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 5):

“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. ”

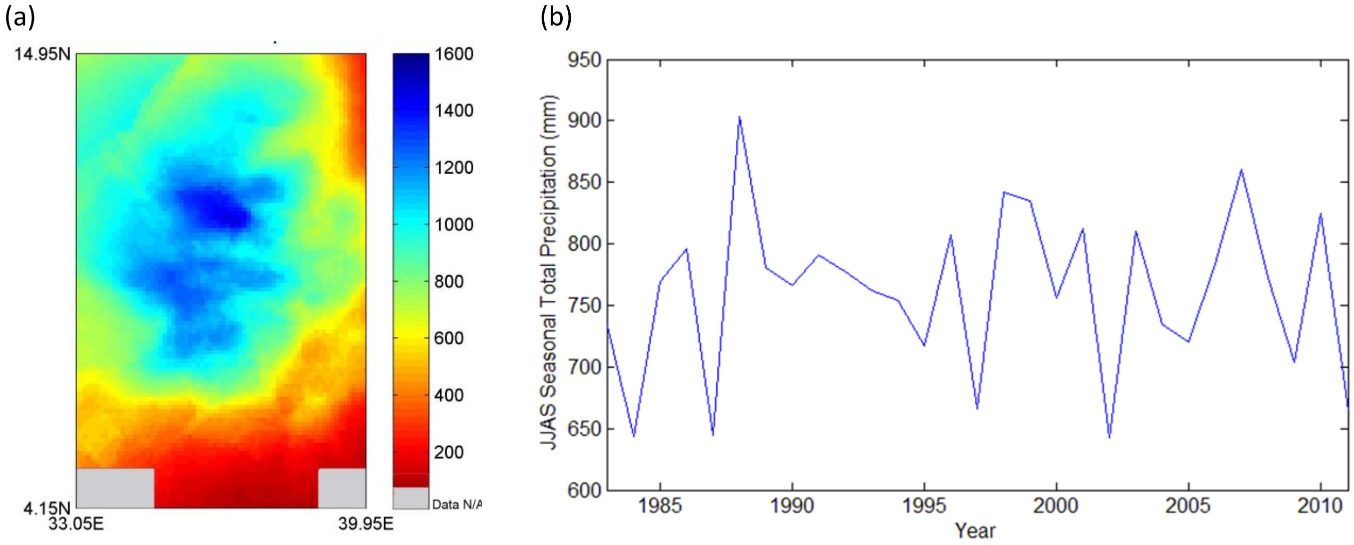

Figure 1: Spatial and temporal variability of June-September seasonal total precipitation in western Ethiopia: (a) spatial pattern of temporal-average, and (b) spatial-average time series.

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 10):

"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. A detailed analysis including a complete correlation table and unique patterns for each cluster-level time series associated with large climate variables is provided in Zhang et al. (2016), which readers are referred to for more details."

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 will be provided based on a combination of the previous concurrent connection to large-climate variable analysis (Zhang et al., 2016) and the prediction results.

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 5 Line 15) and also included here:

“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 statistical modeling process:

(1) Grid-cells within each justifiable region (e.g. equatorial Pacific; Fig. 2) with correlation above the 99% significance level are identified (Fig. 3).
(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).
(5) The top principal components (PCs) from the PCA with 95% variance explained are used as predictors … (the following steps are the same as in the original manuscript)

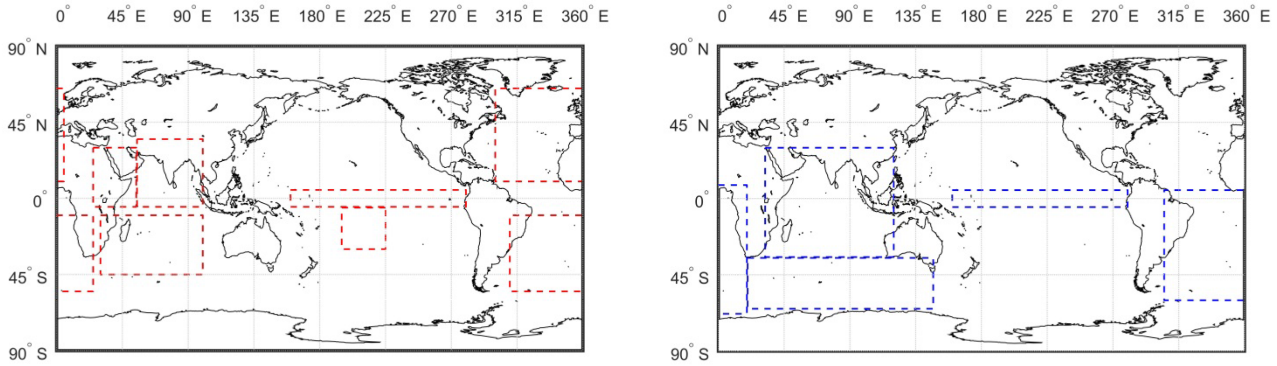

Figure 2: Justifiable climate regions globally for selecting predictors: (a) For SLP and GH at 500 mb with regions including EP, ES, LO, AH, SH, MH, and AM. For SAT, only LO is included. (b) For SST with regions including EP, NI, SI, and AT. *Note: EP - equatorial Pacific region, ES – Tahiti island for 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.*

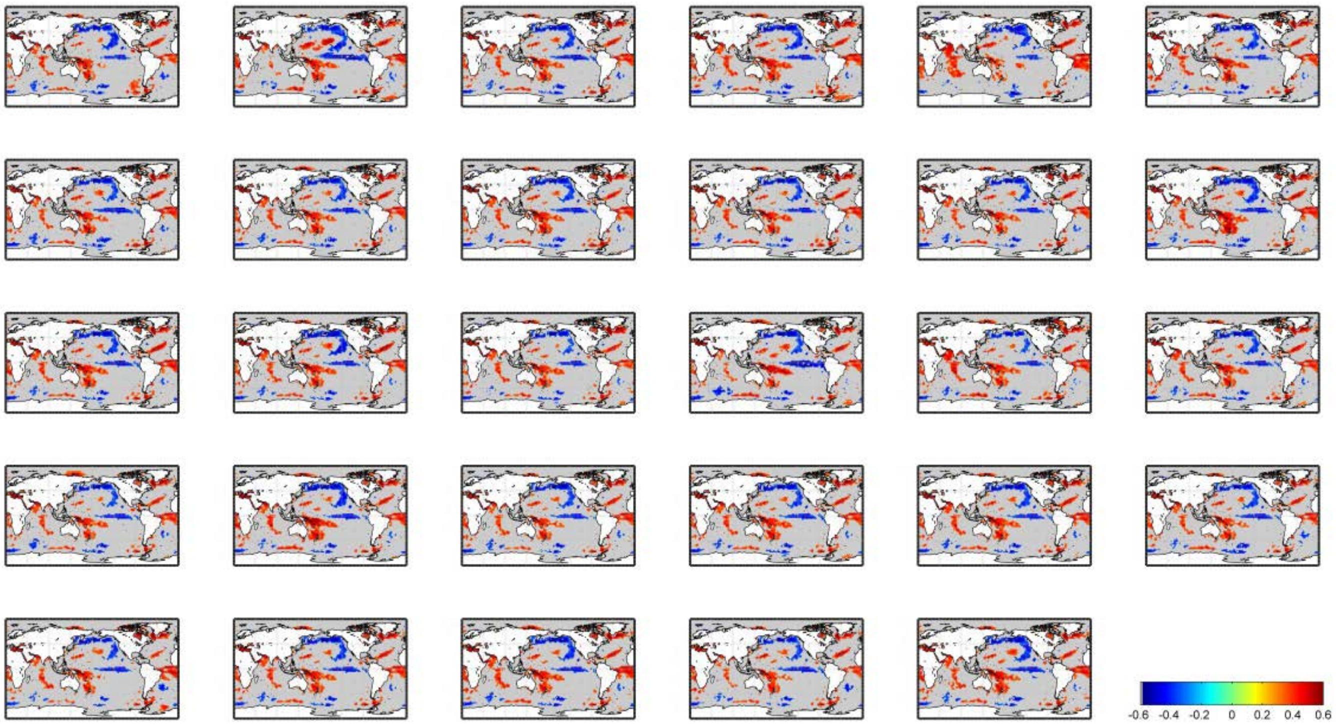

Figure 3: Correlation map between mean JJAS seasonal precipitation time series in Cluster 5 and global SST under cross-validation, with correlations lower than 99% significance level masked out (one-tail test).

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 1). However, we do see improvement over non-cluster scenarios for some of the clusters.

Table 1: 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 2: 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 1 above is included in the revised manuscript together with other updated tables, figures and result analysis, which are also included here:

“Correlations between cluster-level model predictions and observations range from -0.180 to 0.504, with Cluster 5 having the highest correlation and Cluster 6 the lowest (Table 1). In approximately 1/5 of the 29 years, the observation falls outside the prediction envelope (Fig. 4), 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 1). 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. 1 in original manuscript), performs best, with correlation and RPSS values of 0.5 and 27.4%, respectively. ”

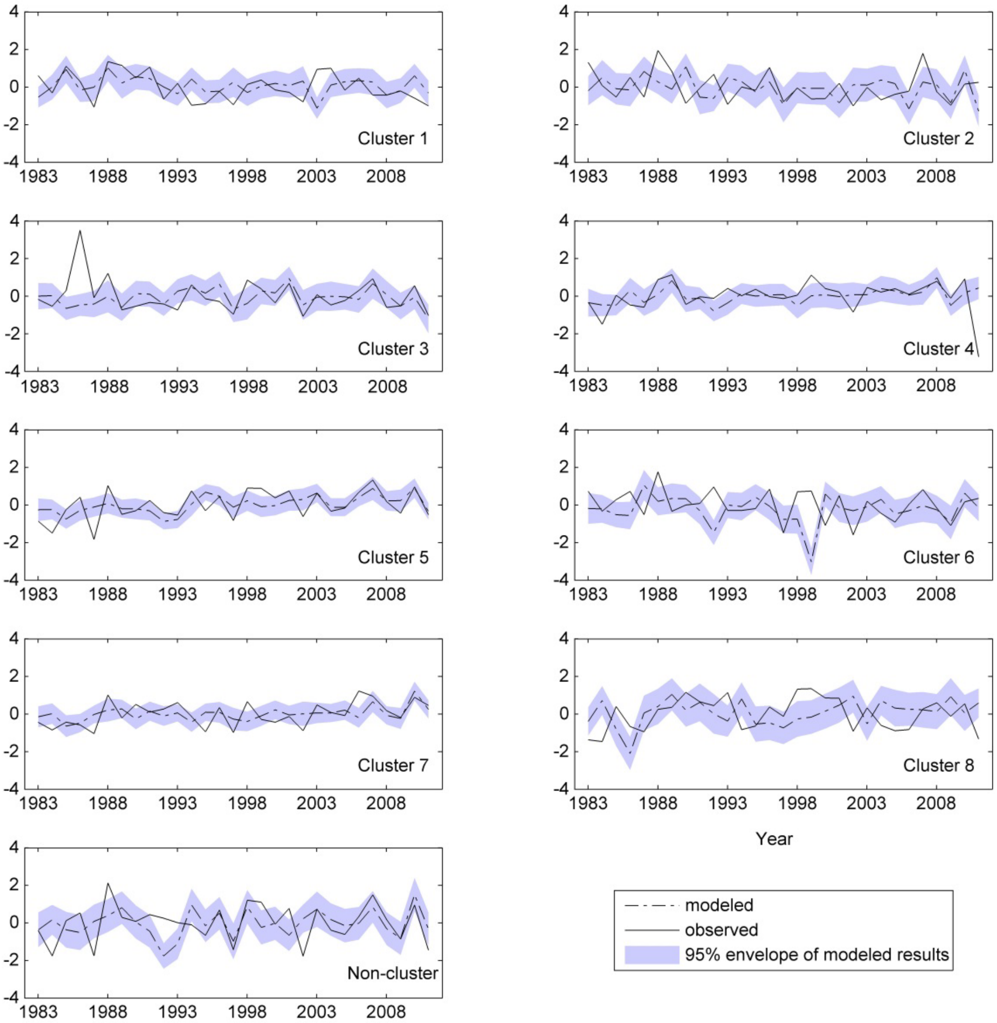

Figure 4: Cluster-level predictions and observations under C-I and NC-I scenario, with drop-one-year cross-validation. The 95% envelope shows the 95% confidence interval constructed using model errors.

“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. 5). 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. 5). 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 of 0.6, such as central-northwestern Ethiopia, which is consistent with the region of high cluster-level prediction skill (Cluster 5). The percentage of grid-cells with correlations passing the 95% significance test is the highest for the NC-D case (Table 3); however, if only comparing within a clustered region, skills can be higher for the *clustered* case."

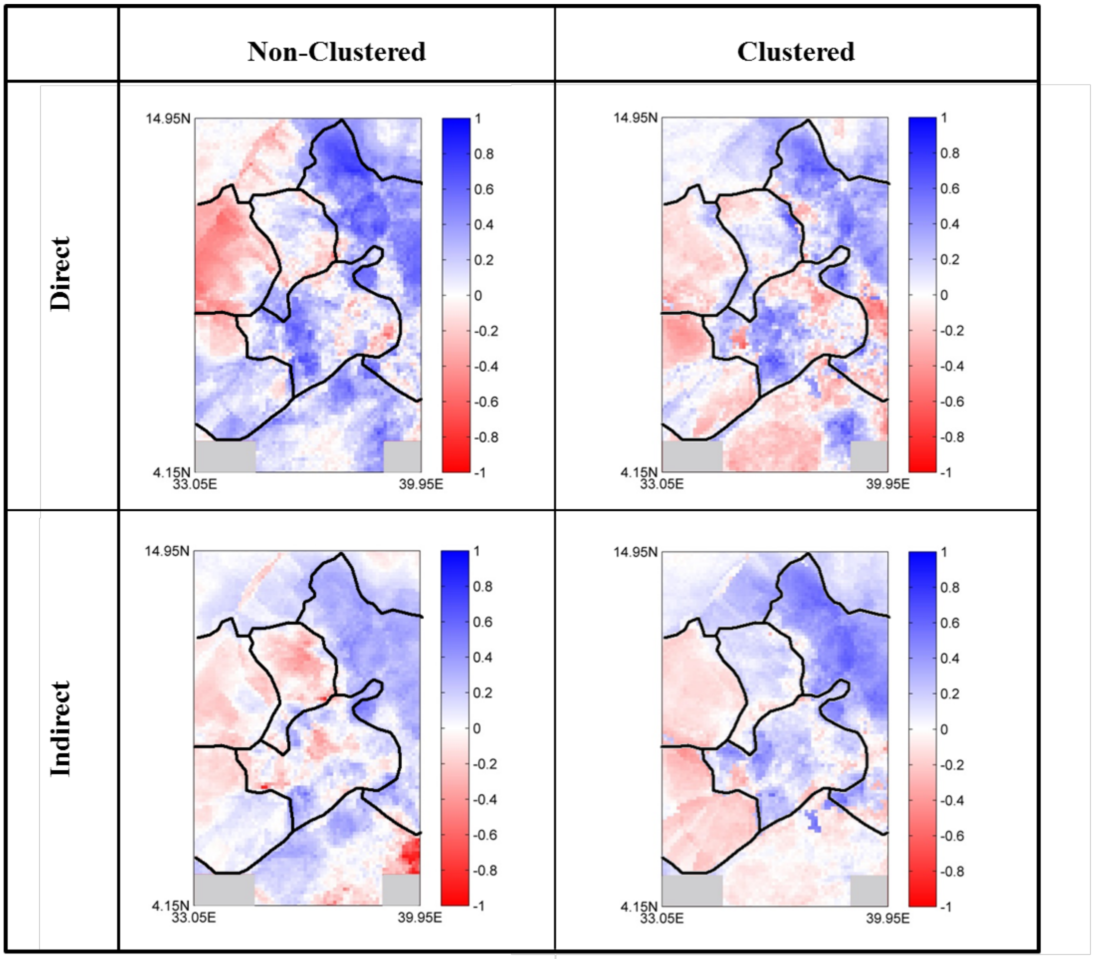

Figure 5: Pearson correlations between grid-level observations and predictions under four scenarios, with the clustering boundary delineated roughly in black.

"Similar findings are evident by evaluating the RPSS. The predictions are most skillful for the same region of central-northwestern Ethiopia (Cluster 5; Fig. 6). The percentage of grid-cells with positive RPSS values is the highest for the C-I case (Table 4), indicating the *clustered indirect* case is superior in terms of RPSS metrics."

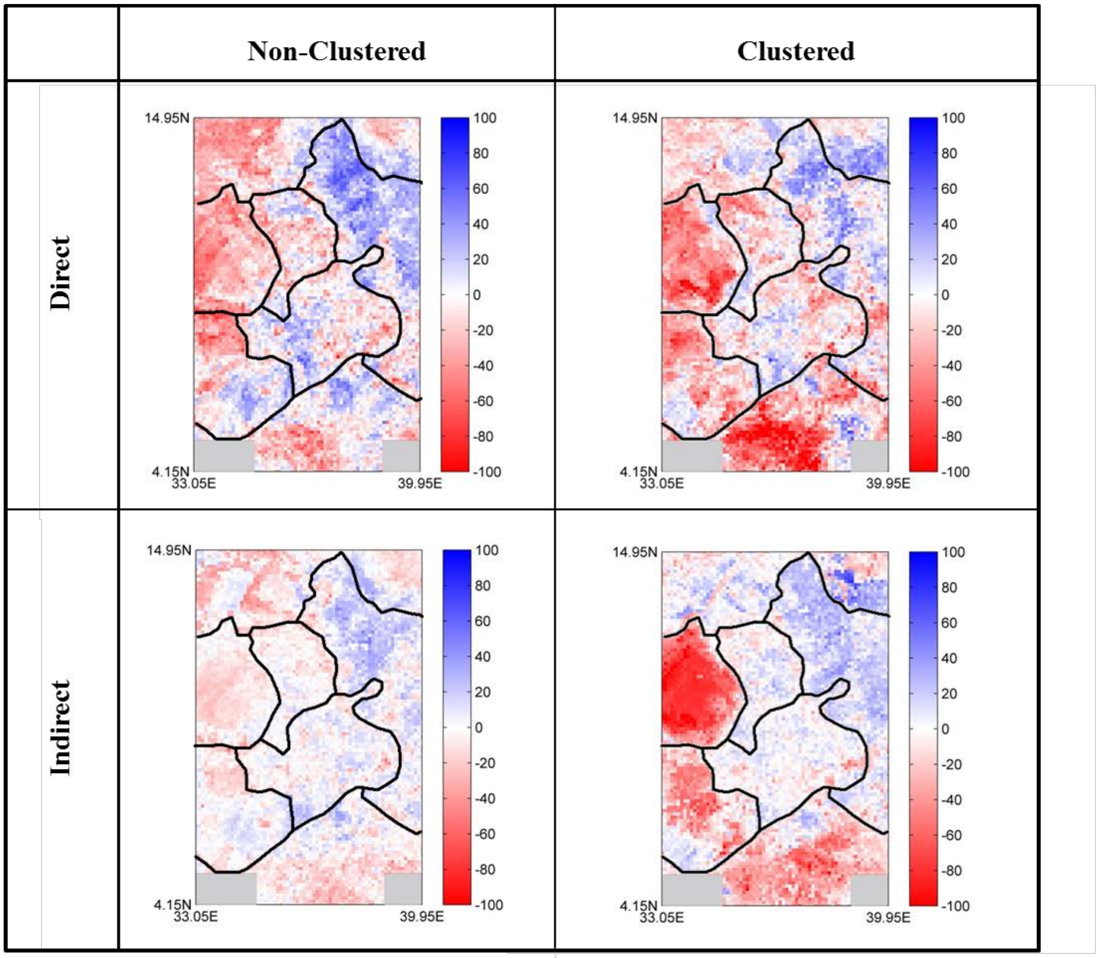

Figure 6: Grid-level RPSS (%) under four scenarios using climate variables as predictors, with the clustering boundary delineated roughly in black.

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	26.97	42.8%
NC-I	0.063	0.186	3.1%	-2.26	14.62	43.9%
C-D	0.055	0.230	10.6%	-13.97	30.97	33.9%
C-I	0.081	0.206	12.4%	-9.60	29.39	44.4%
Dynamical Model						
(9) NASA-GMAO	0.3	0.149	36.10%	2.32	21.2	54.30%
(10) NCEP CFSv2	0.31	0.155	37.30%	3.66	16.61	61.00%

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 5 Line 30):

"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" to provide 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 modeling methodology and resolution, as compared with previous studies. The statistical model prediction results show improvements over non-clustered case for some clusters. Among those clusters, Cluster 5, in agriculturally rich central-northwestern Ethiopia, performs best, with correlation and RPSS values of 0.5 and 27.4%, respectively. The general skill of dynamical models over the entire study region is higher than statistical models, although dynamical models produce predictions at a lower resolution. However, for some specific clustered regions such as Cluster 5, the statistical model outperforms dynamical models for grid-level predictions producing higher correlations and RPSS at a finer 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 4 Line 20) with the following texts inserted:

"Many studies have investigated statistical models for seasonal climate prediction. The variety of those studies lies in the pre-classification of predictor or predictand regime, predictor selection process, as well as statistical methods. For example, Hertig and Jacobeit (2011a) investigate sea surface temperature (SST) regimes as potential predictors for subsequent precipitation and temperature in the Mediterranean area. 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 sufficient 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; Hertig and Jacobeit, 2011b). 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 in combination with principal component analysis (PCA), yet differentiates from other studies by applying predictions to pre-defined homogeneous predictand 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 2 and 3 in the responses above.

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 10. 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:

"The North American Multi-Model Ensemble (NMME) (Kirtman et al., 2014) is an experimental multi-model seasonal forecasting system consisting of dynamical coupled models from various modeling centers in the 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 provided additional result analysis with comparison between them.

"The general skill of dynamical models over the entire study region is higher than statistical models, although dynamic models produce predictions at a lower resolution. However, for some specific clustered regions such as Cluster 5, statistical models outperform dynamical models with grid-level predictions demonstrating higher correlations and RPSS at a higher resolution."

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 7 Line 14. 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., and Rajagopalan, B.: Interannual Variability and Ensemble Forecast of Upper Blue Nile Basin Kiremt Season Precipitation, J. Hydrometeor, 8, 327-343, http://dx.doi.org/10.1175/JHM580.1, 2007.

Camberlin, P., and Philippon, N.: The East African March–May Rainy Season: Associated Atmospheric Dynamics and Predictability over the 1968–97 Period, Journal of Climate, 15, 1002-1019, 10.1175/1520-0442(2002)015<1002:TEAMMR>2.0.CO;2, 2002.

Diro, G. T., Black, E., and Grimes, D. I. F.: Seasonal forecasting of Ethiopian spring rains, Meteorological Applications, 15, 73-83, 10.1002/met.63, 2008.

Diro, G. T., Grimes, D. I. F., and Black, E.: Teleconnections between Ethiopian summer rainfall and sea surface temperature: part II. Seasonal forecasting, Climate Dynamics, 37, 121-131, 10.1007/s00382-010-0896-x, 2011.

Gerlitz, L., Vorogushyn, S., Apel, H., Gafurov, A., Unger-Shayesteh, K., and Merz, B.: 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, 10.5194/hess-20-4605-2016, 2016.

Gissila, T., Black, E., Grimes, D. I. F., and Slingo, J. M.: Seasonal forecasting of the Ethiopian summer rains, International Journal of Climatology, 24, 1345-1358, 10.1002/joc.1078, 2004.

Hertig, E., and Jacobeit, J.: Predictability of Mediterranean climate variables from oceanic variability. Part I: Sea surface temperature regimes, Climate Dynamics, 36, 811-823, 10.1007/s00382-010-0819-x, 2011a.

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, Climate Dynamics, 36, 825-843, 10.1007/s00382-010-0821-3, 2011b.

Jolliffe, I.: Principal component analysis, Wiley Online Library, 2002.

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., and Wood, E. F.: 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, 10.1175/BAMS-D-12-00050.1, 2014.

Korecha, D., and Barnston, A. G.: Predictability of June–September Rainfall in Ethiopia, Monthly Weather Review, 135, 628-650, 10.1175/mwr3304.1, 2007.

Omondi, P., Ogallo, L. A., Anyah, R., Muthama, J. M., and Ininda, J.: Linkages between global sea surface temperatures and decadal rainfall variability over Eastern Africa region, International Journal of Climatology, 33, 2082-2104, 10.1002/joc.3578, 2013.

Segele, Z. T., Richman, M. B., Leslie, L. M., and Lamb, P. J.: Seasonal-to-Interannual Variability of Ethiopia/Horn of Africa Monsoon. Part II: Statistical Multi-Model Ensemble Rainfall Predictions, Journal of Climate, 150129124820009, 10.1175/jcli-d-14-00476.1, 2015.

Suárez-Moreno, R., and Rodríguez-Fonseca, B.: S<sup>4</sup>CAST v2.0: sea surface temperature based statistical seasonal forecast model, Geosci. Model Dev., 8, 3639-3658, 10.5194/gmd-8-3639-2015, 2015.

von Storch, H., and Zwiers, F. W.: Statistical analysis in climate research, Cambridge University Press, Cambridge, 1999.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic press, 2011.

Zhang, Y., Moges, S., and Block, P.: 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, 10.1175/Jcli-D-15-0582.1, 2016.