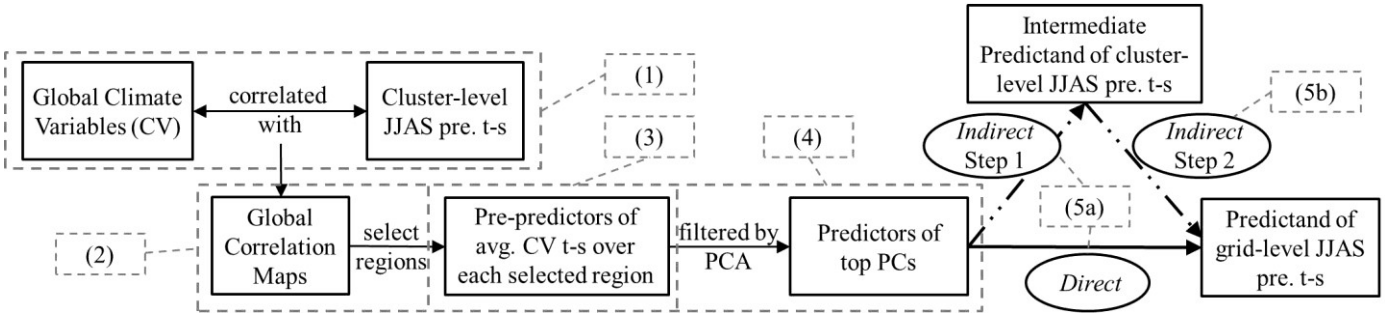


Figure ~~2~~3: Flow chat of data processing for predictors into the statistical model. Numbers framed by dash lines correspond to the procedures listed in the context. Note: pre. – precipitation, t-s – time-series, avg. – average.

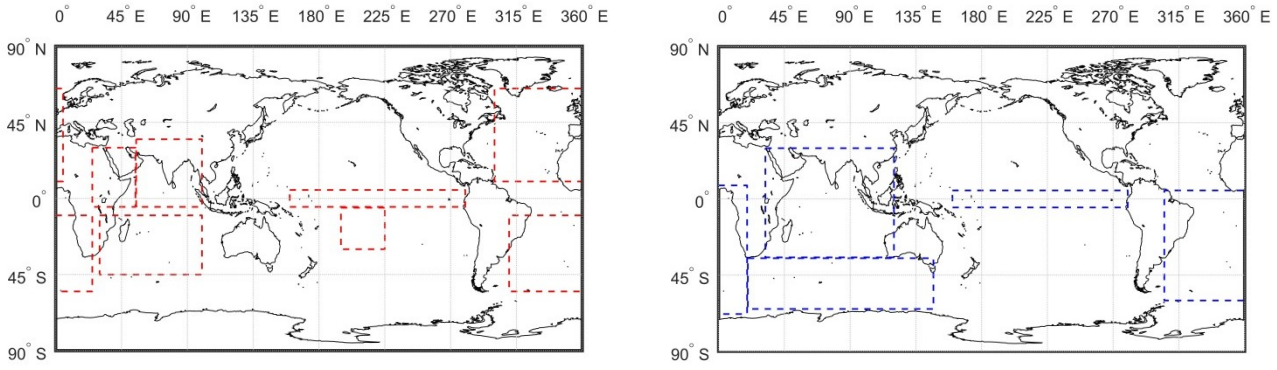


Figure 4: Justifiable climate regions globally for selecting predictors: (a) For SLP and GH at 500 mb with regions including EP, ES, LO, AH, SH, MH, and AM. For SAT, only LO is included. (b) For SST with regions including EP, NI, SI, and AT. Note: EP - equatorial Pacific region, ES – Tahiti island for ENSO measurement, LO - local region, AH - Azores High, SH - St Helena High, MH - Mascarene High, AM - SW Asian Monsoon, NI - North Indian Ocean, SI - South Indian Ocean, AT - Equatorial/South Atlantic Ocean.

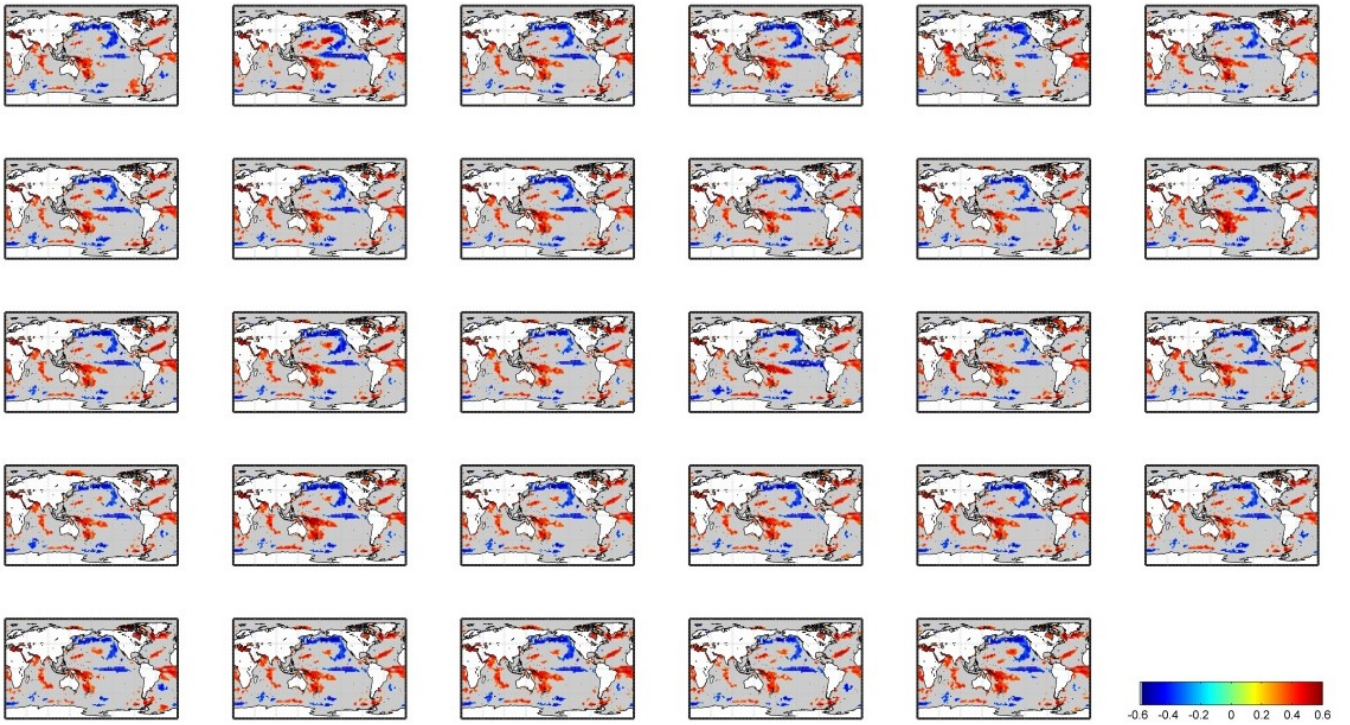
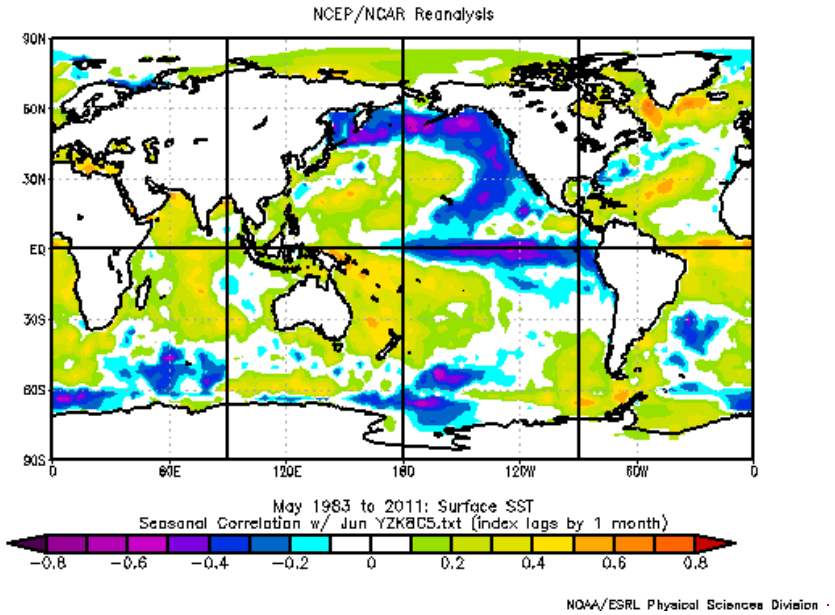


Figure 5: Correlation map between mean JJAS seasonal precipitation time series in Cluster 5 and global SST under cross-validation, with correlations lower than 99% significance level masked out (one-tail test).

Table 1: Climate Variables (C.V.) in May over different regions for each cluster (C1 ~ C8) and region as a whole (non-cluster) used as predictors, with corresponding correlation between the climate variable averaged over the region and the cluster-level JJAS seasonal total precipitation time-series shown (only cells with correlation values shown are used as pre-predictors). After Zhang et al. (2016)

C.V.	SST				SLP						GH at 500mb					SAT	# of pre-predictors
Region	EP	NI	SI	E/SA	LO	EP	AH	SH	MH	AM	LO	EP	AH	SH	MH	AM	LO	
C1	-0.46	-0.46	-0.55	-0.48			0.45	0.45	-0.51			0.52					0.50	9
C2		-0.43	-0.51	-0.43					0.58			0.55			0.50			6
C3			-0.58	-0.59				-0.50					0.57					4
C4			-0.60							-0.58			0.49					3
C5	-0.52	0.52	-0.54	0.59	-0.50	0.61						0.67	0.67	0.54	0.53		0.67	11
C6		0.56		-0.51		0.64						0.66		-0.51				5
C7		0.63	-0.59	0.65				0.44					0.65	0.65			0.44	7
C8			-0.44	0.53	-0.46		0.55						0.63				0.48	6
Non-cluster	-0.47		-0.47	-0.52		0.47				-0.41			0.54	0.58	0.52		0.52	9

Note: EP - equatorial Pacific region, NI - North Indian Ocean, SI - South Indian Ocean, E/SA - Equatorial/South Atlantic Ocean
LO - local region, AH - Azores High, SH - St Helena High, MH - Mascarene High, AM - SW Asian Monsoon



~~**Figure 3: An example of global correlation map built using the correlations between the cluster-level average JJAS precipitation time-series (Cluster 5 in this example) and global sea surface temperature (SST) in May during predictor selection process.**~~

PCA is a common approach in climate modeling to reduce the dimensionality of predictors and remove multi-collinearity, while simultaneously extracting the most dominant signals from the potential predictors, typically reflected in the first few PCs. Since PCA is independent of the predictand, retaining the first few PCs as predictors, in lieu of the original variables, also helps to reduce artificial prediction skill. ~~A scree test (Jolliffe, 2002) is performed to determine the optimal number of PCs to retain as predictors and the amount of variance explained in the predictors.~~

PCR is performed in a "drop-one-year" cross-validation mode to reduce over-fitting effects and therefore avoid overestimation of prediction skill. This requires reconstructing the principal components for the dropped year, and then multiplying the coefficient estimates with each reconstructed PC respectively in order to obtain the final predicted value for the dropped year (e.g. Block and Rajagopalan, 2009, Wilks, 2011). A 95% confidence interval of the cross-validated predictions is also constructed conditioned on model errors. Q-Q plots are evaluated to verify normally distributed residuals (results not included).

For the four scenarios, the model structures are quite similar but have subtle differences which could lead to significantly different outcomes (Table ~~2~~1). Under the NC-D (Eq. (1a, b)) and C-D scenarios (Eq. (2a, b)), the time-series of JJAS seasonal total precipitation in each grid-cell (i.e. at local level) is used as the direct predictand ($Y_{i,t}$); however, the NC-D and C-D scenarios differ, as the former uses the same predictors (X_t) across all the grid-cells, while the latter uses different predictors according to the cluster to which the grid-cell is assigned ($X_{j,t}$). In the indirect case, the cluster-level time-series of JJAS seasonal total precipitation (the time-series averaged over all grid-cells that belong to a given cluster, $Y_{m,t}$ or $Y_{j,t}$) is first predicted (Eq. (3a, b) and (4a, b)). The predicted intermediate product ($\tilde{Y}_{m,t}$ or $\tilde{Y}_{j,t}$) is then used as the only regressor in the second step to estimate the grid-level precipitation ($\tilde{Y}_{i,t}$ or $\tilde{Y}_{i\epsilon j,t}$ for every j; Eq. (3c, d) and (4c, d)). Again, for the C-I scenario, predictors in the first step are unique for each of the eight clusters and grid-cells within that cluster ($X_{j,t}$), while predictors are identical for all grid-cells (X_t) under the NC-I scenario.

Table 21: Equations of linear regression panel models under four scenarios

	Non-clustered		Clustered	
Direct	$Y_{i,t} = \tilde{\alpha}_i + \tilde{\beta}_i X_t + \varepsilon_{i,t}$	..... (1a)	$Y_{i\epsilon j,t} = \tilde{\alpha}_i + \tilde{\beta}_i X_{j,t} + \varepsilon_{i,t}$	..... (2a)
	$\tilde{Y}_{i,t} = \tilde{\alpha}_i + \tilde{\beta}_i X_t$	..... (1b)	$\tilde{Y}_{i\epsilon j,t} = \tilde{\alpha}_i + \tilde{\beta}_i X_{j,t}$	..... (2b)
Indirect	$Y_{m,t} = \tilde{\alpha} + \tilde{\beta} X_t + \varepsilon_t$	..... (3a)	$Y_{j,t} = \tilde{\alpha}_j + \tilde{\beta}_j X_{j,t} + \varepsilon_{j,t}$	..... (4a)
	$\tilde{Y}_{m,t} = \tilde{\alpha} + \tilde{\beta} X_t$	..... (3b)	$\tilde{Y}_{j,t} = \tilde{\alpha}_j + \tilde{\beta}_j X_{j,t}$	..... (4b)
	$Y_{i,t} = \tilde{\eta}_i + \tilde{\gamma}_i \tilde{Y}_{m,t} + \nu_{i,t}$	..... (3c)	$Y_{i\epsilon j,t} = \tilde{\eta}_i + \tilde{\gamma}_i \tilde{Y}_{j,t} + \nu_{i,t}$	..... (4c)
	$\tilde{Y}_{i,t} = \tilde{\eta}_i + \tilde{\gamma}_i \tilde{Y}_{m,t}$	..... (3d)	$\tilde{Y}_{i\epsilon j,t} = \tilde{\eta}_i + \tilde{\gamma}_i \tilde{Y}_{j,t}$	..... (4d)

where Y- predictand of JJAS seasonal total precipitation; X- two predictors of top two PCs;
ε,ν - error terms; $\tilde{Y}$ - predicted values of JJAS seasonal total precipitation; $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\eta}$, $\tilde{\gamma}$- estimated coefficients; i- grid-cell index; t- time (year) index; j- cluster index; $i \in j$- grid-cell i that belongs to clusterj; m- mean over entire study region that is equivalently the only one cluster.

3.3 Dynamical modeling approach

The North American Multi-Model Ensemble (NMME; Kirtman et al., 2014) is an experimental multi-model system consisting of coupled dynamical models from various modeling centers in North America ~~that includes seasonal predictions~~. To our knowledge, it is also the most extensive multi-model seasonal prediction archive. The NMME provides gridded climate predictions that cover regions globally and with different lead times. The hindcasts of monthly mean precipitations are easily accessible through the International Research Institute for Climate and Society (IRI) website (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/), and can be easily aggregated to seasonal totals for comparison with the statistical model results in this study. ~~To compare with statistical model predictions~~Therefore, NMME JJAS seasonal precipitation predictions (1˚×1˚ grid-cells) are extracted from model ensembles that cover the same time period (1983–2011), geographic region (western Ethiopia), and with the same lead time (predictions made on June 1). A subset of 10 NMME models meet these criteria and are retained for further evaluation: (1) COLA-RSMAS-CCSM3, (2) COLA-RSMAS-CCSM4, (3) GFDL-CM2p1, (4) GFDL-CM2p1-are04, (5) GFDL-CM2p5-FLOR-A06, (6) GFDL-CM2p5-FLOR-B01, (7) IRI-ECHAM-AnomalyCoupled, (8) IRI-ECHAM-DirectCoupled, (9) NASA-GMAO, (10) NCEP-CFSv2. The names are kept the same as on the International Research Institute for Climate and Society (IRI) data repository website.

The NMME predictions for each of the 10 models are bias-corrected by applying probability mapping (e.g. Block et al., 2009, Teutschbein and Seibert, 2012, Chen et al., 2013) under cross-validation, subject to the observational dataset from NMA (Fig. 6). This is performed on a grid-cell by grid-cell basis on standardized data (the NMME dataset is reshaped to 0.1°×0.1° grid-cells to match the observational NMA dataset grid-cell size). The basic steps include:

(1) Fit gamma distributions to drop-one-year time-series from each observed and NMME grid-cell; for NMME this is performed on an individual model basis using all ensemble members available. (Goodness-of-fit tests indicate gamma distributions are appropriate; results not shown.)~~.~~

(2) Translate gamma distributions into cumulative distribution functions (CDF).

(3) For any given dynamical model prediction at the grid-cell level, a corrected prediction value is attained by mapping from the modeled CDF to the observed CDF and applying the inverse gamma distribution. This is repeated for all grid-cells, ~~and~~ all NMME models, and all dropped years.

After correction, the gamma CDF of predictions and observations approximately match (Fig. ~~4a~~6a). Additionally, each ensemble still retains its variability over time, though the overall ensemble mean is shifted to closely match observation (Fig. ~~4b~~6b).

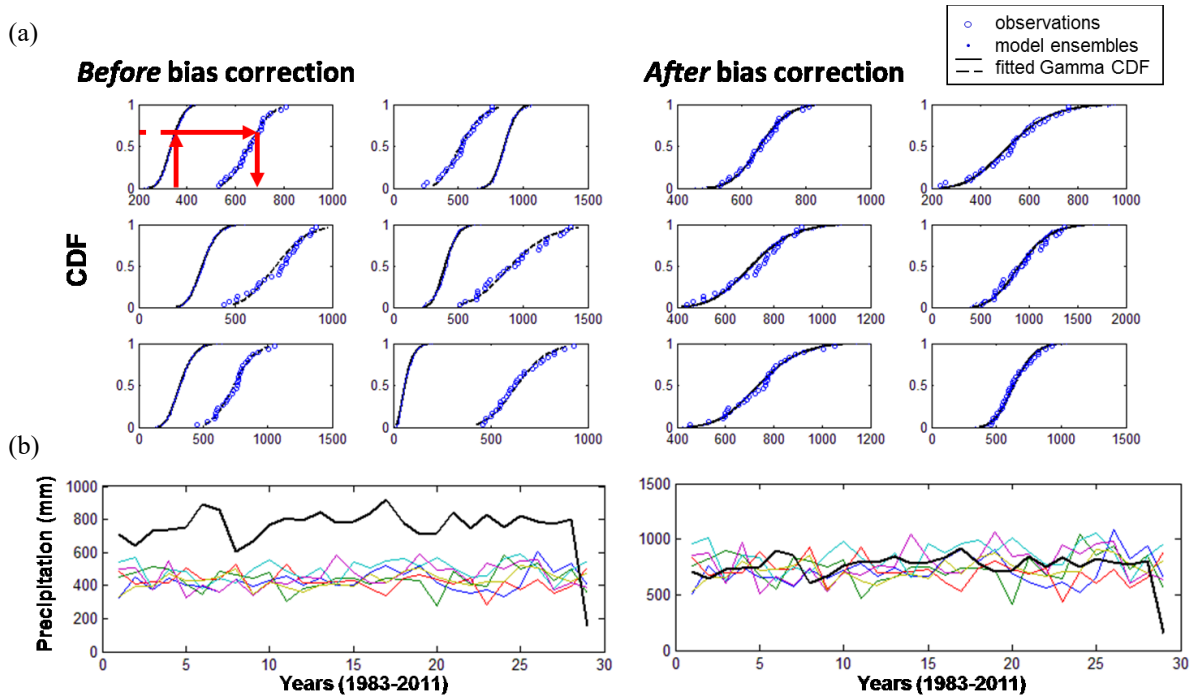


Figure ~~4~~6: (a) bias correction of NMME predictions using probability mapping; (b) precipitation time-series from NMME (colored lines) before and after correction, compared to observations (black line). Examples are shown for randomly selected six grid-cells.

3.4 Performance metrics

Pearson correlations are used to measure the standardized covariance between observations and predictions. Ranked probability skill scores (RPSS; Wilks, 2011) are also evaluated to determine categorical skill based on probabilistic predictions. Here, the data are split into three equal terciles representing below-normal, near-normal, and above-normal conditions. A perfect prediction yields an RPSS of 100%, and a prediction with less skill than climatology (long-term averages) yields an RPSS of less than zero. Median RPSS values from all 29 years are reported.

~~Overall model superiority is evaluated by Akaike information criterion (AIC), Bayesian information criterion (BIC), and generalized cross validation (GCV) scores (Craven and Wahba, 1979, Manning et al., 2008). All metrics reward model parsimony by penalizing models with a larger number of predictors. Smaller AIC, BIC and GCV scores are preferred. The equations used to calculate AIC, BIC, and GCV, respectively, are given by:~~

~~$AIC = N \times \log(RSS/N) + 2 \times K$ ..... (5)~~

~~$BIC = N \times \log(RSS/N) + \log(N) \times K$ ..... (6)~~

~~$GCV = RSS/[N \times (1 - K/N)^2]$ ..... (7)~~

~~where N is the number of years (29 years), K is the number of predictors used in the regression, and RSS is the residual sum of squares (equal to the difference between observations and predictions in each year squared, summed over all the years).~~

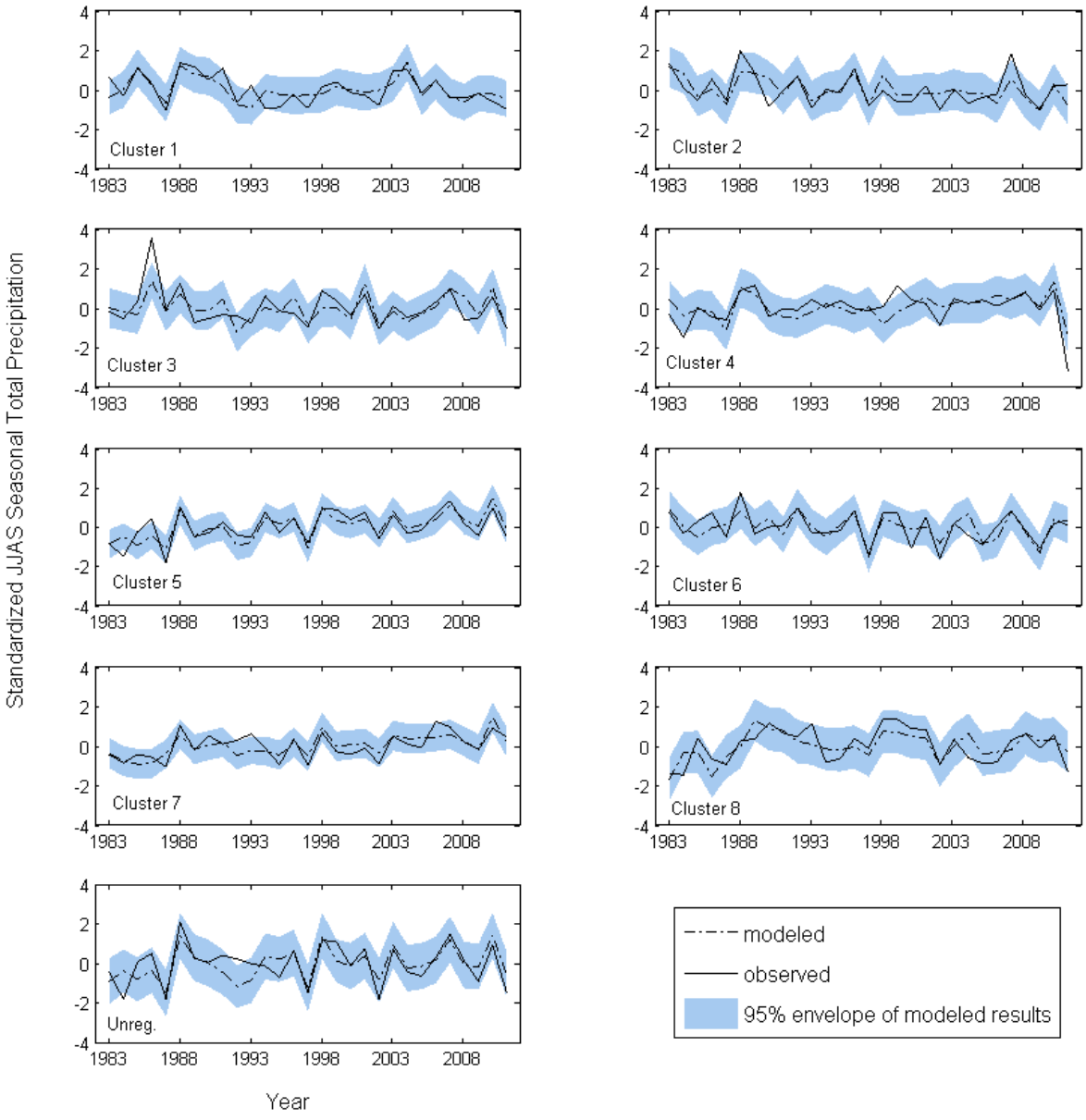
4 Results

4.1 Statistical model predictions

~~Using the scree test (Cattell, 1966), the first two PCs are retained as predictors for each cluster for the *clustered* case, and the first three PCs are retained for the *non clustered* case. In all cases, the total variance explained by the PCs retained is approximately 70%.~~

~~Cluster level model predictions demonstrate good skill, as most observations fall within the predicted 95% confidence envelope (Fig. 5), and strong positive correlations with observations – ranging from 0.68 to 0.84 – are evident across all clusters and the *non clustered* study region (Table 3). Additionally, all RPSS values are positive, indicating superior prediction skill over climatology. Among all clusters, *Cluster 5*, in agriculturally rich central northwestern Ethiopia (Fig. 1), performs best, with correlation and RPSS values of 0.84 and 74.3%, respectively.~~

Correlations between cluster-level model predictions and observations range from -0.18 to 0.50, with Cluster 5 having the highest correlation and Cluster 6 the lowest (Table 2). In approximately 1/5 of the 29 years, the observation falls outside the prediction envelope (Fig. 7), indicating model overfitting and an inability of the predictors to capture precipitation variability. For RPSS, 5 out of 8 clusters indicate superior prediction skill over climatology (Table 2). Improvement in terms of RPSS over the non-cluster scenario is evident for Cluster 1, 3, 5 and 7. Among all clusters, Cluster 5, in agriculturally rich central-northwestern Ethiopia (Fig. 2), performs the best, with correlation and RPSS values of 0.5 and 27%, respectively. Cluster 2, 4, 6 and 8, however, show deteriorated RPSS compared to non-cluster scenario, although those clusters are mainly regions outside Ethiopia and southern Ethiopia (Fig. 2) where water resources and agricultural activities are considerably less (Fig. 1).


Standardized JJAS Seasonal Total Precipitation
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Unreg.
4
2
0
-2
-4
1983
1988
1993
1998
2003
2008
Year
modeled
observed
95% envelope of modeled results

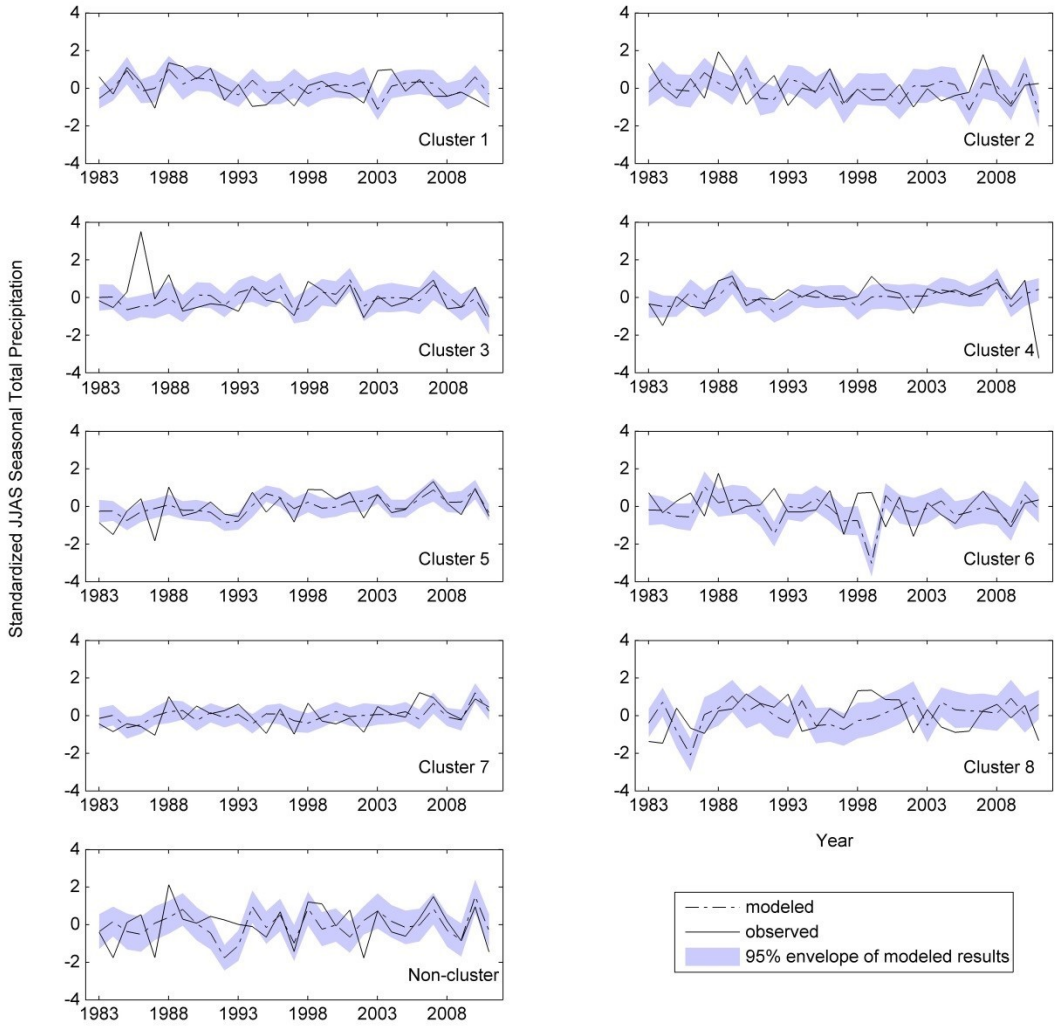


Figure ~~5~~7: cluster-level predictions and observations under C-I and NC-I scenario, with drop-one-year cross-validation. The 95% envelope shows the 95% confidence interval constructed using model errors.

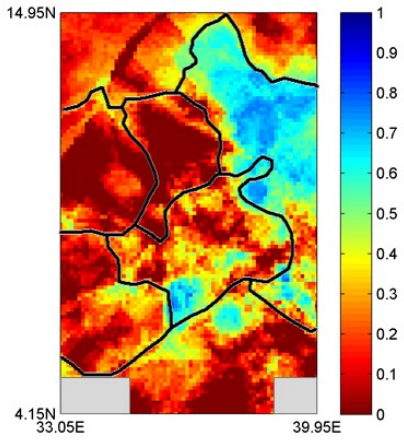
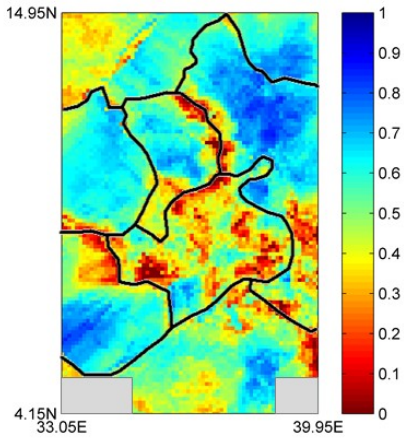
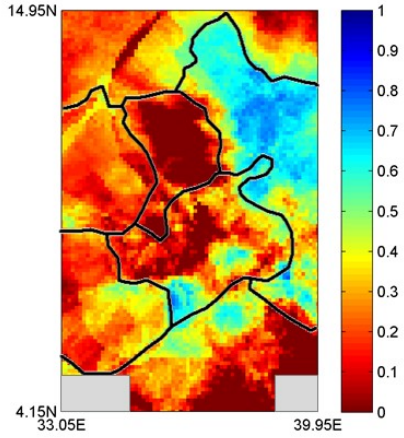
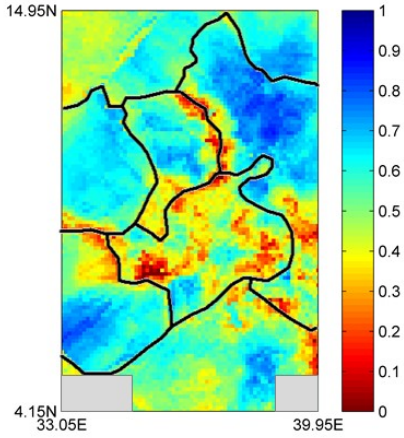
5 ~~**Table 3: Correlation coefficients (Corr.) and RPSS for predictions (drop-one-year cross-validated) at cluster level compared to observations under C-I and NC-I scenario.**~~

~~Cluster~~	~~C1~~	~~C2~~	~~C3~~	~~C4~~	~~C5~~	~~C6~~	~~C7~~	~~C8~~	~~Non-cluster~~
~~Corr.~~	~~0.741~~	~~0.695~~	~~0.711~~	~~0.683~~	~~0.838~~	~~0.744~~	~~0.751~~	~~0.699~~	~~0.739~~
~~RPSS (%)~~	~~45.23~~	~~26.04~~	~~36.16~~	~~19.82~~	~~74.30~~	~~5.44~~	~~51.91~~	~~48.21~~	~~48.72~~

Table 2 Correlation coefficients (Corr.) and RPSS for predictions (drop-one-year cross-validated) at cluster level compared to observations under C-I and NC-I scenario.

Cluster	C1	C2	C3	C4	C5	C6	C7	C8	Non-cluster
Corr.	0.163	-0.010	0.179	0.188	0.504	-0.180	0.351	-0.122	0.297
RPSS(%)	33.41	-21.66	43.01	12.46	27.40	-37.79	20.63	-55.96	13.25

~~At the grid scale, correlations between predictions and observations are clearly superior for the *clustered* case versus the *non-clustered* case (Fig. 6). Some parts of the region reach a correlation of 0.9, such as central northwestern Ethiopia, which is consistent with the region of high cluster level prediction skill (*Cluster 5*). The average correlation over all grid cells is approximately 0.51 (*direct*) and 0.53 (*indirect*) for *clustered* predictions, compared to 0.24 (*direct*) and 0.27 (*indirect*) for the *non-clustered* predictions (Table 4), although spatial differences are clearly apparent (Fig. 6). In addition to higher average correlations, standard deviations of correlations are also lower in the *clustered* case than in the *non clustered* case, indicating a more concentrated correlation distribution for these higher values. The percentage of grid cells with correlations passing the 95% significance test increases from approximately 30% in the *non clustered* case to more than 80% in the *clustered* case (Table 4).~~ At the grid scale, depending on the case (*direct* or *indirect)*, and for different clusters, correlations between predictions and observations can favor the *clustered* case or the *non-clustered* case (Fig. 8). In general, the *indirect* model provides a smoother pattern of correlations, with grid-cells showing a negative correlation in the *direct* case now improved to near or above zero (Fig. 8). For example, Cluster 5 under the *indirect* case illustrates a more consistent positive correlation within the cluster. Some parts of the region reach a correlation over 0.6, such as central-northwestern Ethiopia (Cluster 5), which is consistent with the region of high cluster-level prediction skill. The percentage of grid-cells with correlations passing the 95% significance test is the highest for the NC-D case (Table 3); however, some locations demonstrate the lowest skills among all four scenarios.

	Non-clustered	Clustered
Direct		
Indirect		

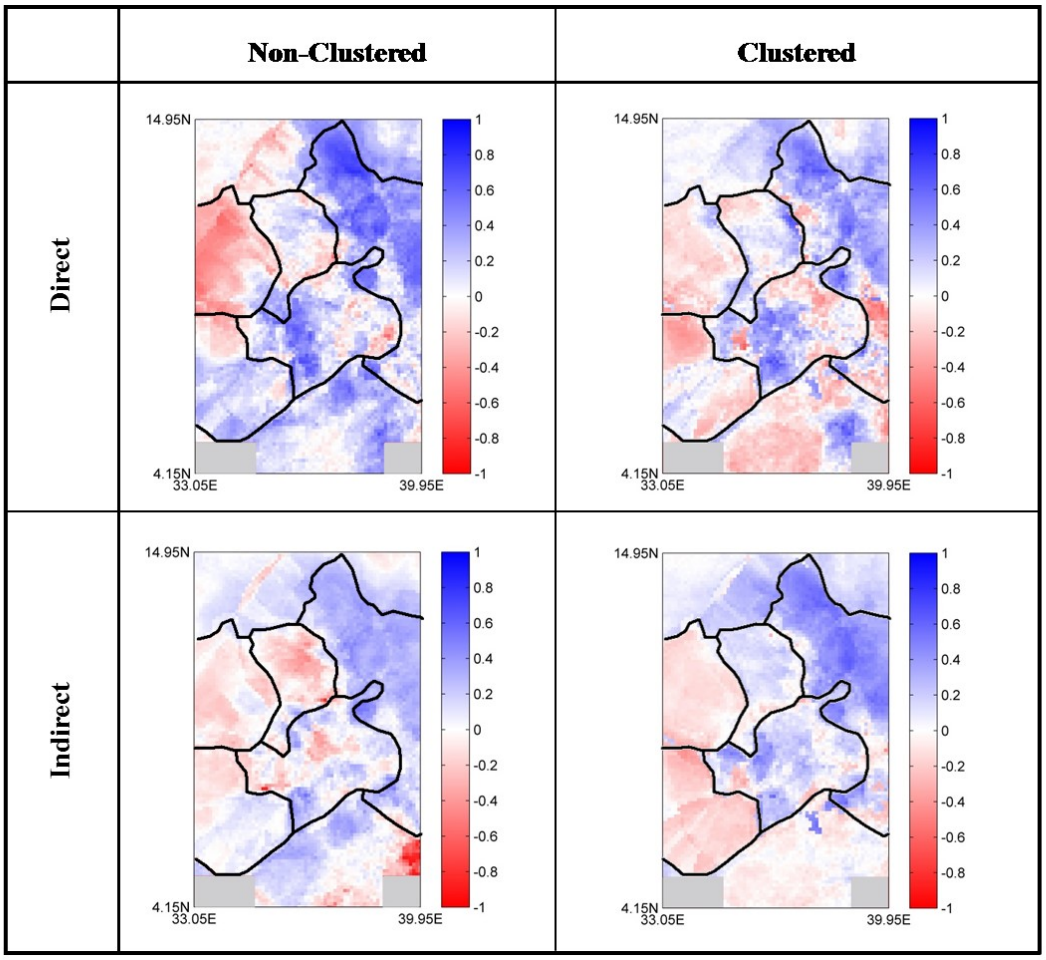


Figure ~~6~~8: Pearson correlations between grid-level observations and predictions under four scenarios, with the clustering boundary delineated roughly in black.

Table ~~4~~3: Grid-level Pearson correlation and RPSS statistics

~~Statistical Model~~	~~Grid-level correlations~~			~~Grid-level RPSS~~		
	~~mean~~	~~stdev~~	~~significant corr %~~	~~mean (%)~~	~~stdev (%)~~	~~positive RPSS %~~
~~NC-D~~	~~0.237~~	~~0.245~~	~~28.1%~~	~~5.42~~	~~18.46~~	~~58.8%~~
~~NC-I~~	~~0.272~~	~~0.247~~	~~32.3%~~	~~5.32~~	~~17.09~~	~~60.7%~~
~~C-D~~	~~0.509~~	~~0.172~~	~~80.8%~~	~~19.16~~	~~19.18~~	~~84.4%~~
~~C-I~~	~~0.532~~	~~0.146~~	~~87.1%~~	~~26.47~~	~~21.47~~	~~90.0%~~
~~Dynamical Model~~						
~~(9) NASA-GMAO~~	~~0.300~~	~~0.149~~	~~36.1%~~	~~2.32~~	~~21.20~~	~~54.3%~~
~~(10) NCEP-CFSv2~~	~~0.310~~	~~0.155~~	~~37.3%~~	~~3.66~~	~~16.61~~	~~61.0%~~

5

Statistical Model	Grid-level correlations			Grid-level RPSS		
	mean	stdev	significant corr %	mean (%)	stdev (%)	positive RPSS %
NC-D	0.128	0.258	19.3%	-5.21	27.0	42.8%
NC-I	0.063	0.186	3.13%	-2.26	14.6	43.9%
C-D	0.055	0.230	10.6%	-14.0	31.0	33.9%
C-I	0.081	0.206	12.4%	-9.60	29.4	44.4%

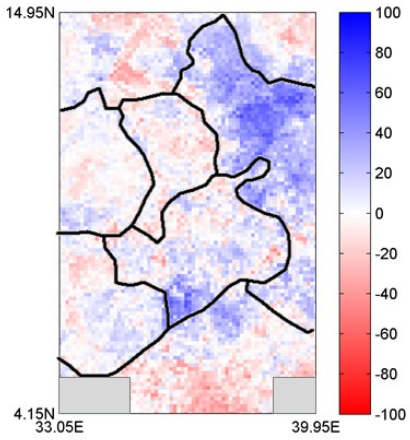
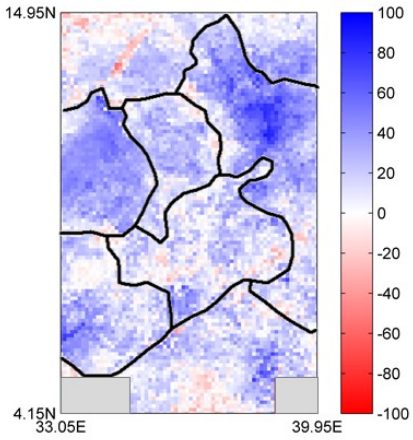
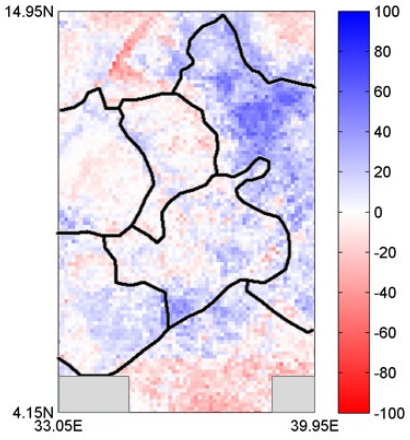
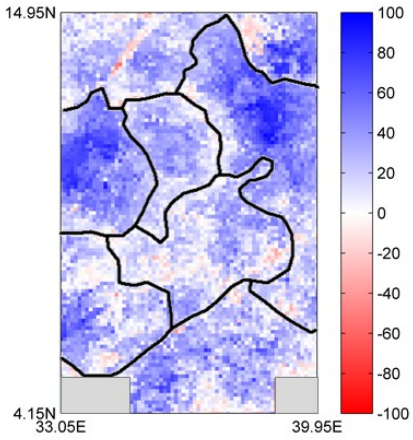
Dynamical Model						
(1)	-0.105	0.209	0.51%	-31.4	25.4	5.70%
(2)	0.133	0.171	6.26%	-14.2	24.6	27.0%
(3)	0.086	0.130	2.08%	-14.9	25.2	26.2%
(4)	0.027	0.156	0.38%	-14.4	19.3	22.6%
(5)	0.067	0.170	1.64%	-9.66	17.0	28.4%
(6)	0.139	0.165	6.53%	-5.66	16.7	38.1%
(7)	0.102	0.130	1.67%	-8.64	17.6	31.7%
(8)	0.009	0.185	0.90%	-10.3	14.8	26.7%
(9)	0.244	0.149	23.1%	-2.33	21.8	46.0%
(10)	0.244	0.149	21.2%	-1.09	16.8	48.9%

~~Similar findings are evident by evaluating the RPSS. The *non-clustered* predictions are modestly skillful, particularly for the same region of central-northwestern Ethiopia (Fig. 7), with an average RPSS of approximately 5.4% (both *direct* and *indirect*) over the entire study region, however RPSS values improve nearly fourfold to 19.2% and 26.5% in the *clustered* case (*indirect* and *direct*, respectively). Additionally, the percentage of grid-cells with positive RPSS values reaches 84.4% – 90.0% in the *clustered* case (Table 4).~~

Similar findings are evident by evaluating the RPSS except for Cluster 8; instead of improving with increased RPSS in the *indirect* case, the grid-scale predictions deteriorate given poor cluster-level prediction (for the C-I case). However, the percentage of grid-cells with positive RPSS values overall is still the highest for the C-I case (Table 3), indicating the *clustered indirect* case is superior in terms of the number of grid-cells with improved skill compared to using climatology, particularly for grid-cells associated with skillful intermediate cluster-level predictions. The predictions are most skillful for the same region of central-northwestern Ethiopia (Cluster 5; Fig. 9) with 87% of its grid-cells showing positive RPSS and a spatial average RPSS value of 15% under the C-I scenario (Table 4).

Table 4: Grid-level Pearson correlation and RPSS statistics for grid-cells within Cluster 5

Statistical Model	Grid-level correlations			Grid-level RPSS		
	mean	stdev	significant corr %	mean (%)	stdev (%)	positive RPSS %
NC-D	0.378	0.211	60.7%	19.1	22.9	80.3%
NC-I	0.265	0.111	12.8%	8.33	14.8	70.3%
C-D	0.229	0.244	30.5%	6.91	24.1	62.3%
C-I	0.345	0.165	55.7%	14.7	13.3	87.1%
Dynamical Model						
(9)	0.353	0.110	46.8%	8.21	18.2	65.7%
(10)	0.248	0.130	18.4%	3.92	16.2	59.5%

	Non-clustered	Clustered
Direct		
Indirect		

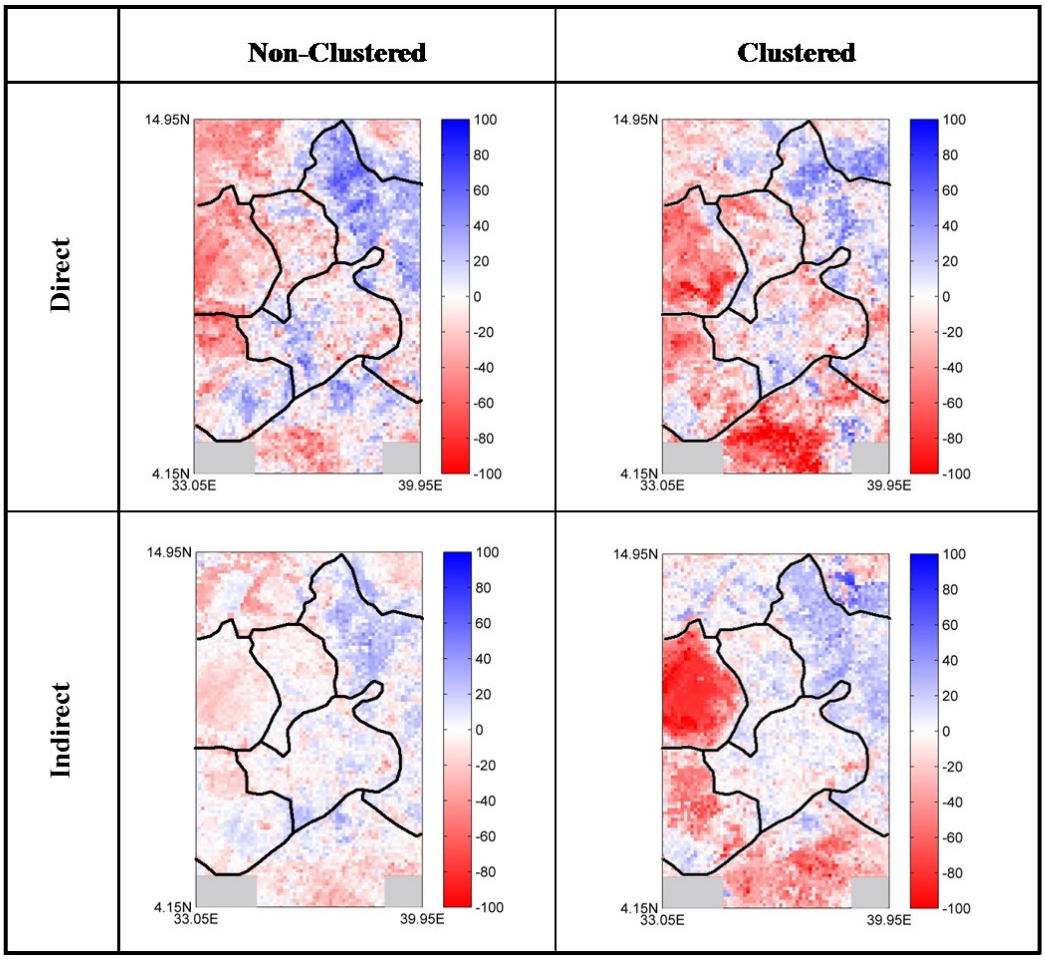


Figure ~~7~~2: grid-level RPSS (%) under four scenarios using climate variables as predictors, with the clustering boundary delineated roughly in black.

~~At the grid scale, predictions by the *indirect* approach generally outperform *direct* approach predictions, based on AIC, BIC and GCV values (Table 5), as well as correlation and RPSS values (Tables 4). Using the predicted cluster level precipitation to predict grid level precipitation (the *indirect* case) appears to help reduce the effect of over fitting and smooth grid scale noise. From another perspective, the results also suggest that precipitation signals at the regional scale are better explained by large scale climate variables, while at highly localized scales the signal is less evident. Obviously, however, this is dependent on cluster size and the degree of spatial coherence within each cluster, as demonstrated in this study.~~

~~**Table 5: grid-level AIC, BIC, and GCV value statistics**~~

~~Statistical Model~~	~~AIC~~		~~BIC~~		~~GCV~~	
	~~mean~~	~~stdev~~	~~mean~~	~~stdev~~	~~mean~~	~~stdev~~
~~NC-D~~	~~281.05~~	~~17.86~~	~~285.15~~	~~17.86~~	~~2.00E+04~~	~~1.46E+04~~
~~NC-I~~	~~277.47~~	~~17.37~~	~~280.20~~	~~17.37~~	~~1.73E+04~~	~~1.20E+04~~
~~C-D~~	~~272.79~~	~~15.38~~	~~276.89~~	~~15.38~~	~~1.41E+04~~	~~7.53E+03~~
~~C-I~~	~~270.05~~	~~15.33~~	~~272.78~~	~~15.33~~	~~1.27E+04~~	~~6.84E+03~~

4.2 Dynamical model predictions

The RPSS values based on the prediction ensembles of each dynamical model improve significantly after bias correction~~., however, t~~The median RPSS values over all the grid-cells are ~~still~~ now close to zero (Fig. ~~8~~10) with~~. Only~~ two models, NASA-GMAO and

NCEP-CFSv2, showing ~~a positive mean~~the highest RPSS value (-2.3~~2~~% and -1.1~~3.66~~%, respectively; Table 3~~4~~). These two dynamical models also exhibit generally higher grid-level correlations over the study region (averaging 0.24 for both models~~0.30 and 0.31, respectively~~; Table 3~~4~~ and Fig. 11~~9~~), as compared with other NMME models. The two best performing dynamical models after bias correction show advantage over statistical models, as assessed by correlation and RPSS metrics; however, all other dynamical models are inferior to the statistical models under NC-D and C-I scenarios, particularly given the percent of grid-cells with significant correlation and positive RPSS metrics (Table 3). ~~However, their overall prediction performance is still clearly inferior to that of the *clustered* statistical models,~~

Within a certain cluster, statistical models may perform better than all dynamical models. For example, for Cluster 5, all statistical models show higher average RPSS values than that of all dynamical models (Table 4). The percentage of grid-cells with significant correlation reaches 61% for the statistical model under NC-D scenario, compared to the highest value of 47% among all the dynamical models. Similarly, the percentage with positive RPSS achieves 87% under C-I scenario as opposed to 66% for dynamical models. Note that the dynamical models also produce predictions in a lower spatial resolution (1°×1°) than the statistical models (0.1°×0.1°).~~as assessed by correlation and RPSS metrics (Table 4).~~

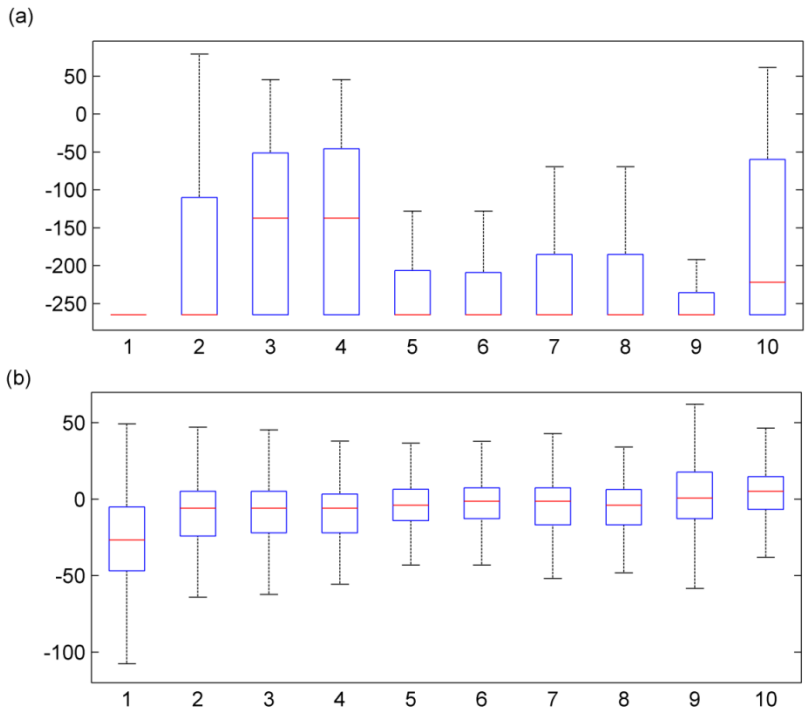


Figure ~~8~~10: Boxplots of grid-level RPSS (%) for 10 dynamical models from NMME (a) before and (b) after bias correction, labeled with the same number as listed in the context. Note: For each box plot, the line inside the box is the median, the box edges represent the 25th and 75th percentiles, and the whiskers extend to the most extreme data points not considered outliers (outliers not shown).

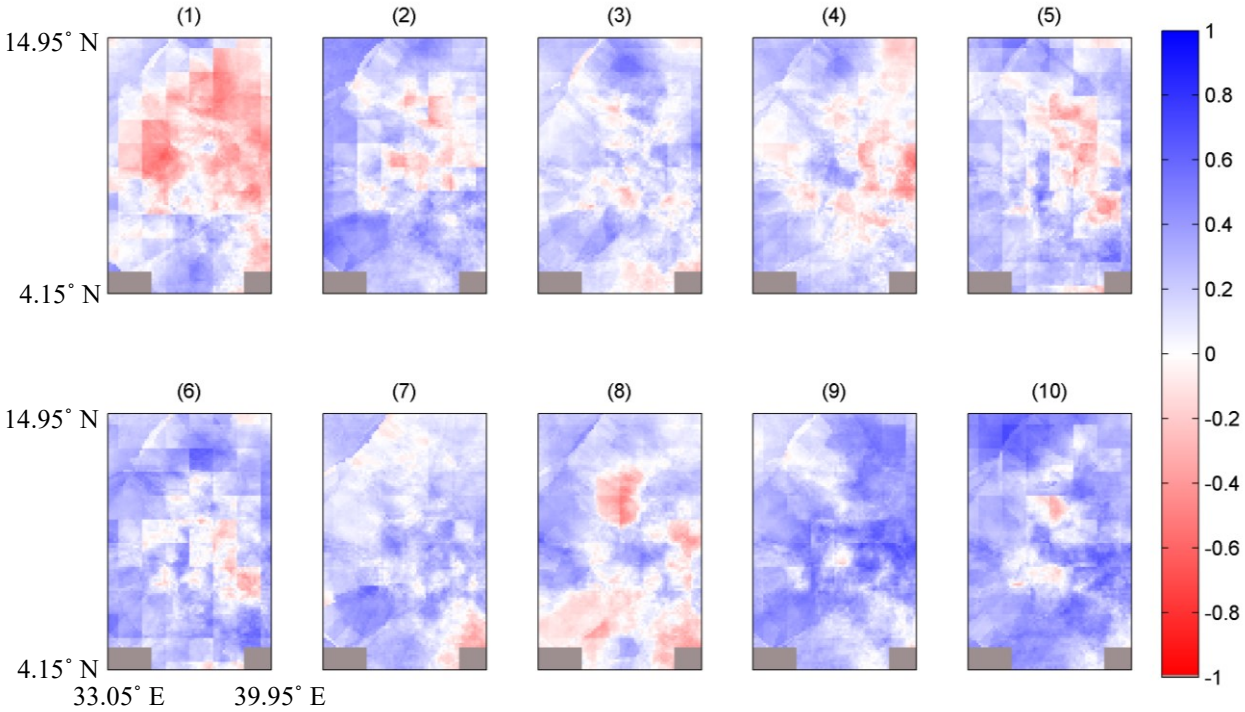


Figure ~~9~~11: Pearson correlations between grid-level observations and ensemble mean of bias-corrected predictions for 10 dynamical models from NMME, labeled with the same number as listed in the context. Note that the scale ranges from -1 to 1.

5 Conclusions and discussion

This study demonstrates the potential for season-ahead large-scale climate information to produce skillful and credible high-resolution precipitation predictions under a *clustered indirect* approach in western Ethiopia. ~~At the regional scale, the approach shows promise, particularly compared to current NMA operational forecasts, which are only moderately more skillful than climatology~~ At the regional scale, the approach shows promise for northwestern Ethiopia (Cluster 1, 3, 5, and 7), particularly compared to current NMA operational forecasts, which are only moderately more skillful than climatology (Korecha and Sorteberg, 2013). The regional average RPSS in this study under the *clustered* case ranges from 21% to 43% for northwestern Ethiopia, as opposed to values under 6% for NMA operational forecast (Korecha and Sorteberg, 2013). The approach adopted here also advances on previous studies (Gissila et al., 2004, Block and Rajagopalan, 2007, Korecha and Barnston, 2007, Diro et al., 2011b, Segele et al., 2015) by first applying an objective cluster analysis and then conditionally constructing high-resolution predictions. A unique set of predictors is applied to each cluster, which contributes to superior prediction performance at ~~both~~ cluster ~~and grid~~ levels in northwestern Ethiopia, as compared with predictions from the *non-clustered* approach. Grid-level prediction under the *clustered indirect* case also reduces the effect of over-fitting relative to the *direct* case and improves negative RPSS values to near or above zeros; that said, the *non-clustered direct* case also illustrates higher correlation and RPSS values on average.

~~Although predictions from the statistical model *clustered* approach are superior to all dynamical predictions for this study,~~ Two out of 10 NMME dynamical models, NASA-GMAO and NCEP-CFSv2, demonstrate overall superior performance to the statistical models; however, for certain regions such as Cluster 5, the performance of statistical models under *clustered indirect* and *non-clustered direct* cases is still superior. It is also worth noting that the statistical model predictions are at a one hundred times finer spatial resolution than the dynamical models providing additional advantages at the local scale, when skillful. Nevertheless,

improvements in dynamical models continue and their application to seasonal precipitation prediction is likely to grow (e.g. Palmer et al., 2004, Saha et al., 2006, Lim et al., 2009). ~~Multi-model combinations of statistical and dynamical models were also investigated for potential improvement of prediction skills through pooling, linear regression, and Bayesian model averaging (BMA; Raftery et al., 1997) using the best statistical model (C-I) and two best dynamical models (NASA-GMAO and NCEP-CFSv2), however, the overall performance was inferior to the single statistical C-I model (results not shown).~~

~~Even though *clustered* statistical model predictions are promising overall, it is worth noting that r~~Relatively poor prediction performance is evident in some locations such as southwestern Ethiopia and regions outside Ethiopia,~~. One such place is along cluster boundaries, where assignment of grid-cells to one cluster versus the neighboring cluster is almost arbitrary, and clearly less certain than grid-cells falling within the central parts of clusters (Fig. 6 and 7). Poor prediction skill is also evident in some of the mountainous regions of the study area,~~ where the hydroclimatic processes that produce precipitation ~~are likely~~ might be driven by ~~orographic and other~~ local factors or other regional climate patterns rather than large-scale climate variables identified in this study. A previous study (Zhang et al., 2016) has shown that the influence of ENSO on JJAS precipitation in western Ethiopia decreases generally from north to south, and is likely one of the reasons why skills are relatively low in southwestern Ethiopia. Cluster 5 was also identified with the strongest connection to equatorial Pacific SST (Zhang et al., 2016), which is consistent with the highest skill found in this study. Other regions with low prediction skill show relatively strong connections to SST and pressure systems in neighboring oceanic regions. However, connections with those climate patterns appear to be less robust than with ENSO, making the predictions in their associated regions less skillful. To test the prospects for improving prediction performance by including season-ahead local variables, soil moisture and spring rains were investigated; however, no significant improvement was found~~, for the *clustered* case, and correlations actually deteriorate for the direct case. Thus adding local predictors in this case may simply serve to introduce more noise and encourage over-fitting.~~ in fact, performance skill deteriorates when adding local predictors may simply serve to introduce more noise and encourage over-fitting.

Additional prediction features also warrant future attention, including longer prediction lead times and evaluation of other relevant characteristics (e.g. intra-seasonal dry spells, seasonal onset or cessation, etc.). As observational datasets continue to grow, data-driven cluster analysis and statistical modeling approaches may be expected to improve. Improving predictive capabilities may not be a complete panacea, but it can continue to be an important part of a decisions-maker's portfolio as they cope with hydroclimatic variability and its inherent risks.

6 Data availability

The National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP/NCAR) reanalysis dataset can be accessed through the National Oceanic & Atmospheric Administration (NOAA) Earth System Research Laboratory (ESRL) website (https://www.esrl.noaa.gov/psd/data/reanalysis/).

The NMME hindcasts are available through the International Research Institute for Climate and Society (IRI) website (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/).

The gridded precipitation dataset in western Ethiopia is available upon request from NMA (http://www.ethiomet.gov.et/).

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This study was supported by NASA Project NNX14AD30G and NSF PIRE Project 1545874. We acknowledge the National Meteorological Agency of Ethiopia for sharing data.

References

ANDERSON, D., STOCKDALE, T., BALMASEDA, M., FERRANTI, L., VITART, F., MOLTENI, F., DOBLAS-REYES, F., MOGENSEN, K. & VIDARD, A. 2007. Development of the ECMWF seasonal forecast System 3. *ECMWF Technical Memoranda,* 503.

AWULACHEW, S. B., YILMA, A. D., LOULSEGED, M., LOISKANDL, W., AYANA, M. & ALAMIREW, T. 2007. *Water resources and irrigation development in Ethiopia*, Iwmi.

BADR, H. S., ZAITCHIK, B. F. & DEZFULI, A. K. 2015. A tool for hierarchical climate regionalization. *Earth Science Informatics***,** 1-10.

BARRETT, C. B. 1993. THE DEVELOPMENT OF THE NILE HYDROMETEOROLOGICAL FORECAST SYSTEM1. Wiley Online Library.

BEKELE, F. 1997. Ethiopian Use of ENSO Information in Its Seasonal Forecasts. *Internet Journal of African Studies*.

BLACK, E., SLINGO, J. & SPERBER, K. R. 2003. An observational study of the relationship between excessively strong short rains in coastal East Africa and Indian Ocean SST. *Monthly Weather Review,* 131**,** 74-94.

BLOCK, P. & GODDARD, L. 2012. Statistical and Dynamical Climate Predictions to Guide Water Resources in Ethiopia. *Journal of Water Resources Planning and Management,* 138**,** 287-298.

BLOCK, P. & RAJAGOPALAN, B. 2009. Statistical–Dynamical Approach for Streamflow Modeling at Malakal, Sudan, on the White Nile River. *Journal of Hydrologic Engineering,* 14**,** 185-196.

BLOCK, P. J., FILHO, F. A. S., SUN, L. & KWON, H. H. 2009. A Streamflow Forecasting Framework Using Multiple Climate and Hydrological Models. *Journal of the American Water Resources Association,* 45**,** 828-843.

BLOCK, P. J. & RAJAGOPALAN, B. 2007. Interannual Variability and Ensemble Forecast of Upper Blue Nile Basin Kiremt Season Precipitation. *J. Hydrometeor,* 8**,** 327-343.

CAMBERLIN, P. 1997. Rainfall Anomalies in the Source Region of the Nile and Their Connection with the Indian Summer Monsoon. *J. Climate,* 10**,** 1380-1392.

CAMBERLIN, P. & PHILIPPON, N. 2002. The East African March–May Rainy Season: Associated Atmospheric Dynamics and Predictability over the 1968–97 Period. *Journal of Climate,* 15**,** 1002-1019.

CATTELL, R. B. 1966. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research,* 1**,** 245-276.

CHEN, J., BRISSETTE, F. P., CHAUMONT, D. & BRAUN, M. 2013. Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research,* 49**,** 4187-4205.

COELHO, C. A. S., PEZZULLI, S., BALMASEDA, M., DOBLAS-REYES, F. J. & STEPHENSON, D. B. 2004. Forecast Calibration and Combination: A Simple Bayesian Approach for ENSO. *Journal of Climate,* 17**,** 1504-1516.

CRAVEN, P. & WAHBA, G. 1979. Smoothing noisy data with spline functions. *Numerische Mathematik,* 31**,** 377-403.

DINKU, T., HAILEMARIAM, K., MAIDMENT, R., TARNAVSKY, E. & CONNOR, S. 2014. Combined use of satellite estimates and rain gauge observations to generate high-quality historical rainfall time series over Ethiopia. *International Journal of Climatology,* 34**,** 2489-2504.

DIRO, G. T., BLACK, E. & GRIMES, D. I. F. 2008. Seasonal forecasting of Ethiopian spring rains. *Meteorological Applications,* 15**,** 73-83.

DIRO, G. T., GRIMES, D. I. F. & BLACK, E. 2011a. Teleconnections between Ethiopian summer rainfall and sea surface temperature: part I—observation and modelling. *Climate Dynamics,* 37**,** 103-119.

DIRO, G. T., GRIMES, D. I. F. & BLACK, E. 2011b. Teleconnections between Ethiopian summer rainfall and sea surface temperature: part II. Seasonal forecasting. *Climate Dynamics,* 37**,** 121-131.

DIRO, G. T., GRIMES, D. I. F., BLACK, E., O'NEILL, A. & PARDO-IGUZQUIZA, E. 2009. Evaluation of reanalysis rainfall estimates over Ethiopia. *International Journal of Climatology,* 29**,** 67-78.

ELAGIB, N. A. & ELHAG, M. M. 2011. Major climate indicators of ongoing drought in Sudan. *Journal of Hydrology,* 409**,** 612-625.

GERLITZ, L., VOROGUSHYN, S., APEL, H., GAFUROV, A., UNGER-SHAYESTEH, K. & MERZ, B. 2016. A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to central and south Asia. *Hydrol. Earth Syst. Sci.,* 20**,** 4605-4623.

GISSILA, T., BLACK, E., GRIMES, D. I. F. & SLINGO, J. M. 2004. Seasonal forecasting of the Ethiopian summer rains. *International Journal of Climatology,* 24**,** 1345-1358.

GODDARD, L. & GRAHAM, N. E. 1999. Importance of the Indian Ocean for simulating rainfall anomalies over eastern and southern Africa. *Journal of Geophysical Research: Atmospheres (1984–2012),* 104**,** 19099-19116.

HAMMER, G. L., NICHOLLS, N. & MITCHELL, C. 2000. *Applications of seasonal climate forecasting in agricultural and natural ecosystems*, Springer Science & Business Media.

HERTIG, E. & JACOBEIT, J. 2011. Predictability of Mediterranean climate variables from oceanic variability. Part I: Sea surface temperature regimes. *Climate Dynamics,* 36**,** 811-823.

INES, A. V. M. & HANSEN, J. W. 2006. Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and Forest Meteorology,* 138**,** 44-53.

JOLLIFFE, I. 2002. *Principal component analysis*, Wiley Online Library.

KALNAY, E., KANAMITSU, M., KISTLER, R., COLLINS, W., DEAVEN, D., GANDIN, L., IREDELL, M., SAHA, S., WHITE, G. & WOOLLEN, J. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society,* 77**,** 437-471.

KASSAHUN, B. 1987. Weather systems over Ethiopia. *Proc. First Tech. Conf. on Meteorological Research in Eastern and Southern Africa.* Nairobi, Kenya: UCAR.

KIRTMAN, B. P., MIN, D., INFANTI, J. M., KINTER, J. L., PAOLINO, D. A., ZHANG, Q., VAN DEN DOOL, H., SAHA, S., MENDEZ, M. P., BECKER, E., PENG, P., TRIPP, P., HUANG, J., DEWITT, D. G., TIPPETT, M. K., BARNSTON, A. G., LI, S., ROSATI, A., SCHUBERT, S. D., RIENECKER, M., SUAREZ, M., LI, Z. E., MARSHAK, J., LIM, Y.-K., TRIBBIA, J., PEGION, K., MERRYFIELD, W. J., DENIS, B. & WOOD, E. F. 2014. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society,* 95**,** 585-601.

KORECHA, D. & BARNSTON, A. G. 2007. Predictability of June–September Rainfall in Ethiopia. *Monthly Weather Review,* 135**,** 628-650.

KORECHA, D. & SORTEBERG, A. 2013. Validation of operational seasonal rainfall forecast in Ethiopia. *Water Resources Research,* 49**,** 7681-7697.

LANDMAN, W. A. & MASON, S. J. 1999. Operational long-lead prediction of South African rainfall using canonical correlation analysis. *International Journal of Climatology,* 19**,** 1073-1090.

LATIF, M., DOMMENGET, D., DIMA, M. & GRÖTZNER, A. 1999. The role of Indian Ocean sea surface temperature in forcing east African rainfall anomalies during December-January 1997/98. *Journal of Climate,* 12**,** 3497-3504.

LIM, E.-P., HENDON, H. H., HUDSON, D., WANG, G. & ALVES, O. 2009. Dynamical Forecast of Inter–El Niño Variations of Tropical SST and Australian Spring Rainfall. *Monthly Weather Review,* 137**,** 3796-3810.

MANNING, C. D., RAGHAVAN, P. & SCHÜTZE, H. 2008. *Introduction to information retrieval*, Cambridge university press Cambridge.

MASON, S. 1998. Seasonal forecasting of South African rainfall using a non–linear discriminant analysis model. *International Journal of Climatology,* 18**,** 147-164.

METZGER, S., LATIF, M. & FRAEDRICH, K. 2004. Combining ENSO Forecasts: A Feasibility Study. *Monthly Weather Review,* 132**,** 456-472.

MUTAI, C. C., WARD, M. N. & COLMAN, A. W. 1998. Towards the prediction of the East Africa short rains based on sea-surface temperature–atmosphere coupling. *International Journal of Climatology,* 18**,** 975-997.

NMSA 1996. Climate and agroclimatic resources of Ethiopia. *NMSA Meteorological Research Report Series.* Addis Ababa, Ethiopia: National Meteorological Services Agency of Ethiopia.

OMONDI, P., OGALLO, L. A., ANYAH, R., MUTHAMA, J. M. & ININDA, J. 2013. Linkages between global sea surface temperatures and decadal rainfall variability over Eastern Africa region. *International Journal of Climatology,* 33**,** 2082-2104.

PALMER, T., ALESSANDRI, A., ANDERSEN, U. & CANTELAUBE, P. 2004. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society,* 85**,** 853.

PARTHASARATHY, B., KUMAR, K. R. & MUNOT, A. A. 1993. Homogeneous Indian Monsoon rainfall: Variability and prediction. *Proceedings of the Indian Academy of Sciences - Earth and Planetary Sciences,* 102**,** 121-155.

RAFTERY, A. E., MADIGAN, D. & HOETING, J. A. 1997. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association,* 92**,** 179-191.

ROECKNER, E., OBERHUBER, J. M., BACHER, A., CHRISTOPH, M. & KIRCHNER, I. 1996. ENSO variability and atmospheric response in a global coupled atmosphere-ocean GCM. *Climate Dynamics,* 12**,** 737-754.

SAHA, S., NADIGA, S., THIAW, C., WANG, J., WANG, W., ZHANG, Q., VAN DEN DOOL, H. M., PAN, H. L., MOORTHI, S., BEHRINGER, D., STOKES, D., PEÑA, M., LORD, S., WHITE, G., EBISUZAKI, W., PENG, P. & XIE, P. 2006. The NCEP Climate Forecast System. *Journal of Climate,* 19**,** 3483-3517.

SCHEPEN, A., WANG, Q. J. & ROBERTSON, D. E. 2012. Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *Journal of Geophysical Research: Atmospheres,* 117**,** n/a-n/a.

SEGELE, Z. T. & LAMB, P. J. 2005. Characterization and variability of Kiremt rainy season over Ethiopia. *Meteorology and Atmospheric Physics,* 89**,** 153-180.

SEGELE, Z. T., RICHMAN, M. B., LESLIE, L. M. & LAMB, P. J. 2015. Seasonal-to-Interannual Variability of Ethiopia/Horn of Africa Monsoon. Part II: Statistical Multi-Model Ensemble Rainfall Predictions. *Journal of Climate***,** 150129124820009.

SHANKO, D. & CAMBERLIN, P. 1998. The effects of the Southwest Indian Ocean tropical cyclones on Ethiopian drought. *International Journal of Climatology,* 18**,** 1373-1388.

SHUKLA, S., ROBERTS, J., HOELL, A., FUNK, C. C., ROBERTSON, F. & KIRTMAN, B. 2016. Assessing North American multimodel ensemble (NMME) seasonal forecast skill to assist in the early warning of anomalous hydrometeorological events over East Africa. *Climate Dynamics***,** 1-17.

SINGH, A., KULKARNI, M. A., MOHANTY, U. C., KAR, S. C., ROBERTSON, A. W. & MISHRA, G. 2012. Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model products. *Meteorological Applications,* 19**,** 179-188.

STONE, R. C., HAMMER, G. L. & MARCUSSEN, T. 1996. Prediction of global rainfall probabilities using phases of the Southern Oscillation Index. *Nature,* 384**,** 252-255.

SUÁREZ-MORENO, R. & RODRÍGUEZ-FONSECA, B. 2015. S⁴CAST v2.0: sea surface temperature based statistical seasonal forecast model. *Geosci. Model Dev.,* 8**,** 3639-3658.

TADESSE, T. 1994. The influence of the Arabian Sea storms/ depressions over the Ethiopian weather. *Proc. Int. Conf. on Monsoon Variability and Prediction.* Geneva, Switzerland: World Meteorological Organization.

TEUTSCHBEIN, C. & SEIBERT, J. 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology,* 456–457**,** 12-29.

WILKS, D. S. 2011. *Statistical methods in the atmospheric sciences*, Academic press.

ZHANG, Y., MOGES, S. & BLOCK, P. 2016. Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial-Temporal Variability: Application to Western Ethiopia. *Journal of Climate,* 29**,** 3697-3717.