ACP-2015-111 – Authors responses to the second round of reviews

Dear Editor,

We thank you and the reviewers for the helpful comments. Please find below our specific responses and the details of the changes we made to the MS following these comments.

Editor comments:

Dear Authors,

Thank you for your revision. Please address points 1 and 2 of Referee #1's new report regarding the parameter tuning and the exposition and I will evaluate the suitability of the revised manuscript for publication in ACP.

Faye McNeill, co-editor ACP

In what follows, we address points 1 and 2 of Referee #1.

Comments of Referee #1:

I commend the authors for their significant additional work and revisions. I vote for acceptance subject to technical corrections and minor revisions to address the following primary concerns. Note: the technical corrections may involve redoing their experiments to address item 1.

We thank the referee for the supportive comments.

1. Parameter tuning: it is not clear whether the parameter tuning was done in a principled way.

2. Exposition needs to be revised to be more nuanced, so that it adequately reflects the observed results. Currently there are various assertions meant to explain the performance of the algorithms, that are unsupported. It would be more scientifically accurate to discuss the results without asserting such strong conclusions.

1. Parameter tuning

The authors state that the parameter eta (learning rate) used in EGA and EWA was "was set to achieve the best performance during the learning period." It is unclear whether the parameter tuning was done in a principled way. Parameter tuning needs to be done on a separate subset of the data set that is not part of the training set or test set. Alternatively, there are also techniques using cross-validation or bootstrap. The authors need be explicit about what they did, and provide details such as how many parameter values were tested, which values were ultimately chosen, and how they formed the validation data set. Also, was the alpha-initialization procedure for LAA done using the regret-optimal technique from MJ03?

For an introduction on how to tune parameters in machine learning algorithms, that addresses the need for a separate "validation set" that is not part of the training set, see for example p. 222 (and all of Chapter 7 on Model Assessment and Selection) of:
Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning: Data

Mining, Inference, and Prediction. Springer. 2009.

These issues need to be addressed in order to make sure that a fair comparison is being made between the various algorithms. This is important since the authors draw conclusions from these comparisons, such as that non-stationarity was not an issue. (Note: this is quite a strong assertion; point 2. advises toning down the exposition to avoid unsupported or weakly supported assertions.) The authors should clarify how they did parameter tuning, and may actually need to re-run the parameter tuning for the algorithms with parameters, and then update their results on the test set.

The tuning of η for the EGA and EWA was done in a principled way. In order to have a long enough validation period, we needed to use the learning period for both learning and parameter tuning. It is important to emphasize that in our work, the learning stops at the end of the training period, and the predictions for the validation period are based on the knowledge gained during the training period. We do not use the model to predict only the next outcome but to predict a time series. The parameter η was set to minimize the RMSE of the forecast during the training period. It is important to note that the EGA differs from the LAA in the fact that the parameter is constant during the entire training period (this is part of the reason why the model is more suitable for a stationary time series). The eta providing the minimal RMSE was found using a recursive search. At first, we calculated the RMSE for all the values of η from 0 to 700 with a resolution of $\Delta\eta=10$. We marked η_1 the value with the minimal RMSE. Then we calculated the RMSE for all the values of η in the range of $(\eta_1-9)-(\eta_1+9)$ with a resolution of $\Delta\eta=1$; We marked by η_2 the value with the minimal RMSE in this range and scanned around this value with a finer resolution and repeated the processes recursively till we found the optimal η with a resolution of $\Delta\eta=0.01$.

It is important to note that we also performed several tests of different methods to scan the possible values of η for the minimal RMSE, and the method described above provided the best results and was also computationally efficient. The smallest possible value of eta is trivial ($\eta=0$ implies no learning), and the highest value was set by the machine precision. However, it is important to mention that the optimal values found (for all the grid cells) were at least an order of magnitude lower than the upper limit.

We modified the text of the MS to better explain the optimization method of η. In addition, we changed the wording of the conclusions we drew from our results to be more specific to the methods we employed (see the responses to point #2).

We added the following text to provide the details of the search method for the optimal value of η:

The optimal value of η was found using a recursive search. We scanned the range $\eta\in[0,700]$ in which the lower limit represents no learning and the upper limit was set by the machine precision. However, our search never reached the upper limit and, in most grid cells, was found to be at least an order of magnitude smaller. In the first scan, we used a coarse resolution of $\Delta\eta=10$ and recursively narrowed the range to reach a resolution of $\Delta\eta=0.01$. Other methods to search for the optimal value of η provided similar results but were less efficient.

For the LAA, the regret-optimal technique of MJ03 was used (as explicitly mentioned in the description of the LAA in the MS).

2. Expository revisions needed to reduce unsubstantiated assertions

"As a successor of the EWA, the LAA also tends to converge to the best model."
This is inaccurate; the fixed-share algorithm is explicitly designed to allow switches between models, and thus need not converge to one model (convergence will depend on the level of non-stationarity in the sequence of expert prediction losses, along with the value of the alpha parameter). Neither does

Learn-alpha, so if it happened to converge to one model in their experiments, evidence should be provided.

We do not claim that the algorithm is designed to converge to the best model for any time series. Although we are not experts in online learning, we do understand the algorithms that we used, including those recommended by the referee. However, we did find that in many grid cells, the LAA converges to the best model for the data analyzed in our research. The weight assigned to the best model (after properly averaging the weights for different values of alpha) is significantly higher than the weight assigned to the other models, and in that sense, it converges.
Please see Figure 1 below which presents the weight assigned to the climatology by the EWA, LAA and EGA after the last time step of the learning period. These weights were used in the prediction period. Clearly, for the EWA, the climatology dominates the predictions over most of the globe. For the LAA, there are more regions where the climatology is not dominant (due to the switching between experts); yet, over most of the globe the climatology dominates the predictions of the LAA. For the EGA, the climatology is much less dominant as reflected in the higher uncertainty associated with the predictions of the EGA. We believe that these results should not be included in the MS in order to keep the MS accessible to climate scientists who are not experts in online learning.
For clarity, we removed this statement.

Earlier in the paper, the following should be revised, to avoid the confusion above:
"The fixed-share algorithm is a generalization of the EWA algorithm that increases the ability to switch between experts"
The terms "generalization" and "increases the ability" are misleading — the fixed-share algorithm explicitly models switches between experts, whereas EWA does not.

We changed the wording accordingly. It now reads:
The fixed-share algorithm is designed to switch between *experts* (or between climate models in our case) in response to changes in their performances.

Currently such claims appear multiple places in the paper, and at times are asserted as the explanation for certain results, e.g.
"Note that the smaller uncertainty of the EWA and the LAA forecasters is simply due to the fact that these forecasters converge to the best model in each grid cell (if the ensemble includes a model that is always the best)."
Not only is this statement inaccurate for LAA, but also it is not scientifically sound to assert a sole cause for an outcome, without significant evidence.

Please see our response to the previous points. We do have strong evidence that both the EWA and the LAA assign a very high weight to one of the models (the climatology) in most grid cells (see Figure 1). In addition, the experts are the same experts for all the learning algorithms considered in our research. Therefore, the only way to minimize the weighted variance is to assign a high weight to one expert and much lower weights to the others (or if several models are very close, they may all get high weights compared with those with very different predictions). We believe that the statement above is correct and reflects our findings.

Along these lines, this strong assertion is also unsupported:
"The small improvement achieved by the EWA and LAA, when the climatology was added as an expert to the ensemble, is only associated with the fact that they tracked the climatology in regions where it

performed the best."

This point was tested again and found to be precise.
Please see Figure 1 below.
For clarity and in order to focus the statement, we changed the wording of this sentence. It now reads:
We believe that the small improvement achieved by the EWA and LAA, when the climatology was added as an expert to the ensemble, stems from the fact that over most of the globe, the climatology dominated the predictions of these SLAs.

It seems that statistical significance was studied for EGA only, and the differences between the performance of the 3 algorithms was not tested for statistical significance. If they were tested this should be clarified; currently the following statement seems like an assertion.
"The improvement of the LAA over the EWA is too small to be considered significant."
As mentioned above, given these issues and the reduction of uncertainty by LAA, the strong assertion of lack of non-stationarity does not seem sufficiently supported. Either further evidence should be provided, e.g. reporting statistics at a higher granularity than global, or else this can be addressed by modifying the exposition.

Following the referee's comment, we tested the statistical significance of the improvement achieved by the EWA and LAA. We found that our previous statement was indeed wrong. As shown below, the small improvement is statistically significant. We changed the wording accordingly. In Figures 2 and 3, we show the same information as in Figure 8 of the MS but for the EWA and LAA predictions.
Following this analysis, we changed the wording of the text in the MS. It now reads:
The improvement, relative to the climatology and the equally weighted ensemble, achieved by the LAA and the EWA, although small, was found to be statistically significant.

Minor comments:

- "optimal discretization" could sound misleading — this is the discretization that optimizes the performance guarantees in MJ03.

We changed the wording to address this comment. It now reads:
performance-optimized discretization

- It might be helpful to use the term "online learning," to introduce these types of algorithms to this audience. “SLA" is non-standard in machine learning.

We added the term “online learning” when first introducing the term “sequential learning algorithm” (SLA).

- Did you find the result for Figure 4 to hold for LAA? With modeling non-stationarity, this might not hold.

Please see Figure 4 below which shows the RMSE vs. the learning period for the LAA. We agree that for non-stationarity data, a deviation from this behavior is expected.

To summarize, we thank the referee for the helpful and constructive criticism.

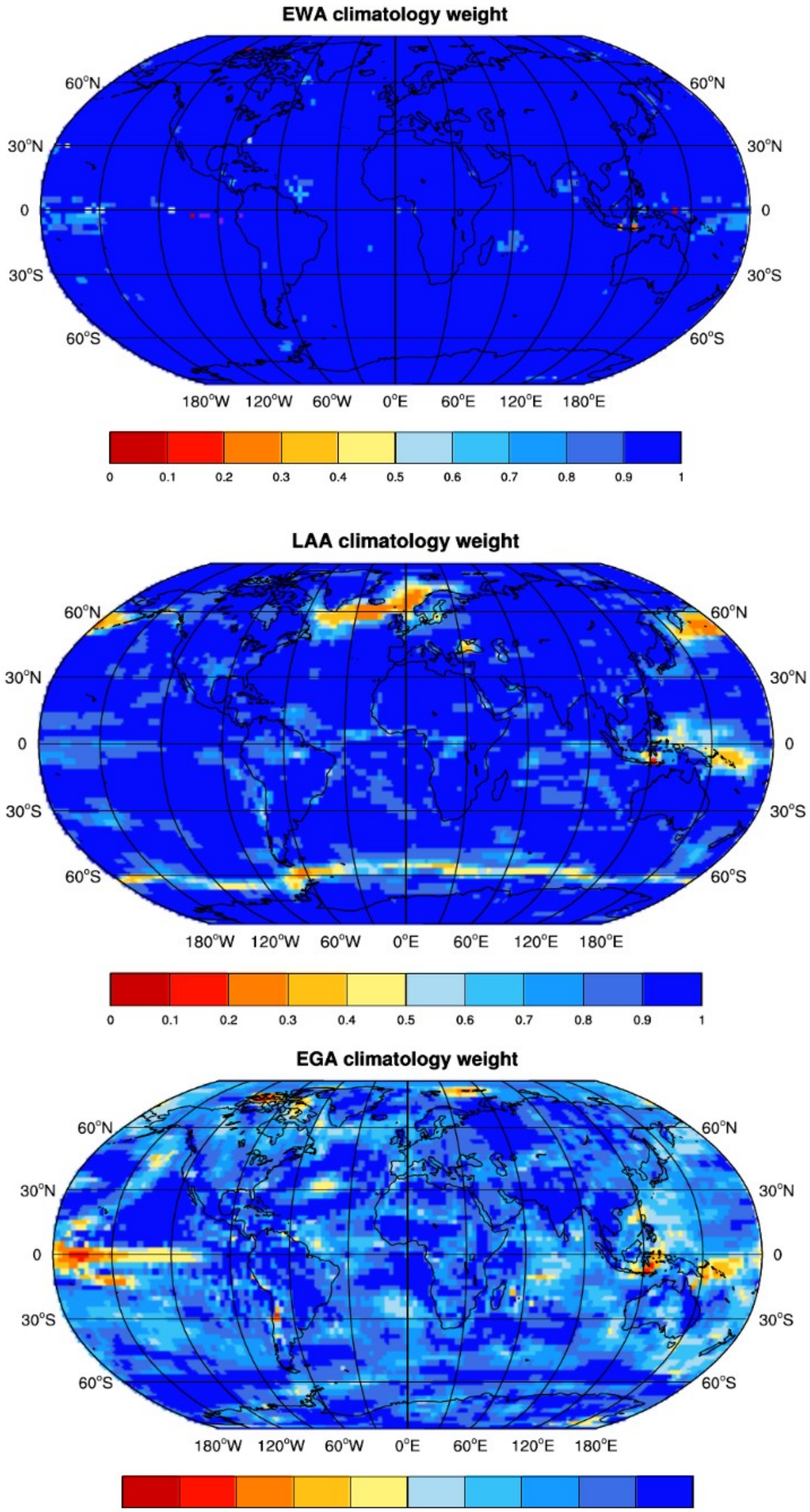

Figure 1: The spatial distribution of the weight assigned to the climatology for the prediction period by the EWA (upper panel), LAA (middle panel), and the EGA (lower panel).

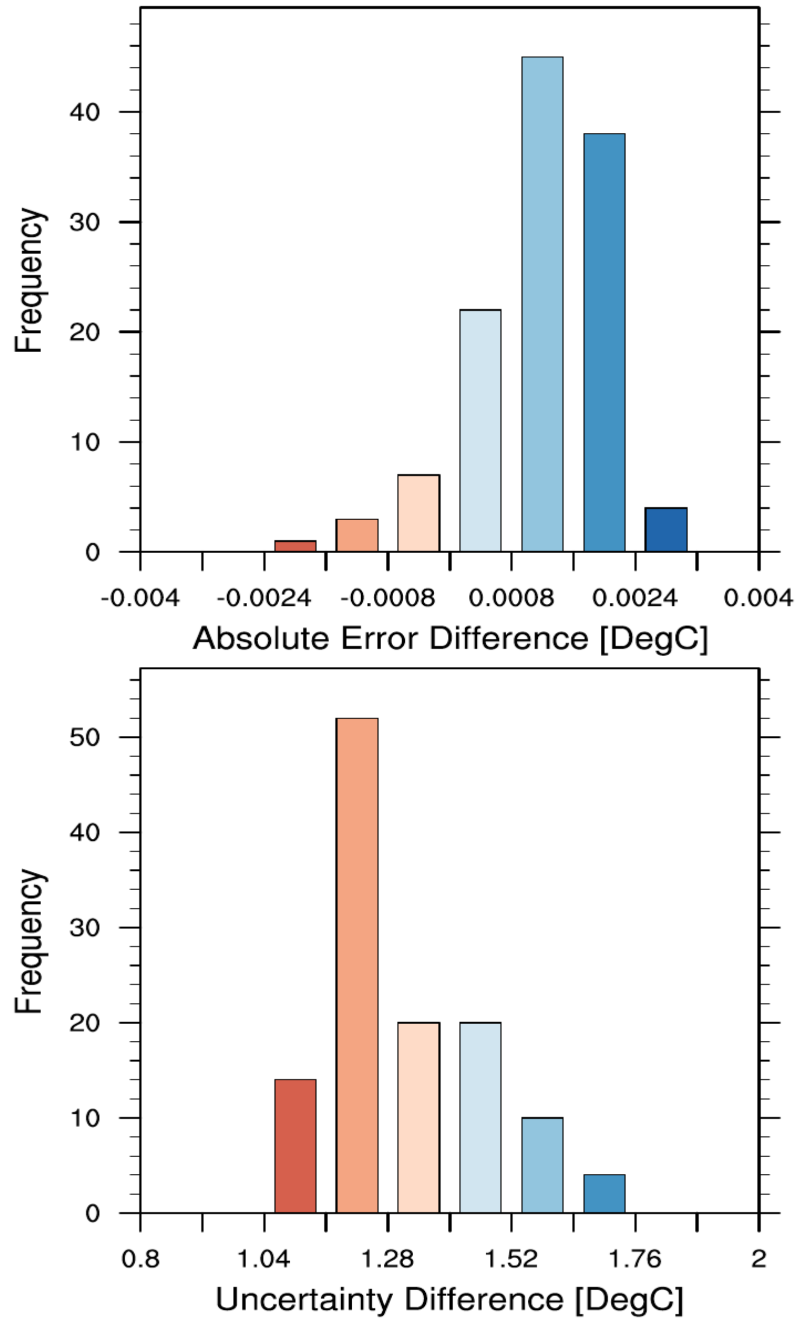

Figure 2: The histograms of the globally averaged differences of absolute error and uncertainty. The upper panel shows the histogram of the globally averaged difference between the absolute error of the climatology and that of the EWA forecaster. *The lower panel shows the histogram of the difference between the uncertainties of equally weighted and EWA weighted ensembles. Both quantities show that the small improvement of the EWA is statistically significant.*

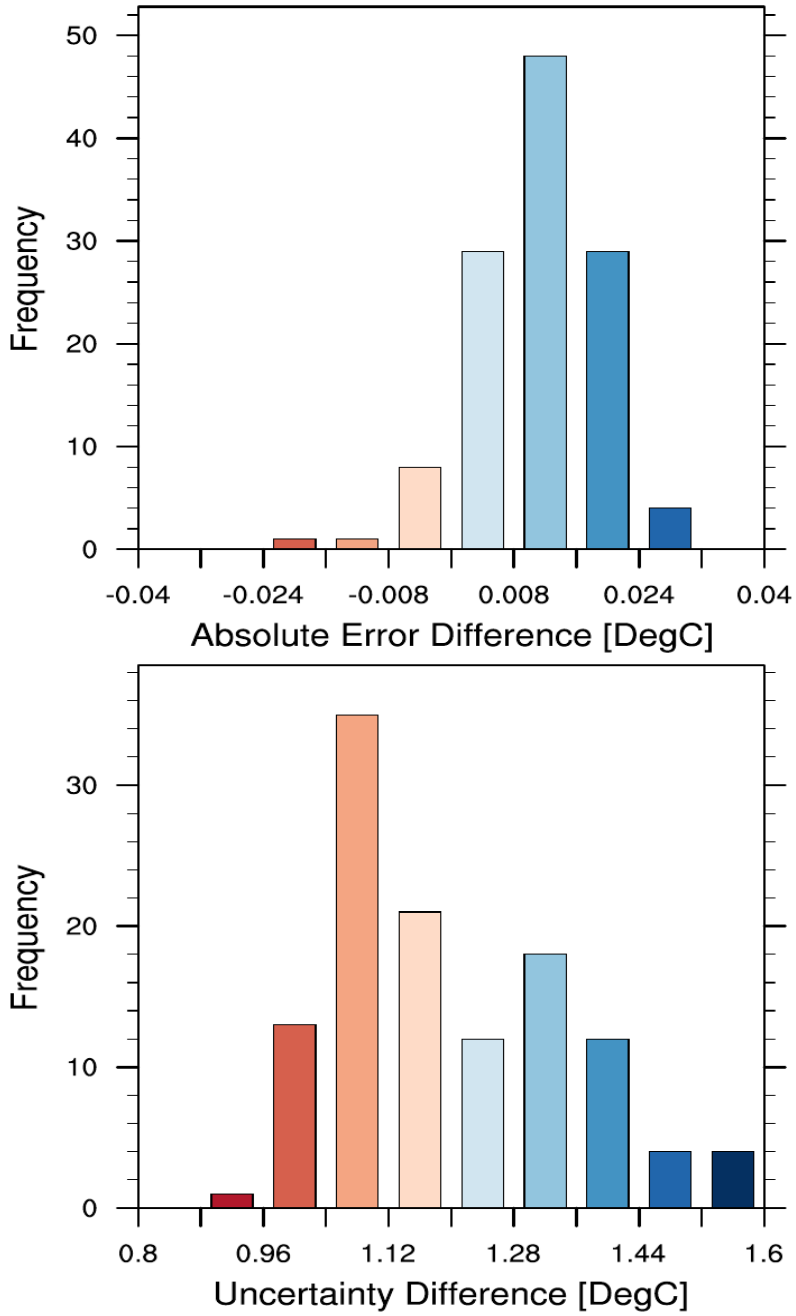

Figure 3: The histograms of the globally averaged differences of absolute error and uncertainty. The upper panel shows the histogram of the globally averaged difference between the absolute error of the climatology and that of the LAA forecaster. The lower panel shows the histogram of the difference between the uncertainties of equally weighted and LAA weighted ensembles. Both quantities show that the small improvement of the LAA is statistically significant.

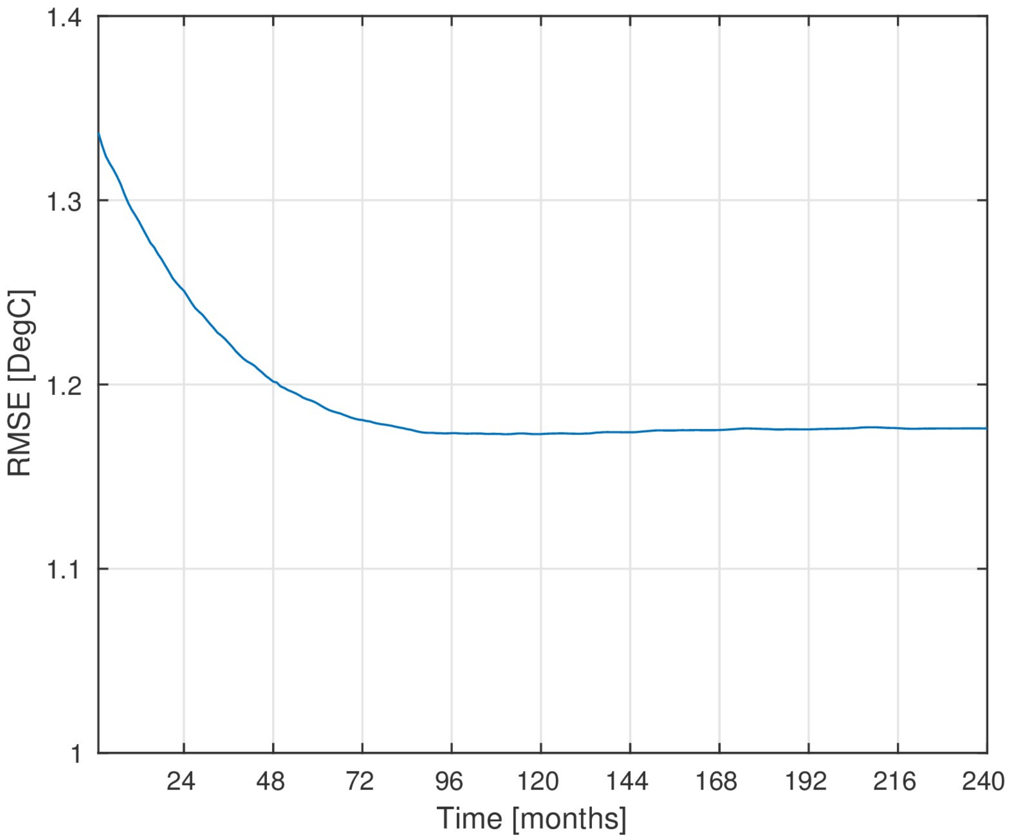

Figure 4: Global, area-weighted RMSE of the 2m-temperature, during the 10-year validation period, as a function of the learning time. The presented RMSE was calculated for the LAA forecaster. In general, a longer learning period improves the forecaster predictions, though for non-stationary data, this is not necessarily the case.