Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-49-AC2, 2019

Nonlinear Processes
in Geophysics
Discussions
Open Access

Interactive comment on "Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression" *by* Moritz N. Lang et al.

Moritz N. Lang et al.

moritz.lang@uibk.ac.at

Received and published: 10 December 2019

This work investigates the effect of different types of training periods on predictive performance of postprocessing models at different types of locations (plain, alpine foreland, alpine). The presentation is concise, the aims of the work and the used methods are presented in a clear way. Especially the graphical illustration of the different types of training periods and of the situations in the considered data situations is very helpful. This comparative study is highly relevant for applications. The approaches for constructing training data presented here are all discussed in individual papers and applied to quite different situations, even based on different types of postprocessing

models. Therefore, it is quite interesting to have a unified study of the effects of these training periods under the same conditions. However, some other settings might be included in the study, and some more details in the already presented results could be interesting, see below.

We want to thank you for your fruitful and constructive review. We have carefully addressed each of your comments and due to your suggestions we have additionally performed the comparison of the different time-adaptive training schemes for daily precipitation sums employing a non-Gaussian response distribution. According to your suggestions and the comments of reviewer 1, the most substantial changes in the manuscript are the following:

- The main goal of the article is now more clearly stated in the manuscript. The objective is to cover a wide range of methods as proposed in the literature – rather than finding the universally best method – in order to provide guidance on strengths and weaknesses of the underlying strategies. Therefore, to show a wide spectrum of possible approaches in a unified setup, we consider typical basic applications of these training schemes and refrain from more elaborate tuning or combinations. We have adjusted the introduction (Sect. 1), the conclusion (Sect. 4) and the corresponding paragraphs in the methodology (Sect. 2.2.2) accordingly.
- We have added more information on the 2016-03-08 change in the horizontal resolution of the ECMWF EPS (cycle 41r2). This specific change was chosen to construct the data sets A-C because it is likely to affect coefficient estimates more substantially. We also now clarify that, in fact, further model changes occurred in the time periods considered but that these did not affect the horizontal resolution and hence can be expected to have much smaller effects on the coefficient estimates.
- An additional comparison of the different time-adaptive training schemes has

been performed on daily precipitation sums employing a left-censored Gaussian model for post-processing. All results are very similar to the analyses for the 2 m temperature forecasts presented in the manuscript and hence nicely support the conclusions given in the paper. Therefore, we feel that it is not necessary to report these additional results in the main manuscript but we do include them in an online supplement.

Our reply to your comments can be found on the following pages.

General comment:

The presented study is only based on NR for the Gaussian case. It would be useful to include at least one other (NR) scenario with quite different behavior to see whether in a case like precipitation or wind (gust) speed the results concerning the performance of the different training data sets is the same. Both precipitation and wind speed are more heavy tailed than temperature, and there can be much more localized phenomenons on maybe sub-model-grid scales. Investigation of a non-Gaussian scenario is therefore recommended.

As the original implementation of the *regularized sliding-window* approach is based on post-processing precipitation forecasts, we have decided to present an additional analysis on post-processing daily precipitation sums. We employ the same 15 measurement sites as presented in the manuscript, and use observations and de-accumulated EPS daily precipitation sums forecasted by the ECMWF EPS. In order to remove some of the positive skewness, we follow Stauffer *et al.* (2017a) and apply a power-transformation with an ad-hoc chosen power parameter of 2 to the observations and to every ensemble member. As an appropriate response distribution for daily precipitation sums, we employ the zero left-censored Gaussian distribution (Stauffer *et al.* 2017a) with a sliding window length of 80 days.

Figure 1 and Fig. 2 illustrate the temporal evolution of the regression coefficients for

the validation period in data set A at forecast step $+24$ h for Innsbruck and Hamburg, respectively. Figure 3 provides the counterpart of Fig. 5 in the manuscript, showing CRPS skill scores with the classical *sliding-window* approach as a reference. Due to the employed power-transformation, CRPS values are computed by quantile sampling with $n = 1000$; for a more detailed description compare Stauffer *et al.* (2017b).

The results for post-processing daily precipitation sums, depicted in Fig. 1–3, can be summarized as followed:

- For both Innsbruck and Hamburg, the *sliding-window* and *regularized sliding-window* approaches show very strong fluctuations in the evolution of the regression coefficient without a clear seasonal pattern comparing the consecutive years with each other (Fig. 1 and 2).
- The coefficient paths for the *sliding-window plus* approach and the *smooth model* look comparable with quite low seasonal variation in all coefficient paths. For Hamburg, the seasonal variability in the scale parameter is slightly larger than for Innsbruck (Fig. 1 and 2).
- The *sliding-window plus* and the *smooth model* approaches show the highest improvements over the classical *sliding-window* approach with a slightly better performance of the *sliding-window plus* approach for data set C in comparison to data sets A and B (Fig. 3).

All presented results are very similar to the analyses for 2 m temperature forecasts presented in the manuscript and support the conclusions given in the paper. Hence, we suggest to include these analyses in an online supplement in addition to the manuscript.

Figure 5, possible extensions: The boxplots are aggregations of all scores over the 5 stations and over all forecast horizons. It would be interesting to see these boxplots *with values aggregated over the stations but for a specific forecast horizon only, e.g. exemplarily for 12h and 72h ahead. It could be interesting to see whether different forecast horizons affect the predictive performance in different ways – in conjunction with the situations (model change included or not) in datasets A, B, C.*

Figure 4 shows aggregated CRPS skill scores for groups of five respective stations classified as topographically plain, mountain foreland, and alpine sites regarding solely data set A but conditional on the forecast steps from $+12$ h to $+72$ h on a 12 hourly temporal resolution. As it can be seen, the variability of the predictive performance for the various setups is rather similar between different forecast steps. Exceptions are visible for the *smooth model* for stations located in the plains at forecast steps $+24$ h and $+48$ h (0000 UTC) and for stations located in the foreland at forecast steps $+36$ h and $+60$ h (1200 UTC). While the variance for the plain sites increases and the predictive performance slightly decreases, the performance for the foreland sites show the exact opposite.

As the variability between the different forecast steps is overall within a reasonable range and does not show any distinct pattern, we think that Fig. 4 provides no significantly new insights to the research question of the manuscript.

It seems that both, SW plus and the smooth model tend to improve the forecast skill, in some scenarios in Figure 5 there is not so much difference between the two. On the contrary, the smooth model exhibits much more variation in the skill. Therefore it might be interesting to include a table or figure regarding the computation time of the different approaches. In case e.g. that the smooth model takes much more computation time than the SW and SW plus approach, then this could maybe lead to a recommendation/rule of thumb for practical use, like the more sophisticated smooth model does not provide so much more improvement than the SW plus, but has much higher computation time, so for practical use the SW plus suffices.

The computation time for the various sliding-window approaches is in the order of seconds, whereas the estimation of the *smooth model* takes a few minutes. The latter is however estimated using MCMC sampling, which allows drawing inferential conclusions about the selected terms but is not mandatory for practical use.

In addition, for the sliding-window approaches the NR models must be re-estimated every day, whereas the same *smooth model* is valid for several years. We have included these times in the end of Sect. 2.3.2, but want to point out that these are only valid for the employed estimation software listed in the section “code availability”.

In that regard, the question could be addressed whether these two models indeed do not significantly differ. You might consider adding p-values of some statistical (student-t, wilcoxon, or diebold mariano) test comparing whether the average performance is significantly different or not.

The purpose of the evaluation is to reveal patterns in the performance of the different strategies and to build a better awareness of possible constraints of the various methods, rather than to evaluate if the performance differences are significant. As summarized in Sect. 4 of the manuscript, all four training schemes have their advantages in particular applications; at the end it’s up to the user to select the appropriate training-scheme for his/her specific application. Thus, we have have decided not to include an additional evaluation of the predictive performances of the different methods as we think that it may distract the readers from the main objective of the article.

Technical comments:

Section 2.2.2.: You introduce the regularized sliding window approach of Scheuerer (2014). You only mention that the approach yielded better results in case of precipitation. But you do not really mention that another distribution was used in Scheuerer (2014). As your case study is only based on the normal distribution, it should be explicitly stated that the results in Scheuerer (2014) are for a non-Gaussian distribution.

Thank you for pointing that out. We agree and have rephrased “post-processing precipitation amounts employing a left-censored generalized extreme value distribution” in Sect. 2.2.2.

Figure 3 and 4: The two validation years in data set A are both plotted in each of the panels representing a specific sliding window approach, both as dashed lines. It is really difficult to distinguish the lines belonging to the different years. Maybe you could try two different line types, and/or line thicknesses, so that one can distinguish the trajectories of the two years more easily.

Thank you for pointing out this possible confusion. For clarity, we now use different line types for the consecutive calendar years in both Fig. 3 and 4.

Figure 5: The flat bar representing the “boxplot” for the standard sliding window approach could removed from the figure. As the standard SW approach is the reference model for the skill scores, this flat boxplot does not really provide any additional information, but it confuses at first sight

We agree that the flat box-whiskers provide no additional information. We had included these as a visible reference, but this has apparently not added to the clarity of this figure. Hence, we have removed the reference from Fig. 5 in the manuscript.

Stauffer R, Mayr GJ, Messner JW, Umlauf N, Zeileis A (2017a). Spatio-Temporal Precipitation Climatology over Complex Terrain Using a Censored Additive Regression Model. *International Journal of Climatology*, **37**(7), 3264–3275. 10.1002/joc.4913.

Stauffer R, Umlauf N, Messner JW, Mayr GJ, Zeileis A (2017b). Ensemble Postprocessing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies. *Monthly Weather Review*, **145**(3), 955–969. 10.1175/MWR-D-16-0260.1.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2019-49, 2019.

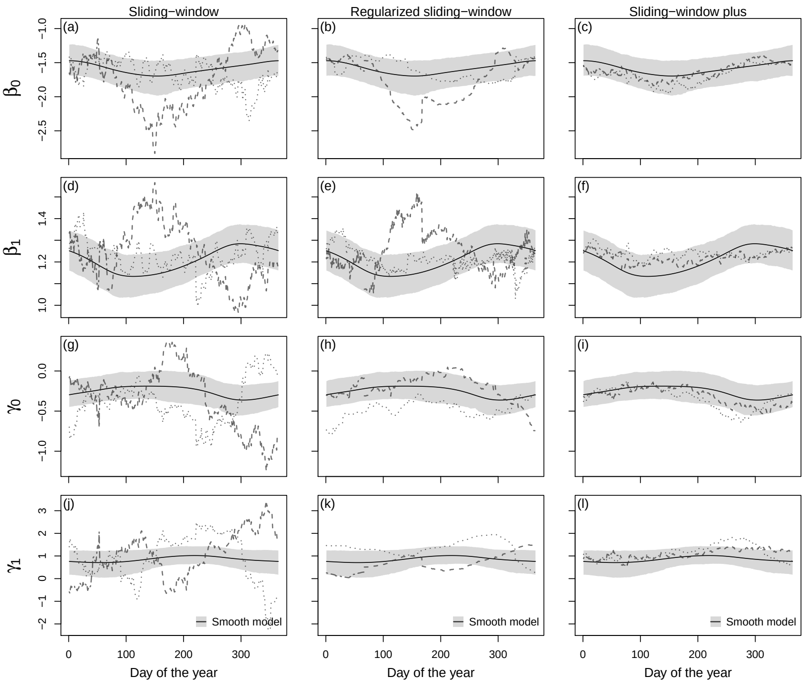

Fig. 1. Similar to Fig. 3 in the paper, the temporal evolution of regression coefficients is shown for the validation period in data set A for Innsbruck at forecast step +24h (valid at 0000 UTC). Contrary to Fig. 3 in the paper, regression coefficients are presented for post-processing daily precipitation sums, employing a left-censored Gaussian response distribution with a sliding window length of 80 days.

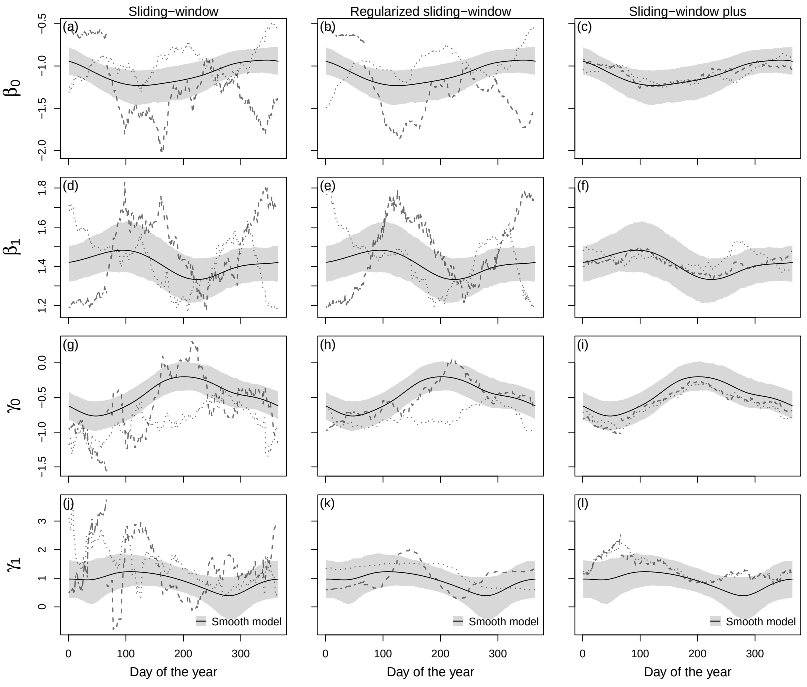

Fig. 2. As Fig. 1 in this rebuttal letter, but for Hamburg at forecast step +24h (valid at 0000 UTC).

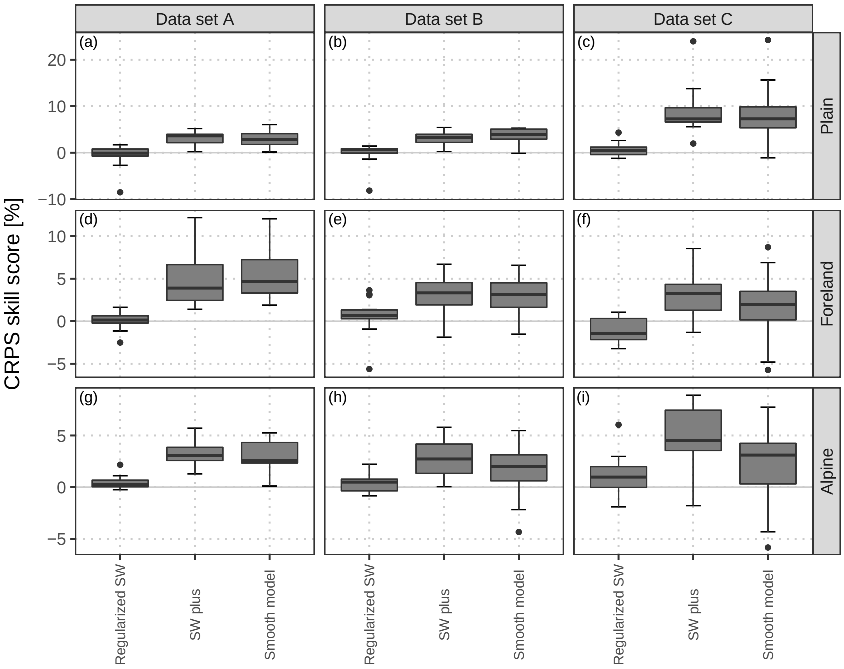

Fig. 3. Similar to Fig. 5 in the paper, CRPS skill scores are shown with the classical sliding-window approach as a reference. Contrary to Fig.3 in the paper, regression coefficients are presented for post-processing daily precipitation sums, employing a left-censored Gaussian response distribution with a sliding window length of 80 days. Each box-whisker contains aggregated skill scores over the forecast steps from +24h to +72h on a 24 hourly temporal resolution and over five respective weather stations.

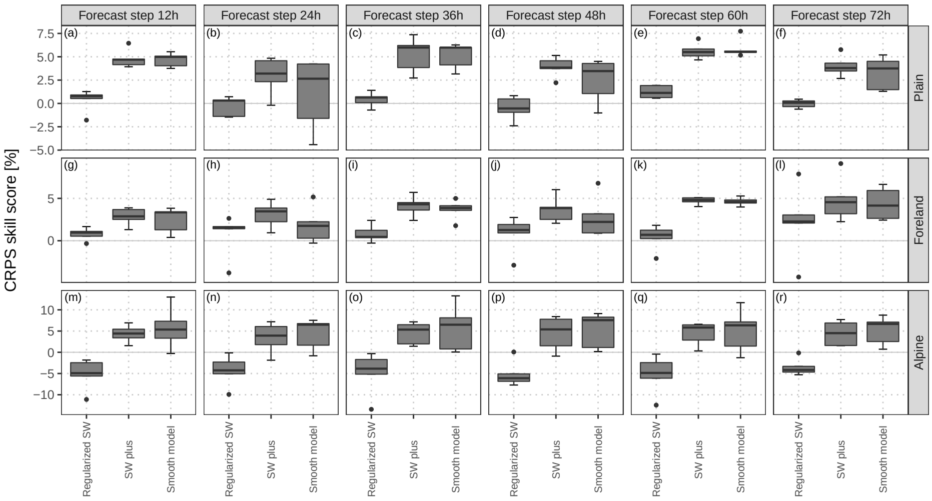

Fig. 4. As Fig. 5 in the paper, CRPS skill scores are shown with the classical sliding-window approach as a reference. Contrary to Fig. 5 in the paper, all scores are shown only for data set A but conditional on the forecast steps from +12h to +72h on a 12 hourly temporal resolution.