ACPD Detailed response to referees

We thank the reviewers for their comments. These comments helped us to relate our work to previous works in the field and to consider an additional learning algorithm that was designed for non-stationarity sequences. Please find below our detailed response to the referees' comments along with a summary of the changes made in the revised version.

Anonymous Referee #1

We thank reviewer #1 for a very detailed and informative review.

Strong similarities to past work

It appears the authors were unaware of prior work, but unfortunately this submission cannot be published without major revision, because both the proposed problem (using learning algorithms to compute weightings to improve the predictions of the multi-model ensemble of AOGCMs) and the type of machine learning approach ("sequential learning algorithms") were already presented in previous work.

• C. Monteleoni, G. Schmidt, and S. Saroha. Tracking Climate Models. In NASA Conference on Intelligent Data Understanding (CIDU), pages 1–15, 2010.

• C. Monteleoni, G. Schmidt, S. Saroha, and E. Asplund. Tracking Climate Models. In Journal of Statistical Analysis and Data Mining: Special Issue: Best of CIDU 2010. Volume 4, Issue 4, pages 72–392, August 2011.

• S. McQuade and C. Monteleoni. MRF-Based Spatial Expert Tracking of the Multi-Model Ensemble. In New Approaches for Pattern Recognition and Change Detection, session at American Geophysical Union (AGU) Fall Meeting, 2013.

The first 2 are the most similar to the present submission; the 3rd is also highly related and was presented at AGU.

We thank the referee for pointing out these important works.

In the revised version, we cite and discuss these works (the published works that we could read and refer to) and the differences between them and our work.

Moreover, while the stated task is identical to the first two works listed above, i.e. to use "sequential learning algorithms" (also known as "online learning with expert advice"), to combine the temperature projections of a CMIP ensemble (CMIP3 was used in the 2010 paper), the machine learning methods from the 2010 paper are more recent than the ones in the present submission and essentially subsume the methods used here.

To see this, note that while the authors cite a textbook to refer to the algorithms (Cesa-Bianchi and Lugosi 2006, which we will refer to as CB06), the actual algorithms used are from the 1990's and earlier.

The algorithm denoted as EWA in the submission is originally due to:

• N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In IEEE Symposium on Foundations of Computer Science, pages 256–261, 1989.

• V. Vovk. Aggregating strategies. In Proc. 3rd Annu. Workshop on Comput. Learning Theory, pages 371–383. Morgan Kaufmann, 1990.

The algorithm denoted as EGA in the submission is originally due to:

• Jyrki Kivinen and Manfred K. Warmuth: Exponentiated gradient versus gradient descent for linear predictors. Information and Computation 132(1):1-63, 1997.

In contrast, the algorithm in the 2010 paper (which we will hereby refer to as M10) is from the early 2000's:

• C. Monteleoni and T. Jaakkola. Online Learning of Non-stationary Sequences. In Advances in Neural Information Processing Systems 16. pages 1093–1100, 2003.

In the intervening time, the field of learning algorithms had advanced significantly (and subsequently it has advanced even further). In particular, relevant to the study of climate change and the problem in question, online learning algorithms were designed expressly for non-stationary data. The algorithms applied in the present submission were not designed to handle non-stationary data. Their performance guarantees (regret bounds) are with respect to the best fixed expert. However many climate scientists have shown that in any CMIP ensemble, some climate models perform better at some times and others at other times. An influential paper on online learning algorithms (extending algorithms with exponentiated weight updates for a variety of prediction loss functions) provided modified weight updates that allow for "switching" among "best" experts:

• M. Herbster and M. K. Warmuth, Tracking the best expert, Mach Learn 32 (1998), 151 – 178.

The algorithm used in M10 was a further extension from that work which actually learns the switching rate online (sequentially), simultaneous to updating the weights over the experts. It can still track the best fixed-expert as a special case, and in that sense it is significantly more flexible than the algorithms used in the present submission.

In our field, it is common to cite a book (or a review article) that summarizes previous results.
Note that the application of the EWA and EGA in our work involves optimization of the learning rate. In order to compare the results of these algorithms to the more recent LAA, we considered the LAA in the revised version, and we present the results of all the learning schemes for an easy comparison of their performances.

In the revised version, we added citations of the original works related to the EWA, EGA, fixed-share, and LAA . In addition, we added the results of the LAA to Figures 3 and 5 and changed the Results and Discussion sections to reflect these changes and additions.

More recently, there have been several more machine learning approaches to the multi-model ensemble prediction problem proposed in the present submission. Note that the present submission treats geospatial locations separately. This was also the case in M10, but subsequently, they did an extension to explicitly model geospatial neighborhood influence in the online learning weight updates. This is another highly related work to the present submission.

• S. McQuade and C. Monteleoni. Global Climate Model Tracking using Geospatial Neighborhoods. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Special Track on Computational Sustainability and AI, pp. 335–341, 2012.

The above was on a regular grid over latitude and longitude. However there has also been some recent work on learning the geospatial structure (need not be a grid) from the data such that regional features such as mountain ranges may cause some geospatially neighboring regions to have lesser spatial influence on one another.

• A. R. Goncalves, P. Das, S. Chatterjee, V. Sivakumar, and A. Banerjee. Global climate models combination using Multitask Sparse Structure Learning. Fourth Workshop on Understanding Climate Change Through Data, At National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA, 2014.

The above is not a "sequential learning" technique. Additionally, there have been other kinds of machine learning algorithms applied to the problem, e.g.

• M. Ghafarianzadeh and C. Monteleoni. Climate Prediction via Matrix Completion. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Late-Breaking Papers Track, 2013.

Geospatial correlations do exist in the climate system and accounting for these correlations have the potential to improve future climate predictions as discussed by the referee. However, this is beyond the scope of the current work. Moreover, the results show a relatively smooth spatial distribution of the RMSE and uncertainty, which suggests that the geospatial correlations are well captured by the models in the ensemble.

Note, when the authors discuss using methods that predict "according to the best model," they should see the related literature on learning algorithms. (In CB06, see Follow the Leader and Follow the Perturbed Leader).

The meaning of the "best" model is explicitly explained in the text. It is important to note that we only used this type of prediction to show that the predictions of the SLAs outperform those of the models that obtain the highest weights.

In summary, the main problem with the present submission is its very strong similarity to previous work, along with a significant gap in the discussion of related work on using learning algorithms in climate predictions, in particular for the problem in question.

The works mentioned by the referee analyzed long-term climate predictions, most of them based on the CMIP3. In our work, we focused only on decadal climate predictions. These are different simulations initialized with real observed conditions unlike the long-term climate predictions. These experiments were first introduced in the CMIP5. The introduction of the paper elaborates on the differences between the two types of predictions.

Our results show that for decadal climate predictions, adding the climatology as an additional expert in the ensemble significantly improves the predictions. For obvious reasons, adding the climatology to long-term climate projections is not expected to have the same effect.

The revised versions of Figures 3 and 5 (and the globally averaged results provided in the text) show that the LAA performs similarly to the EWA. However, the EGA outperforms both the EWA and the LAA. The similar performance of the EWA and the LAA suggests that for decadal climate predictions, the non-stationarity nature of the dynamics is not significant. None of the previous works presented such a comparison between different SLAs. All these details are discussed in the revised version.

Previous works focused on improving the predictions using ensembles of models and learning algorithms. Here, we show that SLAs not only improve the predictions but they also reduce the associated uncertainties. The uncertainties in future climate predictions are crucial for evaluating the predictions and, clearly, for any practical use of the results.

In the revised manuscript, we attempt to better emphasize these differences from previous works.

Technical corrections

There's a mistake in the equations. While it is more standard to state the weights in terms of the losses, L, than the regret, R, the authors chose to use R, and in the process a mistake was introduced into the equation for weights. When using L, there's a negative sign before the learning rate η in the exponent. However since R is defined as the loss of the expert subtracted from the loss of the learning algorithm, when using R in the exponent, there should not be negative sign. Intuitively, experts' weights should be decreased not increased, proportionally to their prediction losses. So the authors should either switch to L or else delete the negative in the exponent of Equation (3). (See pages 14, 72, and 304 of CB06).

We thank the referee for pointing out this typo.

In the revised version, we use the loss rather than the regret in order to maintain consistency with previous works.

Differences from past work, and positive aspects of the submission
On the positive side, the paper is very thoughtful and well-written. Moreover, there are also some novel experiments and ways of looking at the results, along with interesting insights from climate science. These should be emphasized in a revision.
• The input ensemble is from CMIP5 and has other differences from the ensembles used in several of the past works.
• Learning is done at a much higher spatial granularity than in M10 and some of the past works.
• The visualizations are different than M10 (although not all are different than all the past work on the problem, cf. Goncalves).
• There are some interesting findings, e.g. they point out that their results show that there is more model disagreement at polar regions.
• The authors use a variety of different metrics to evaluate their results instead of just prediction loss, e.g. RMSE (which however is closely related to the chosen prediction loss, also used in M10), as well as global, area-weighted uncertainty, and some interesting significance tests.
• The algorithm used in M10 has a per-time-iteration complexity that's on the order of a factor of T times that of the algorithm used in the present submission, where T is the desired number of temporal epochs (e.g. number of months for the whole experiment). This is for the optimized version of the algorithm in M10; one can also use lower complexity for slightly degraded results. This increased complexity is used in handling non-stationary data.
This is a difference, although not necessarily an advantage, from M10: The evaluations are done using a training and a test period. However this might not be the best technique, given that the observations and climate models exhibit non-stationarity. In such a setting, online learning algorithms are typically evaluated using progressive validation error and/or cumulative prediction loss (see M10).

We thank the referee for the positive comments.

Recommendations and suggested resources
It is very encouraging to see climate scientists using machine learning in their research. However this reviewer recommends resubmission, taking these major concerns into account. Given that there is already significant prior work in this area, the authors are strongly advised to re-think the framing of their submission, discussing past work, emphasizing the differences, and focusing much more on what is novel in their submission. Perhaps given the strong overlaps with past work, the authors will decide to run additional experiments with other learning algorithms with different properties, and compare them.

We followed the referee's advice and revised the paper to include a better review of previous works in the field, and in particular, we added the most recent learning algorithm and compared its performance to the older algorithms used in the original submission. We also now emphasize the difference between decadal climate predictions and long-term predictions, as well as the finding that on decadal time scales, the non-stationarity of the climate predictions do not seem to be crucial.

There is a field of interdisciplinary research using learning algorithms in climate science called "climate informatics" (see http://climateinformatics.org for a variety of links and a recent video tutorial surveying work in the field). There is an annual international climate informatics workshop (since 2011; the next workshop will be held be at NCAR in Colorado, in September 2015). These authors are doing work in this area and might benefit from this growing research community.

We thank the referee for informing us of this important workshop.

Anonymous Referee #2

We thank the reviewer for the helpful report.

This paper presents a machine learning-based approach for improving model-generated projections of past and future climate. The work is well motivated citing appropriate literature, and the technical approach is presented in sufficient detail.

We thank the referee for the positive comments.

Unfortunately, the authors appear to have missed some important literature in this area, most notably work by C. Monteleoni and colleagues over the past several years, see e.g.,:
C. Monteleoni, G. Schmidt, and S. Saroha. Tracking Climate Models. In NASA CIDU, 2010. S. McQuade and C. Monteleoni. Global Climate Model Tracking using Geospatial Neighborhoods. In The 2nd Int'l Workshop on Climate Informatics, 2012. C. Monteleoni, G. Schmidt, and S. McQuade, Climate Informatics: Accelerating Discovery in Climate Science with Machine Learning. IEEE Computing in Science and Engineering (CISE) Magazine, Special Issue on Machine Learning. 15(5), 32–40, 2013.

Indeed, in the previous version of the manuscript, we missed these important works. Please see our detailed response to reviewer #1 for a thorough discussion of the differences between our work and previous ones.

In the revised manuscript, the learn-α algorithm was considered and compared with the other learning algorithms.

After much deliberation, this omission demands rejection of the manuscript. In order to be reconsidered, the authors should thoroughly familiarize themselves with this and any other prior work (in both climate / Earth science and machine learning as it is an interdisciplinary contribution), re-frame the presentation of their approach in the context of the literature, and qualitatively and/or quantitatively compare their approach to existing methods where appropriate.

In the revised manuscript, we discuss previous works and the differences between them and our work. In addition, we consider the algorithm previously used and compare its performance to the other sequential learning algorithms.

List of changes:

1. We added a brief discussion of the differences between decadal climate predictions and long-term climate projection to the Introduction.

2. We added paragraphs to the Introduction discussing previous works in which SLAs were applied to an ensemble of climate models.

3. In the description of the SLAs, we now use the loss rather than the regret, and we fixed the typo pointed out by reviewer #1.

4. In the SLA section, we introduced the learn-α algorithm (LAA).

5. In sections 3-5, we added the results of the LAA for the CMIP5 ensemble that we considered. The results are compared with the results of the other SLAs.

6. Figures 3 and 5 were modified to include the results of the LAA. Their captions were changed accordingly.

7. The Discussion and Summary section was modified to include a discussion of the differences between the LAA and the other SLAs considered. In addition, we now discuss the implications of the results found for decadal climate predictions.

8. We added references to:

Herbster, M. and Warmuth, M. K.: Tracking the best expert, Mach Learn, 32, 151–178, 1998.

Kivinen, J. and Warmuth, M.: Exponentiated gradient versus gradient descent for linear predictors. Information and Computation, 132(1):1–63, 1997.

Littlestone, N. and Warmuth, M.: The weighted majority algorithm. Information and Computation, 108:212–261, 1994.

Monteleoni, C. and Jaakkola, T.: Online Learning of Non-stationary Sequences, in Advances in Neural Information Processing Systems, 16, 1093–1100, 2003.

Monteleoni, C., Saroha, S., and Schmidt, G.: Tracking Climate Models, in NASA Conference on Intelligent Data Understanding (CIDU), pages 1–15, 2010.

Monteleoni, C., Saroha, S., Schmidt, G., and Asplund, E.: Tracking Climate Models, in Journal of Statistical Analysis and Data Mining: Special Issue: Best of CIDU 2010, Volume 4, Issue 4, pages 72–392, 2011.