universität innsbruck

To Maxime Taillardat

Name / Email
Moritz Lang
Moritz.Lang@uibk.ac.at

Date
December 10, 2019

Revision of 'npg-2019-49'

Dear Maxime Taillardat

We thank both reviewers for the positive and constructive feedback regarding our manuscript "**Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression**".

We have carefully revised our manuscript according to their comments and suggestions. The most substantial changes 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.

On the following pages a point-to-point response to both reviewers will be given. The attached manuscript highlights the changes in the text in blue color. In addition, we have added a supplement on post-processing daily precipitation sums after the revised manuscript.

Your sincerely,

Moritz Lang
Corresponding Author

Reviewer 1

This manuscript compares the effect of different schemes to compose training data for statistical post-processing methods (here: non-homogeneous regression) on the performance of the resulting forecasts. It is well written and highly relevant to operational forecasting where availability of reforecast data may be limited and the consequences of changes in the NWP model on forecast calibration must be understood in order to decide whether forecasts from an older NWP model version can be used to fit the parameters defining the post-processing model. This last point is the only one where I feel the manuscript could benefit from a more detailed discussion.

We want to thank you for your fruitful and constructive review. We have been carefully going trough your comments to address each point individually including your general comment on NWP model changes.

Below, you can find a detailed point-to-point reply to your comments and suggestions.

Specifically:

The CRPS skill scores in Fig. 5 h) suggest that the regularization scheme struggles with the adjustment to the NWP model upgrade and to the annual cycle, but also the SW plus and the smooth model have an overall neutral effect on skill even though these schemes increase the training sample size significantly. It would be interesting to better understand the causes of this result.

Figure 5 h shows the CRPS skill scores for alpine sites for data set B. Thus, the *smooth model* is trained on the 'old EPS version' while the predictions are for the 'new EPS version' which explains its relative performance loss. This is also true for the *sliding-window plus* which is, in large parts, based on data from the 'old EPS version'. The classical *sliding-window* approach and the *regularized sliding-window* approach both adjust to the 'new EPS version' more rapidly at the same pace and, hence, show a similar predictive performance difference as for data set A (Fig. 5 g).

Figure 3 gives some good idea about the problems with the regularization scheme (parameters adjust very slowly to changes) but it is not ideal to illustrate problems with 'SW plus' and the 'Smooth model' since no NWP model upgrade happens in data set A. Wouldn't it be better to use data set B for this figure, where we can expect some adjustment during the first days/weeks of the validation period? Also, is Innsbruck the best location to illustrate the effect of a NWP model upgrade? As 'best' alpine location in this context I would consider the one that is most strongly impacted by the horizontal resolution change (this could be studied by considering changes in biases in the raw ensemble forecasts) in the ECMWF model and therefore presents a worst case scenario in terms of adjustment to a NWP model upgrade.

Figure 1 shows estimated coefficient paths for the weather station Altdorf for the first calendar year of the validation period of data set B. Altdorf is the station which is most strongly affected by the model change with a height difference in the model topography of 47.4 m. The first 40 days within the validation, which correspond to the period directly after the EPS change, is highlighted in pink. The variability of the coefficients within the first 40 days compared to the rest of the year are in the same order and hence no clear adjustment can be detected.

Despite the well reasoned comment, the analyses for data set B provide no further insights to the adjustment phases of the *sliding-window* and *regularized sliding-window* approaches. Hence, to restrict the presented analysis to the error sources (v) and (vi) described in Sect. 2.1 of the manuscript, we suggest to keep showing the coefficient paths for data set A in the paper. Consequently, as for data set A no EPS change has to be considered, we have kept the results for Innsbruck in the manuscript.

I would encourage the authors to provide some more discussion along these lines, since NWP model upgrades have been the main argument to justify the need for reforecasts, and I am not aware of any previous study that looks at the effect of NWP model upgrades on the performance of post-processed ensemble forecasts in a quantitative way.

We agree that analyzing the effects of EPS changes would be a very interesting research question on its own. We have tried to account for this issue by our study design (data set A, B, C), however, for a comprehensive perspective on this topic one would need to perform an extensive analysis on more than 15 stations. This is beyond the objective of this paper which mainly aims at presenting how time-adaptive post-processing scheme are related to each other and how these perform under specific restrictions, such as the EPS change which affects the horizontal resolution investigated in our current study. Other studies focus more on investigating the effect of model changes themselves such as, e.g., Demaeyer and Vannitsem (2019) by studying the impact of changes in a quasi-geostrophic model on post-processing.

To address your initial comment on when it is beneficial to use data from a previous NWP model for forecast calibration: As the results in our study show, the time-adaptive training schemes using multiple years of data are superior to the ones using the most recent days only, even in case of the EPS change investigated. However, this might look different

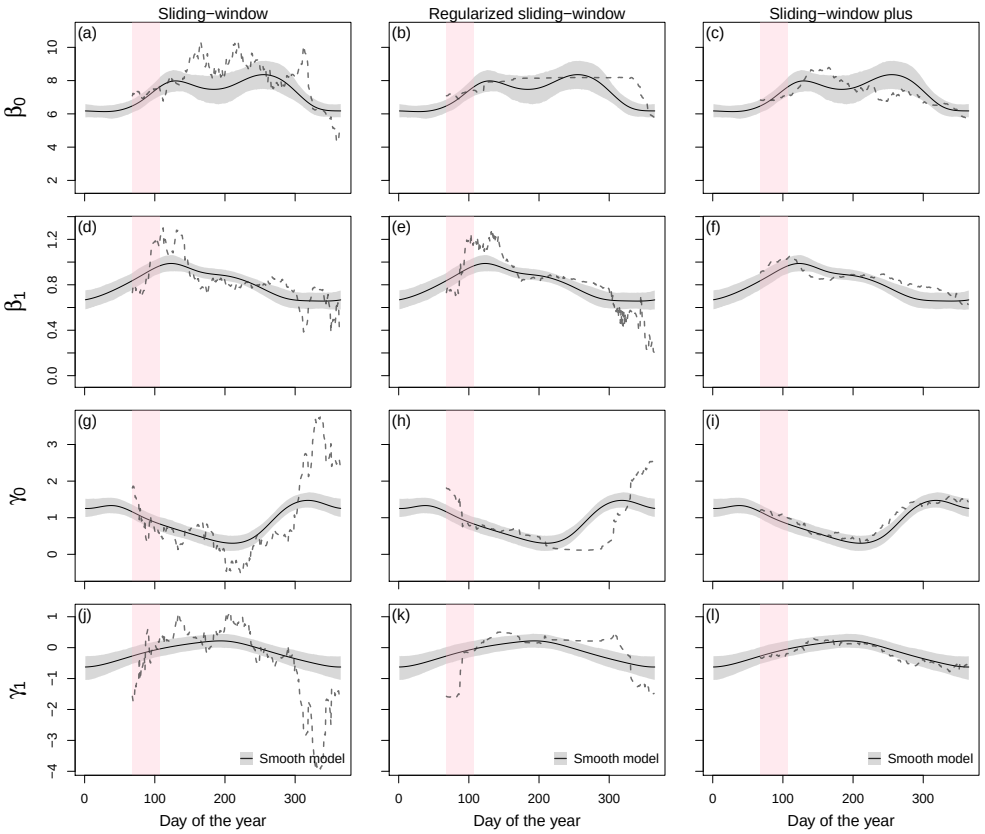

Figure 1: *As Fig. 3 and 4 in the paper, but for station Altdorf for the first calendar year during the validation period in data set B. The first 40 days within the validation period, which correspond to the period directly after the change in the horizontal resolution of the ECMWF EPS on March 8, 2016, is highlighted in pink.*

for a different NWP model and/or future model changes and must be evaluated individually in each case. This is now explicitly stated in the conclusion of the manuscript (Sect. 4).

It is interesting to see in Figs. 3 and 4 that this type of regularization seems to introduce too much inertia, i.e. the parameters only adapt with a certain delay or sometimes not at all. Have the authors tested alternative stopping criteria? A simple and obvious variant would be to perform 2 or 3 iterations on each (except the first) new day.

Figure 2 shows the temporal evolution of regression coefficients exemplary for Innsbruck using three instead of only one iteration. In comparison to the *regularized sliding-window* model version presented in Fig. 3 of the manuscript, the temporal evolution of the coefficients for the modified *regularized sliding-window* approach is much more comparable to the evolution of the classical *sliding-window* approach. This increased similarity is also visible in the aggregated CRPS skill scores shown in Fig. 3. In comparison to original *regularized sliding-window* approach discussed in the manuscript, the modified *regularized sliding-window* approach has both less profound performance losses for alpine stations, and less visible performance gains for stations in the plain and foreland with the classical *sliding-window* approach as a reference.

In summary, three iterations used in the estimation process for each new day is not generally superior to a single iteration for the employed data set. To show a wide spectrum of commonly-used training schemes in a unified setup, we kept the different approaches as close as possible to the originally proposed version to show their advantages and possible disadvantages Thus, we refrain from introducing further modifications such as e.g., additional hyperparameter tuning as now stated in the conclusion (Sect. 4). In addition, we now explicitly emphasize in Sect. 2.2.3 of the paper that an increased number of iterations might be more appropriate for the *regularized sliding-window* approach depending on the employed data.

Minor comments:

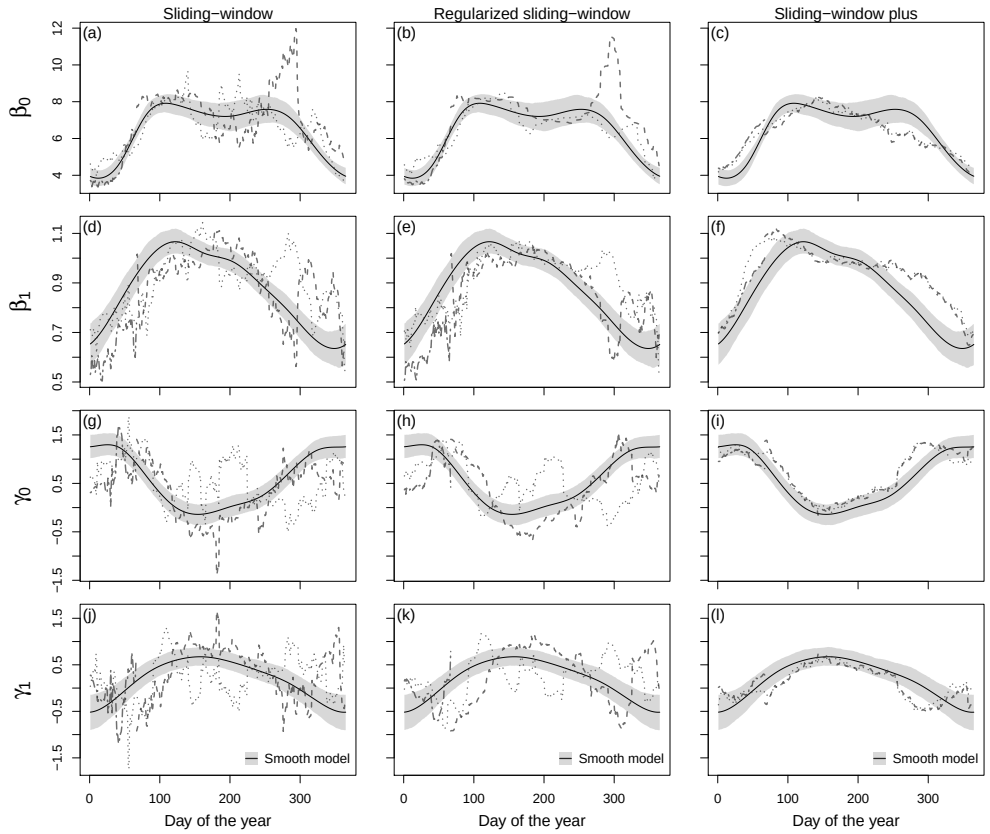

Figure 2: *As Fig. 3 in the paper, the temporal evolution of regression coefficients is shown for the validation period in data set A for Innsbruck at forecast step $+36$ h (valid at 1200 UTC). Contrary to Fig. 3 in the paper, the modified* regularized sliding-window *approach uses three iterations in the estimation process before the optimizer is stopped. The coefficient paths are plotted for the consecutive calendar years 2014, 2015, and 2016 as dashed, dotted, and two-dashed line, respectively.*

244-245: While it's possible (even likely) that a larger slope coefficient is due to higher skill of the EPS temperature forecasts, one cannot be sure if at least to some extent the larger slope coefficient is due to an amplitude bias of the raw ensemble forecasts, i.e. the ensemble underpredicts high temperatures and overpredicts low temperatures, and increasing the slope coefficient compensates for that.

A very good remark! We have corrected the statement in Sect. 3.1 according to your comment.

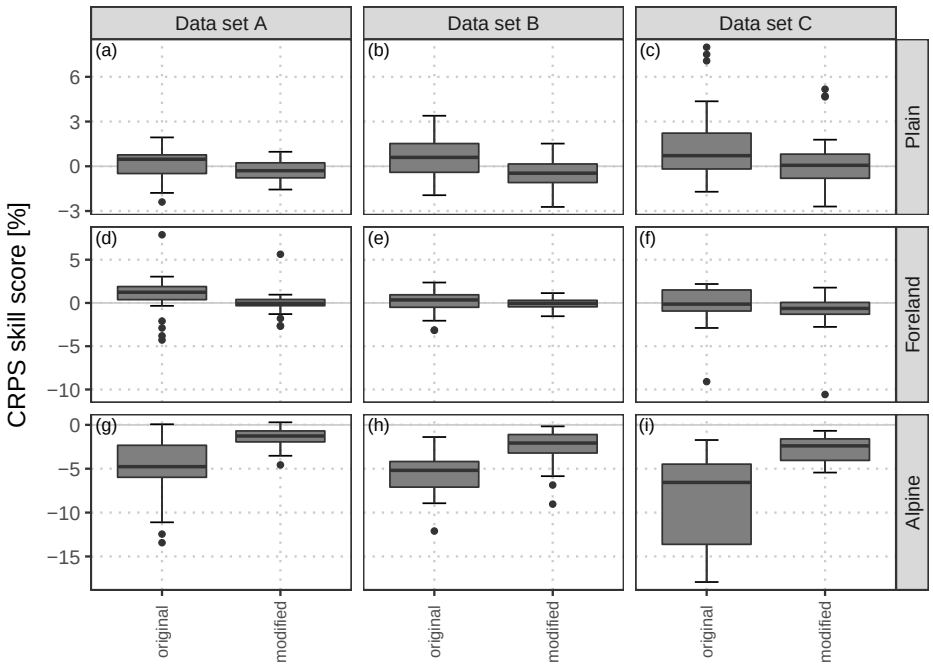

Figure 3: *Similar to Fig. 5 in the paper, but only for the* regularized sliding-window *approaches with the classical* sliding-window *approach as a reference. The* original *version as presented in the manuscript uses a single iteration, whereas the* modified *version uses three iterations in the estimation process before the optimizer is stopped.*

Reviewer 2

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.

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.

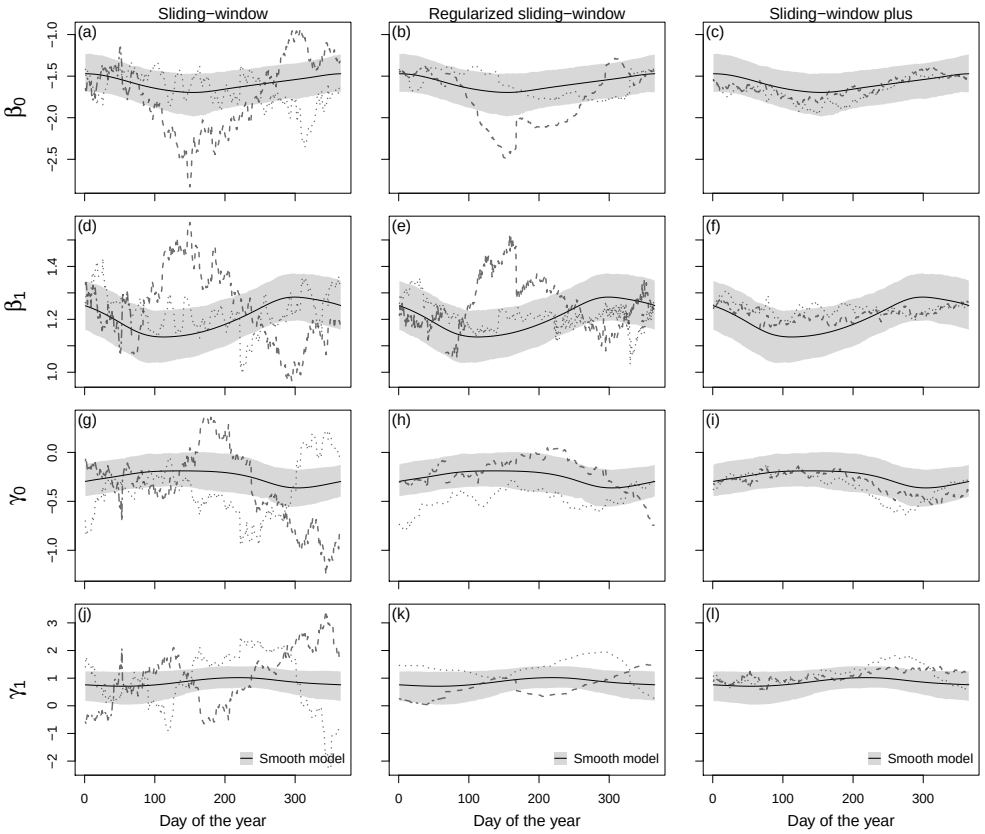

Figure 4: *Similar to Fig. 3 in the paper, the temporal evolution of regression coefficients is shown for the validation period in data set A for Innsbruck at forecast step $+24$ h (valid at 0000 UTC). Contrary to Fig. 3 in the paper, regression coefficients are presented for post-processing daily precipitation sums, employing a left-censored Gaussian response distribution with a sliding window length of 80 days.*

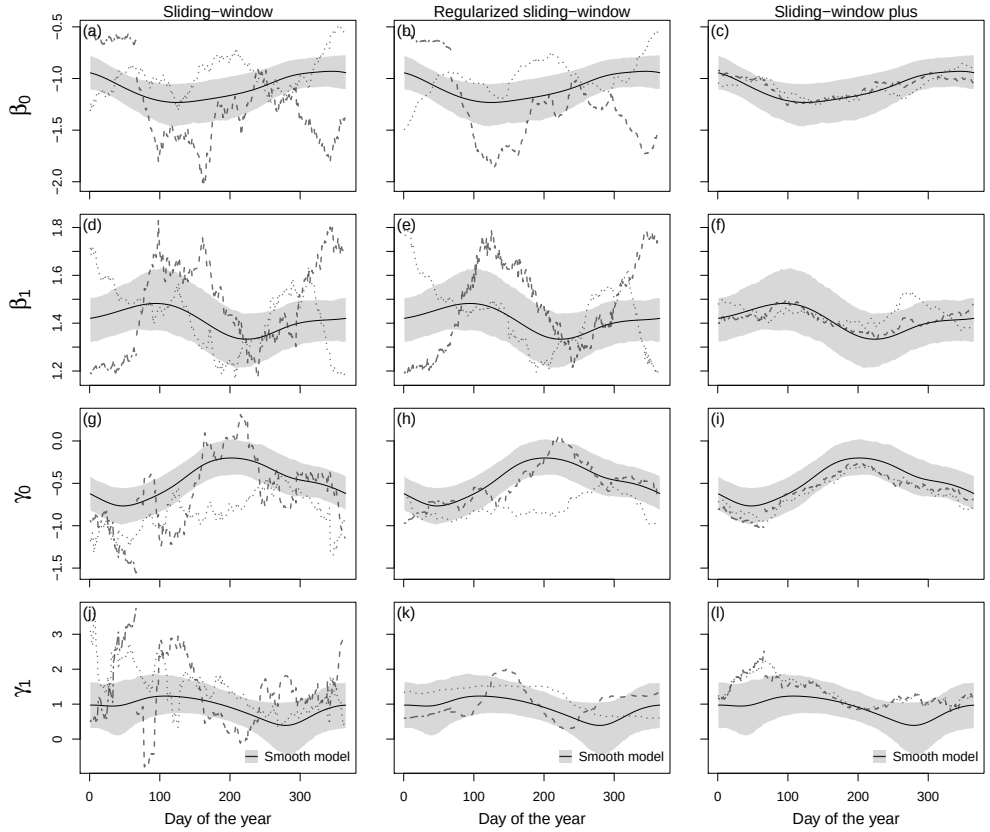

Figure 5: *As Fig. 4 in this rebuttal letter, but for Hamburg at forecast step $+24\,h$ (valid at 0000 UTC).*

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 4 and Fig. 5 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 6 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. 4–6, 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. 4 and 5).
- 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. 4 and 5).
- 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. 6).

All presented results are very similar to the analyses for 2 m temperature forecasts presented in the manuscript and

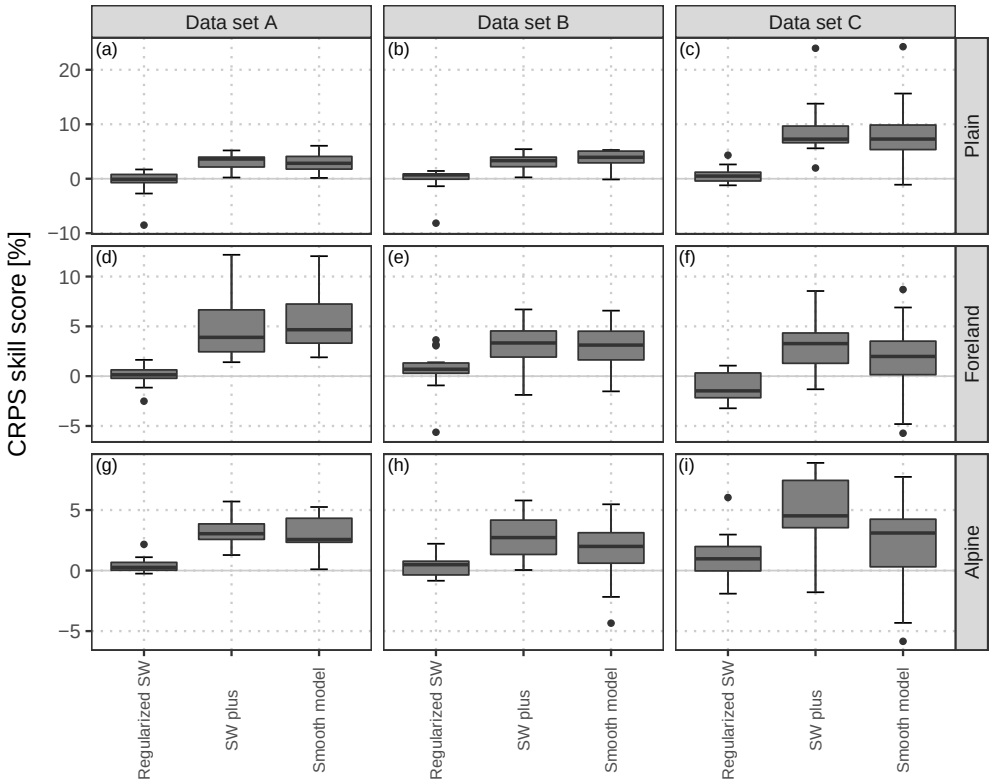

Figure 6: *Similar to Fig. 5 in the paper, CRPS skill scores are shown with the classical* sliding-window *approach as a reference. Contrary to Fig. 3 in the paper, regression coefficients are presented for post-processing daily precipitation sums, employing a left-censored Gaussian response distribution with a sliding window length of* 80 *days. Each box-whisker contains aggregated skill scores over the forecast steps from* $+24$ *h to +72 h on a 24 hourly temporal resolution and over five respective weather stations.*

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 7 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. 7 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

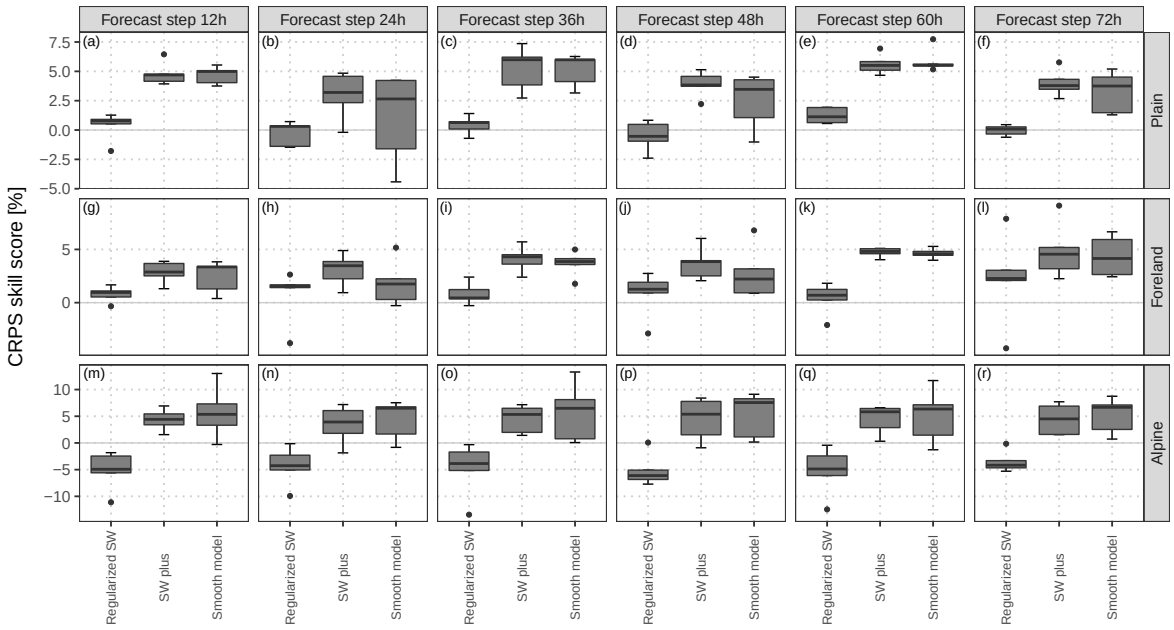

Figure 7: *As Fig. 5 in the paper, CRPS skill scores are shown with the classical* sliding-window *approach as a reference. Contrary to Fig. 5 in the paper, all scores are shown only for data set A but conditional on the forecast steps from* $+12$ *h to* $+72$ *h on a* 12 *hourly temporal resolution.*

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.

* References

Demaeyer J, Vannitsem S (2019). "Correcting for Model Changes in Statistical Post-Processing – An approach based on Response Theory." *Nonlinear Processes in Geophysics Discussions*, **2019**, 1–27. doi:10.5194/npg-2019-57.

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. doi: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. doi:10.1175/MWR-D-16-0260.1.

Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression

Moritz N. Lang[1,2], Sebastian Lerch[3], Georg J. Mayr[2], Thorsten Simon[1,2], Reto Stauffer[1,4], and Achim Zeileis[1]

[1]Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2]Department of Atmospheric and Cryospheric Sciences, Universität Innsbruck, Innsbruck, Austria
[3]Institute for Stochastics, Karlsruher Institut für Technologie, Karlsruhe, Germany
[4]Digital Science Center, Universität Innsbruck, Innsbruck, Austria

Correspondence: Moritz N. Lang (moritz.lang@uibk.ac.at)

Abstract. Non-homogeneous regression is a frequently-used post-processing method for increasing the predictive skill of probabilistic ensemble weather forecasts. To adjust for seasonally varying error characteristics between ensemble forecasts and corresponding observations, different time-adaptive training schemes, including the classical sliding training window, have been developed for non-homogeneous regression. This study compares three such training approaches with the sliding-window approach for the application of post-processing near-surface air temperature forecasts across Central Europe. The predictive performance is evaluated conditional on three different groups of stations located in plains, in mountain foreland, and within mountainous terrain, as well as on [..[1]]a specific change in the ensemble forecast system of the European Centre for Medium-Range Weather Forecasts (ECMWF) used as input for the post-processing.

The results show that time-adaptive training schemes using data over multiple years stabilize the temporal evolution of the coefficient estimates, yielding an increased predictive performance for all station types tested compared to the classical sliding-window approach based on the most recent days only. While this may not be surprising under fully stable model conditions, it is shown that "remembering the past" from multiple years of training data is typically also superior to the classical sliding-window when the ensemble prediction system is affected by certain model changes. Thus, reducing the variance of the non-homogeneous regression estimates due to increased training data appears to be more important than reducing its bias by adapting rapidly to the most current training data only.

1 Introduction

The need of accurate probabilistic weather forecasts steadily increases, because reliable information about the expected uncertainty is crucial for optimal risk assessment in agriculture and industry, or for personal planning of outdoor activities. Therefore, most forecast centers nowadays issue probabilistic forecasts based on ensemble prediction systems (EPSs). To quantify the uncertainty of a specific forecast, an EPS provides a set of numerical weather predictions using slightly perturbed initial conditions and different model parameterizations (Palmer, 2002). However, due to various constraints and required simplifications in the

[1]removed: changes

EPS, these forecasts often show systematic biases and capture only parts of the expected uncertainty; especially when EPS forecasts are directly compared to point measurements (Gneiting and Katzfuss, 2014). In order to increase the predictive skill of the forecasts for specific locations, statistical post-processing is often applied to correct for these systematic errors in the forecasts' expectation and uncertainty.

One of the most frequently used parametric post-processing methods is 'ensemble model output statistics' (EMOS) introduced by Gneiting et al. (2005). To emphasize that not only the errors in the mean but also the errors in the uncertainty are corrected, the method is often referred to as 'non-homogeneous regression' (NR). In the statistical literature, this type of model is also known as distributional regression (Klein et al., 2014) since all parameters of a specific response distribution are optimized simultaneously conditional on respective sets of covariates.

As the error characteristics between the covariates, typically provided by the EPS, and the observations often show seasonal dependencies and might change inter-annually over time, different time-adaptive training schemes have been developed for NR models. Gneiting et al. (2005) proposed the so-called 'sliding training window' approach where the training data set consists of EPS forecasts and observations of the most recent 30–60 days only. As soon as new data become available, the training data set and the statistical model are updated so that the estimated coefficients automatically evolve over time and adjust to changing error characteristics. This makes it very handy for operational use, however, little training data can sometimes yield unrealistic jumps in the estimated coefficients over time, especially if events which show a significantly different error characteristic enter the training data set. Therefore, to stabilize the temporal variability of the coefficient estimates, several approaches have been proposed in the literature. Scheuerer (2014) regularizes the estimation by only allowing the optimizer to slightly adjust the coefficient from day to day. In an alternative approach, Möller et al. (2018) extend the training data by using not only the days prior to estimation, but also the days centered around the same calendar day over all previous years available. This idea of using a rolling centered training data set over multiple years is similar to the concept of using annual cyclic smooth functions to capture seasonality as employed by Lang et al. (2019). These smooth functions are also known as regression splines (Wood, 2017), where the estimate of each point in the function only depends on data in its closer neighborhood; this allows for a smooth and stable evolution of the coefficients over the year.

Alternative time-adaptive models are based on historical analogs or non-parametric approaches. For approaches employing analogs (Junk et al., 2015; Barnes et al., 2019), training sets are selected to consist of past forecast cases with atmospheric conditions similar to those on the day of interest. Such methods may lead to models that are able to account for the flow-dependency of EPS errors (Pantillon et al., 2018; Rodwell et al., 2018). However, the definition and computation of similarity measures is far from straightforward, and substantial methodological developments may be required to obtain suitably extensive training data sets for stable model estimation (Hamill et al., 2008; Lerch and Baran, 2017). For non-parametric approaches (Taillardat et al., 2016; Henzi et al., 2019) or semi-parametric approaches (Rasp and Lerch, 2018; Schlosser et al., 2019), time-adaptive choices of the training data are typically abandoned as well, as interactions between the day of the year and other covariates can capture the potential time-adaptiveness. Therefore, analog-based and non-parametric approaches will not be pursued further in the context of this work.

In addition to the training scheme employed, an important data-specific aspect which has to be considered in post-processing is that the EPS may change over time (Hamill, 2018). [..[2]]This also motivates the recent study of Demaeyer and Vannitsem (2019), which introduces the promising concept of a post-processing method specifically dealing with model changes in a simplified physical setup. However, as stated by the authors, more research would be required to transfer their findings to real case scenarios. When using data of an operational EPS, changes in the underlying numerical model such as, e.g., an increased horizontal resolution, can [..[3]]typically lead to sudden transitions in the predictive performance of the EPS and hence affect the error characteristics of the data. If the training data set used to estimate the statistical post-processing model contains data of a previous EPS version which significantly differs from the current one, it can result in a loss of the predictive performance.

This paper presents a comparison of four widely-used different time-adaptive training schemes proposed in [..[4]]the literature that employ alternative strategies to account for varying error-characteristics in the data. To show a wide spectrum of possible approaches in a unified setup – rather than finding the universally best method – we consider typical basic applications of these training schemes and refrain from more elaborate tuning or combinations. A case study is shown for post-processed 2 m temperature forecasts for three different groups of stations across Central Europe in the midlatitudes, namely stations in the plain, in the foreland, and within mountainous terrain (Fig. 1). The study highlights the advantages and drawbacks of the different approaches in different topographical environments and investigates the impact of [..[5]]a change in the horizontal resolution of the EPS, which is expected to have a particularly pronounced effect on the predictive performance[..[6]].

The structure of the paper is as follows: Section 2 explains the different methods and the comparison setup including the underlying data. In Sect. 3, the different time-adaptive training schemes are compared in terms of their coefficient paths and their predictive performance. Finally, a summary and conclusion is given in Sect. 4.

2 Methodology and comparison setup

The different training schemes for NR models proposed in the literature try to adapt to various kinds of error sources that can occur in post-processing, both in space and time. In order to provide a unifying view and to fix jargon, we first discuss these different error sources and then introduce the training schemes considered along with the comparison setup employed.

2.1 Sources of errors in post-processing

NR models aim to adjust for errors and biases in EPS forecasts but, of course, the NR models can be affected by errors and misspecifications themselves. Therefore, we try to carefully distinguish the two different models involved with their as-

[2] removed: Changes
[3] removed: possibly
[4] removed: literature
[5] removed: changes in the EPS
[6] removed: in different topographical environments

sociated errors, i.e., the numerical weather prediction model underlying the EPS vs. the statistical NR model employed for post-processing.

The skill of the EPS can be quantified in EPS forecast biases and variances which (i) typically vary for different locations conditional on the surrounding terrain, (ii) often show cyclic seasonal patterns, and (iii) can experience non-seasonal temporal changes, e.g., due to changes in the EPS itself.

In addition to the error sources in the employed EPS, the performance of the statistical post-processing itself will typically also (iv) differ at different measurement sites, (v) strongly depend on the amount of training data used, and (vi) whether it is affected by effects that are not accounted for in the NR specification.

Clearly, larger training samples (v) will lead to more reliable predictions when the NR specification (vi) – in terms of response distribution, covariates and corresponding effects, link functions, estimation method, etc. – appropriately captures the error characteristics in the relationship between EPS forecasts and actual observations. However, when these error characteristics differ in space (i and iv) and/or in time (ii and iii), it is not obvious what the best strategy for training the NR is. Extending the training data (v) in space or time will reduce the variance of the NR estimation but might also introduce bias if the NR specification (vi) is not adapted. Thus, this is a classical bias-variance trade-off problem and we investigate which strategies for dealing with this are most useful in a typical temperature forecasting situation.

To fix jargon, we employ the terms "model" and "bias" without further qualifiers when referring to the NR model in post-processing. Whereas when referring to the numerical weather prediction model we employ "EPS model" and "EPS bias". Moreover, we refer to a statistical model whose estimates have small bias and variance as stable.

2.2 Non-homogeneous regression with time-adaptive training schemes

Non-homogeneous regression as originally introduced by Gneiting et al. (2005) is a special case of distributional regression, where a response variable y is assumed to follow a specific probability distribution $\mathcal{D}$ with distribution parameters $\theta_k, k = 1, \ldots, K$:

$$y \sim \mathcal{D}(\theta_1, \ldots, \theta_K) = \mathcal{D}(\mathrm{h}_1(\eta_1), \ldots, \mathrm{h}_K(\eta_K)), \tag{1}$$

where each parameter of the distribution is linked to an additive predictor η_k via a link function h_k to ensure its appropriate co-domain. In case of post-processing air temperatures, the normal distribution is typically employed (Gneiting and Katzfuss, 2014), and Eq. (1) can be rewritten as

$$y \sim \mathcal{N}(\mu, \sigma). \tag{2}$$

In the classical NR (Gneiting et al., 2005), the two distribution parameters location μ and scale σ are expressed by the ensemble mean m and ensemble variance or standard deviation s, respectively:

$$\mu = \eta_\mu = \beta_0 + \beta_1 \cdot m, \tag{3}$$

$$\log(\sigma) = \eta_\sigma = \gamma_0 + \gamma_1 \cdot s, \tag{4}$$

with $\beta_\bullet$ and $\gamma_\bullet$ being the corresponding intercept and slope coefficients. Here, we use the logarithm link to ensure positivity of the scale parameter σ, however, a quadratic link with additional parameter constraints for the coefficients as used by Gneiting et al. (2005) would also be feasible. In this study, we regard the statistical model specifications according to Eq. (2)–(4), but all concepts of time-adaptive training schemes could easily be transferred to other response distributions $\mathcal{D}$, to alternative link functions $\mathrm{h}(\cdot)$, or to more complex additive predictors η with additional covariates.

The regression coefficients $\beta_\bullet$ and $\gamma_\bullet$ are estimated by minimizing a loss function over a training data set containing historical pairs of observations and EPS forecasts. In this study, we employ maximum likelihood estimation, which performs very similar to minimizing the continuous ranked probability score (CRPS, Gneiting and Raftery 2007) as used by Gneiting et al. (2005) when the response distribution is well specified (Gebetsberger et al., 2018). For a single observation y, the log-likelihood L of the normal distribution is given by

$$L(\mu, \sigma | y) = \log \left\{ \frac{1}{\sigma} \phi \left(\frac{y - \mu}{\sigma} \right) \right\}, \tag{5}$$

where $\phi(\cdot)$ is the probability density function of the normal distribution. The coefficients $\beta_\bullet$ and $\gamma_\bullet$, specified in Eq. (3) and (4), are derived by minimizing the sum of negative log-likelihood contributions L over the training data. The larger the training data the more stable is the estimation in case the statistical model is well specified; however, if the covariate's skill varies either seasonally or non-seasonally over time, this leads to the bias-variance trade-off between preferable large training data sets for stable estimation and the benefit of shorter training periods which allow to adjust more rapidly to changes in the data or, to be precise, in the error characteristics of the data (see Sect. 2.1). In the following, four approaches are discussed how to gain informative time-adaptive training data sets while ensuring a stable estimation.

2.2.1 Sliding-window

The *sliding-window* approach originally introduced by Gneiting et al. (2005) uses the most recent days prior to the day of interest as training data for estimation. For post-processing 2 m temperature forecasts, Gneiting et al. (2005) found the best predictive performance for training periods between 30 and 45 days with substantial gains in increasing the training period beyond 30 days and slow but steady performance losses for training lengths beyond 45 days. According to Gneiting et al. (2005), the latter is presumably a result of seasonally varying EPS forecast biases.

In this study, we use a period of 40 days for the *sliding-window* approach, which is a frequently used compromise (e.g., Baran and Möller 2017; Gneiting et al. 2005; Wilson et al. 2007). However, as discussed in Gneiting et al. (2005), different training periods might perform better for distinct weather variables, locations, forecast steps, or model specifications. Common choices in the literature include training lengths between 15 and 100 days, for example depending on whether the estimation of regression coefficients is performed station-specific or jointly for multiple locations at once.

2.2.2 Regularized sliding-window

A regularized adaption of the classical *sliding-window* approach was introduced by Scheuerer (2014) in order to stabilize the estimation based on early stopping in statistical learning. The motivation is that gradient-based optimizers adjust the starting

values by iteratively taking steps in the direction of the steepest descent of a distinct loss function until some convergence condition is fulfilled. These steps are largest in the first iteration and getting smaller towards the optimum. Thus, the most important adjustments are made during the first steps, while further adjustments often improve the fit to unimportant or even random features in the data which can lead to wiggly coefficient paths over time and ultimately to an overfitting (Scheuerer, 2014).

Therefore, Scheuerer (2014) proposes to use the coefficients of the previous day as starting values and to stop the optimizer after a single iteration to stabilize the evolution of the coefficient estimates. A drawback of his approach is that it implies that the estimation never converges and in case of poor starting values or strong truly observed temporal changes in the data the obtained coefficients might be incorrect (Scheuerer, 2014). For post-processing precipitation amounts employing a left-censored generalized extreme value distribution, Scheuerer (2014) obtained better results with regularized coefficients than without regularization.

For the *regularized sliding-window* approach used in this study, we employ the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm as in Scheuerer (2014) and stop the optimizer after one single iteration. For the first time, we let the BFGS algorithm perform 10 iterations and use $(\beta_0, \beta_1)^\top = (0, 1)^\top$ as starting values in the location parameter μ and $(\gamma_0, \gamma_1)^\top = (0.1, 1)^\top$ as starting values in the scale parameter σ. According to Scheuerer (2014) a single iteration might not always provide the optimum degree of regularization, however, for the presented comparison study a single iteration yields a regularized setup which is on the opposite side of the possible model spectrum compared to the classical *sliding-window* approach which runs until convergence. In comparison to Scheuerer (2014), we perform maximum likelihood estimation instead of CRPS minimization.

2.2.3 Sliding-window plus

As already pointed out by Gneiting et al. (2005), training data from previous years could additionally be included in the *sliding-window* approach to address seasonal effects. This should reduce the variance in the estimation of the regression coefficients, which stabilizes the evolution of the coefficients similar to the *regularized sliding-window* approach.

This idea has recently been pursued by Vogel et al. (2018) for the construction of climatological reference forecasts, and by Möller et al. (2018) for a post-processing approach based on D-vine copulas in which much more coefficients than in classical NR need to be estimated, making a more extensive training data set necessary. Their so-called 'refined training data set' consists of the most 45 recent days prior to the day of interest, plus 91 days centered around the same calendar day over all previous years available. Including multiple years yields more stable estimates while, on the other hand, there is the trade-off of losing the ability to quickly adjust to non-seasonal temporal changes in the EPS forecast biases. The approach of Möller et al. (2018) can be seen as time-adaptive version of the seasonal training proposed by Hemri et al. (2016) who consider training data sets comprised of days from all previous years within the same season (winter/summer) as the day of interest.

In this study, to be comparable to the *sliding-window* approach we use the most recent 40 days prior to estimation and a respective 81 days interval centered around the day of interest over the previous years available in the training data.

Table 1. Overview of time-adaptive training schemes, distinguished by model specification/estimation and training data selection corresponding to errors sources (vi) and (v), respectively. The basic model specification refers to Eq. (3)–(4) in contrast to the extended Eq. (6)–(7).

	Model		Data	
Name	Specification	Estimation	Years	Seasons
Sliding-window	Basic	Maximum likelihood	Current	Current
Regularized sliding-window	Basic	Early stopping	Current	Current
Sliding-window plus	Basic	Maximum likelihood	Multiple	Current
Smooth model	Extended	Penalized	Multiple	All

2.2.4 Smooth model

If we reformulate the *sliding-window plus* approach, it is very similar to fitting an annual cyclic smooth function where the points of the function only depend on data points in the closer neighborhood, specified by the sliding window length.

Cyclic smooth functions belong to the broader model class of generalized additive models (GAMs, Hastie and Tibshirani, 1986), which allow one to include potentially nonlinear effects in the linear predictors η. Smooth functions are also referred to as regression splines and are directly linked to the model parameters as additive terms in η. Introductory material for cyclic smooth functions conditional on the day of the year can be found in Lang et al. (2019) and a comprehensive summary on GAMs is given in Wood (2017).

To account for seasonal variations we only need to fit one single model, here called *smooth model*, over a training data set with several years of data. The effects included allow the coefficients to smoothly evolve over the year, which leads to the following adaptions in Eq. (3) and (4) for the location μ and scale σ, respectively:

$$\mu = \eta_\mu = \beta_0 + f_0(\text{doy}) + (\beta_1 + f_1(\text{doy})) \cdot m, \tag{6}$$

$$\log(\sigma) = \eta_\sigma = \underbrace{\gamma_0 + g_0(\text{doy})}_{\substack{\text{seasonally varying} \\ \text{intercept}}} + \underbrace{(\gamma_1 + g_1(\text{doy}))}_{\substack{\text{seasonally varying} \\ \text{slope}}} \cdot s, \tag{7}$$

with m and s being the ensemble mean and ensemble standard deviation, respectively; $\beta_\bullet$ and $\gamma_\bullet$ are regression coefficients, and $f_\bullet(\text{doy})$ and $g_\bullet(\text{doy})$ employ cyclic regression splines conditional on the day of the year (Wood, 2017). The regression coefficients β_0 and γ_0, and β_1 and γ_1 are unconditional on the day of the year and can be interpreted as global intercept or slope coefficients, respectively.

2.3 Comparison setup

2.3.1 NR training schemes

The NR training schemes presented in Sect. 2.2 deal with the potential temporal error sources from Sect. 2.1 in different ways (see Table 1 for an overview). The classic *sliding-window* employs the basic NR model equations from Eq. (3)–(4) and avoids potential biases in the NR model estimation by using only very recent data from the same year and season. Compared to this,

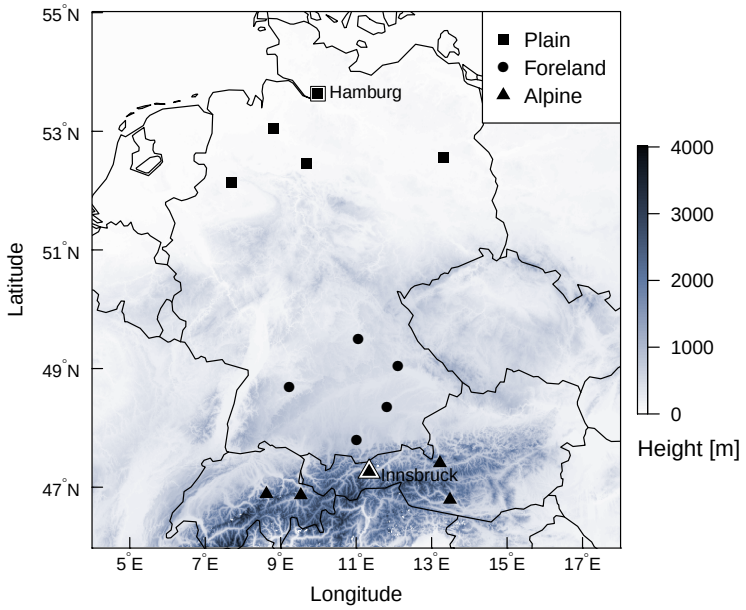

Figure 1. Overview of the study area with selected stations classified as plain, foreland, and alpine station sites. The two highlighted and labeled stations, Hamburg and Innsbruck, are discussed in detail in Sect. 3.1. Elevation data are obtained from the SRTM-30 m digital elevation model (NASA JPL, 2013).

the *regularized sliding-window* and *sliding-window plus* both try to stabilize the coefficient estimates by reducing the variance – either through regularized estimation (vi) or by considering multiple years (v). The *smooth model* differs from all of these by modifying both the model (vi) and data (v) specification, using the extended model specification from Eq. (6)–(7) fitted by penalized estimation to a large data set comprising several years and all seasons.

Potential spatial differences (i) and (iv) are handled for all training schemes in the same way: The NR models are estimated separately for each station and subsequently evaluated in groups of terrain types (plain, foreland, alpine). The underlying EPS data – described subsequently – is the same for all NR training schemes and thus affected by the same seasonal (ii) and non-seasonal changes (iii).

2.3.2 Data sets

For validation of the training schemes we consider 2 m temperature ensemble forecasts and corresponding observations at 15 measurement sites located across Austria, Germany, and Switzerland. The sites are chosen to investigate the impact of potential error sources in space (i) and (iv), e.g., through varying discrepancies between the real and the EPS topography. The data comprises three groups of five stations located either in plains, mountain foreland, and within mountainous terrain (see Fig. 1). The estimated statistical models for the stations Hamburg and Innsbruck, highlighted by symbols with white borders, are discussed in more detail in Sect. 3.1.

As covariates for Eq. (3)–(7), we employ the ensemble mean m and the ensemble standard deviation s of bilinearly interpolated 2 m temperature forecasts issued by the global 50-member EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF). We assess forecast steps from $+12$ h to $+72$ h ahead on a 12 hourly temporal resolution for the EPS run initialized at 0000 UTC and use data from March 8, 2010 to March 7, 2019.

This period has been selected in order to investigate the impact of non-seasonal long-term changes in the EPS model (iii) that is not reflected in the NR model specifications. Namely, the horizontal resolution of the ECMWF EPS changed from the the previous version (cycle 36r1; January 26, 2010) to the new version on March 8, 2016 (cycle 41r2). [..[7]]This specific model change was chosen among various others as it modifies the height of the terrain and, thus, likely introduces an EPS bias for temperature forecasts directly affecting the coefficient estimates; other changes such as modified model parameterizations or improvements in the analysis-scheme are expected to have a minor impact on the post-processing of 2 m temperatures. It is of specific interest how the *sliding-window plus* and the *smooth model* are affected if the training period comprises data from both the 'old EPS version' before the change in the horizontal resolution as well as the 'new EPS version'. Thus, we construct three data sets with different validation period that are either (A) not affected by [..[8]]this EPS model change at all, (B) start immediately after the model change, or (C) with some time lag after change.

To understand how this affects the different training schemes, we first illustrate in Figure 2a how training and validation period are selected for each scheme. For the three sliding-window approaches, the NR models are re-estimated every day as the validation date rolls through the validation period (hatched area). In contrast, the *smooth model* is estimated only once for the entire validation period based on a fixed training data period of four years prior to the validation period. For a fair comparison, the training data for the *sliding-window plus* model is also restricted to four years prior to each validation date.

Now Fig. 2b illustrates how the three data sets A, B, and C are selected in relation to the EPS change on March 8, 2016:

- Data set A: All models are trained and evaluated without being affected by the EPS change.
- Data set B: All models start with a training period entirely before the EPS change but a validation period entirely after the change. However, for the *sliding-window* and *regularized sliding window* approaches the training period quickly rolls across the change point and after 40 days they are not affected by it anymore. For the *sliding-window plus* the training data also rolls into the new EPS version but still partially uses data from the old EPS version. Finally, as the *smooth model* is only estimated once it cannot adapt at all to the new EPS version.
- Data set C: Effects from A and B are mixed so that the *smooth model* and the *sliding-window plus* model use data from both the old and new EPS version, while the classical *sliding-window* and *regularized sliding-window* models already use only data from the new EPS version.

The validation period is 2 years for A and B and 1 year for C. A total number of 731/730/365 NR models has to be estimated for the three sliding-window approaches, while only 1/1/1 *smooth model* is required for data sets A/B/C per station and forecast

[7] removed: Specifically, it is
[8] removed: the

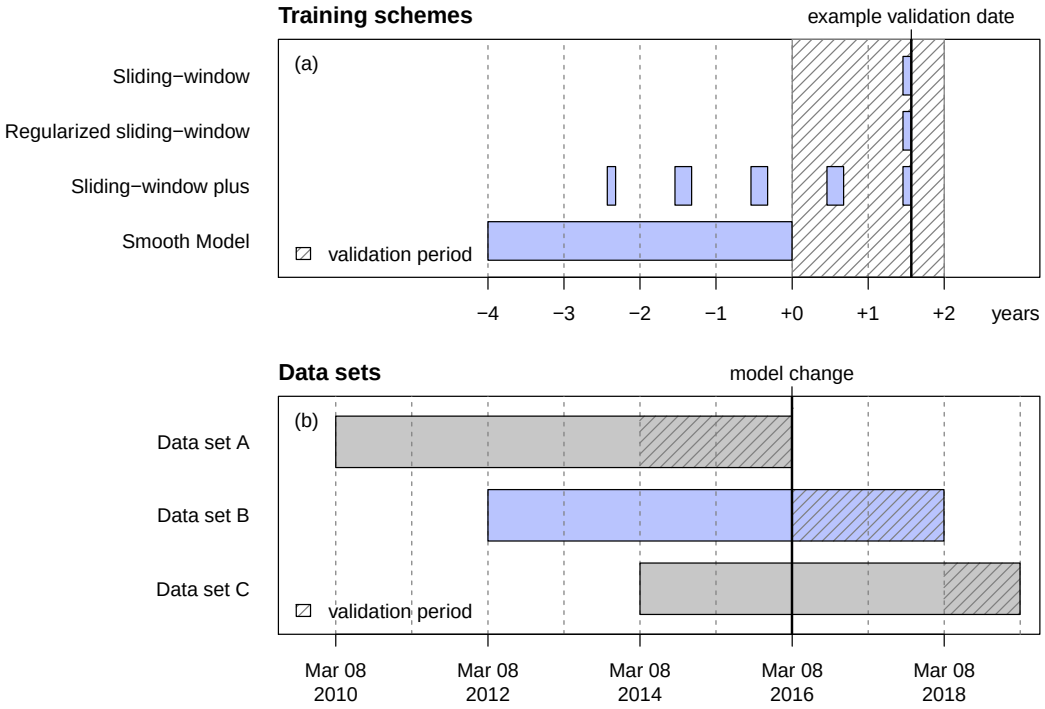

Figure 2. (a) Illustrative example of how the training data sets are composed for the four different time-adaptive training schemes. (b) Schematic overview of the training and validation data sets employed in this study with regard to the change in the horizontal resolution of the ECMWF EPS on March 8, 2016 (cycle 41r2). For training, up to four years of data are used in all data sets; for validation, two years of data are used for data sets A and B, and one year for data set C.

step. The computation time for the various sliding-window approaches is in the order of seconds, whereas the estimation of the *smooth model*, including full MCMC sampling, is in the order of minutes on a standard computer.

3 Results

This section assesses the performance of the different time-adaptive training schemes. First, the temporal evolution of the estimated coefficients are shown for two stations representative for one measurement site in the plains and one in mountainous terrain. Afterwards, the predictive performance of the training schemes is evaluated in terms of the CRPS conditional on the three data sets with and without [..[9]]the change in the horizontal resolution of the EPS (Fig. 2) and grouped for stations classified as topographically plain, mountain foreland, and alpine sites (Fig. 1).

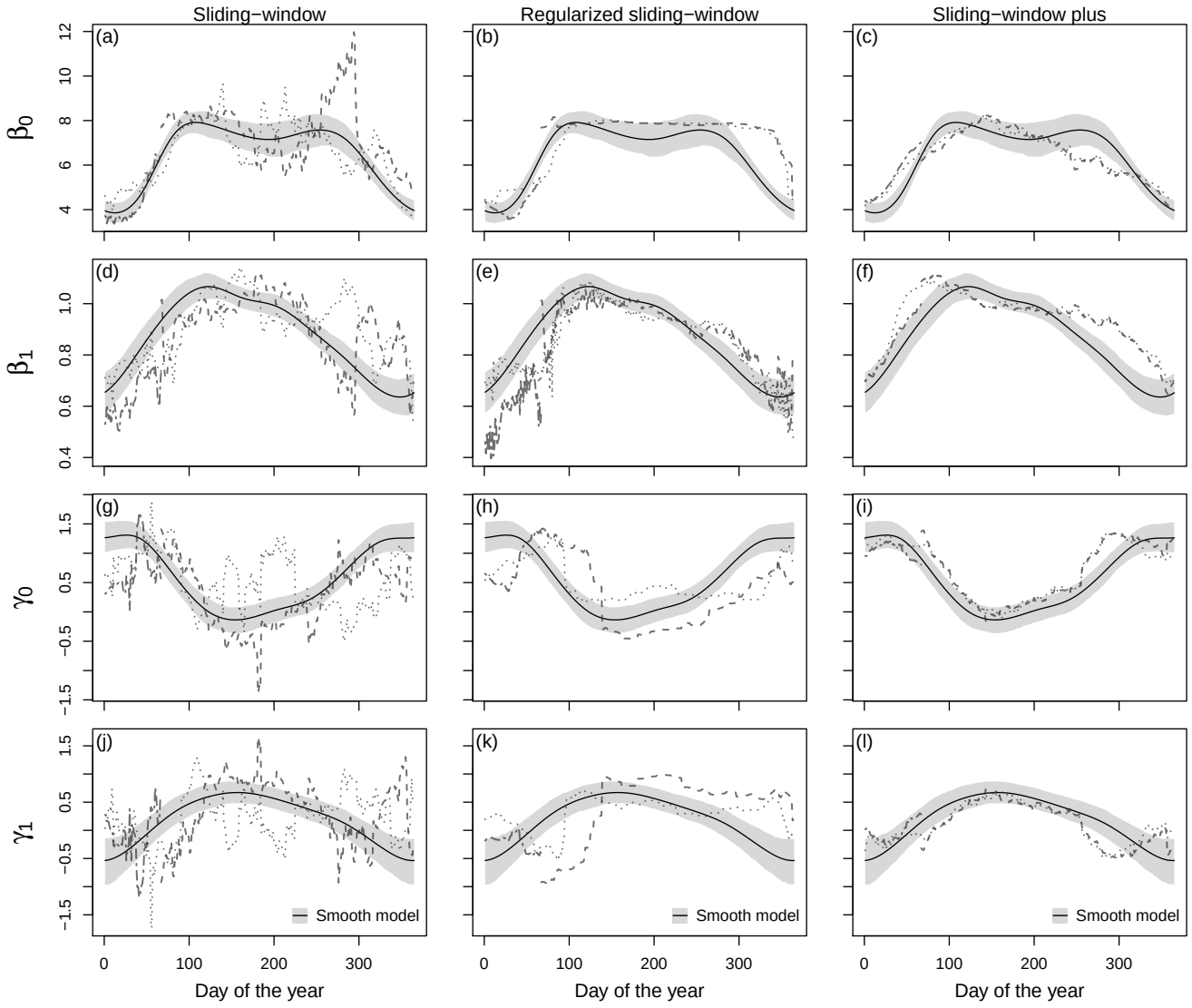

Figure 3. Temporal evolution of regression coefficients for the validation period in data set A for Innsbruck at forecast step $+36$ h (valid at 1200 UTC). The coefficient paths are shown for the coefficients β_0 (a–c) and β_1 (d–f) in the location parameter μ, and for the coefficients γ_0 (g–i) and γ_1 (j–l) in the scale parameter σ based on the *sliding-window*, *regularized sliding-window*, and *sliding-window plus* approach (dashed, from left to right) compared to the *smooth model* approach (solid line). The [..[10]]coefficient paths are plotted for the consecutive calendar years 2014, 2015, and 2016 as dashed, dotted, and two-dashed line, respectively. The grey shading represents the 95% credible intervals of the coefficients in the *smooth model* based on MCMC sampling.

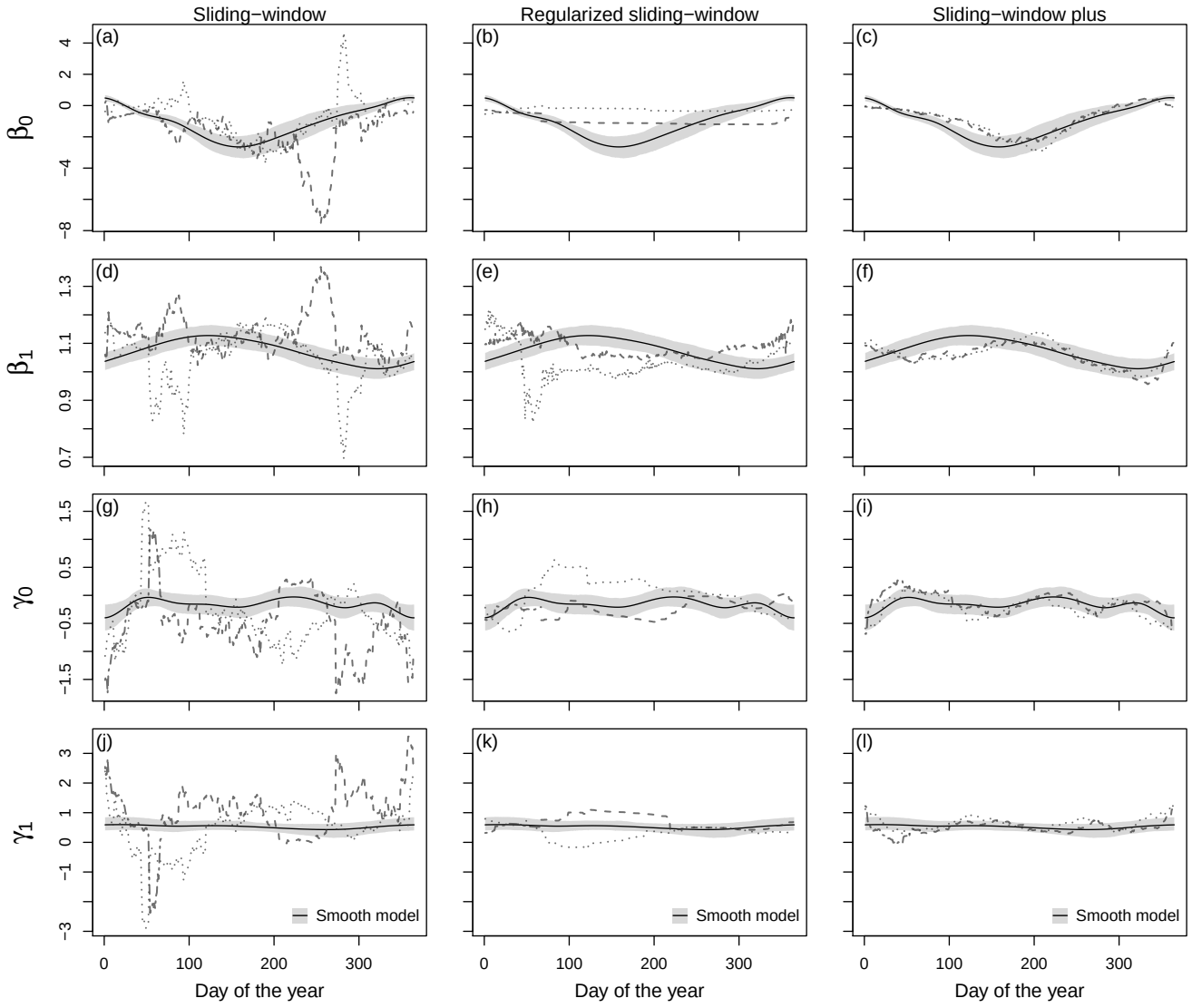

Figure 4. As Fig. 3, but for Hamburg at forecast step +36 h (valid at 1200 UTC).

3.1 Coefficient paths

Figure 3 shows the estimated coefficients for Innsbruck at forecast step +36 h conditional on the day of the year. The coefficient paths are plotted for the different time-adaptive training schemes for two years included in the validation period of data set A. The pronounced seasonal evolution of the coefficients for all training schemes [..[11]]shows that the EPS' forecast bias and

[9]removed: EPS changes

[11]removed: indicates

skill varies seasonally which makes a time-adaptive training scheme mandatory to capture these characteristics in the post-processing. [..[12]]During summer, a slope coefficient β_1 close to one in the location parameter μ and a high slope coefficient γ_1 in the scale parameter σ [..[13]]indicate a better performance of the EPS compared to the cold season.

In comparison to the other time-adaptive training schemes, the classical *sliding-window* approach (Fig. 3a, d, g, j) shows very strong outliers and an unstable temporal evolution for all coefficients with distinct differences during the two subsequent validation years; this is more pronounced for the scale parameter σ where the estimates seem to be more volatile than for the location parameter μ. All strategies extending the classical *sliding-window* approach smooth the temporal evolution of the coefficients to a certain extent while maintaining the overall seasonal cyclic pattern. For the *regularized sliding-window* approach (Fig. 3b, e, h, k), the stabilization strongly differs for the individual coefficients and some of the estimated coefficients seem to need rather long to adapt during the transition periods; the latter could indicate that a single iteration step might not be sufficient in this study. The coefficient paths for the *sliding-window plus* approach (Fig. 3c, f, i, l) and for the *smooth model* (Fig. 3a–l; solid line) look very similar with minor distortions during the cold season. Due to the definition of the *smooth model*, its coefficient paths show the most stable evolution but with the lowest ability to react to abrupt changes in the error characteristics.

For Hamburg (Fig. 4) by contrast to Innsbruck, the information content of the mean EPS temperature forecast is quite high throughout the year. This yields a lower bias correction and an almost one-to-one mapping of the ensemble mean to the location parameter μ indicated by a coefficient β_1 close to one. Despite the different post-processing characteristics, the temporal evolution of the coefficient paths is similar to the one for Innsbruck which confirms our previous findings: For the extended sliding-window approaches the coefficients have indeed very little seasonal variability, while for the classical *sliding-window* approach the coefficients show unrealistically strong fluctuations over time without a clear seasonal pattern (Fig. 4a, d, g, j). As for Innsbruck, the *regularized sliding-window* approach has a rather unrealistic stepwise evolution for some coefficients (Fig. 4b, e, h, k). The coefficient paths for the *sliding-window plus* approach (Fig. 4c, f, i, l) and the *smooth model* (Fig. 4; solid line) look comparable. These results support the bias-variance trade-off that regularizing or smoothing stabilizes the coefficient paths, while loosing the ability to rapidly react to temporal changes in the data.

3.2 Predictive performance

After the illustrative evaluation of the coefficients' temporal evolution for the different time adaptive training schemes, Fig. 5 shows aggregated CRPS skill scores for groups of five respective stations classified as topographically plain, mountain foreland, and alpine sites (Fig. 1) regarding the data sets A, B, and C (Fig. 2). In all panels the *regularized sliding-window* approach, the *sliding-window plus* approach, and the *smooth model* is compared to the classical *sliding-window* approach as a reference.

- For data set A, the *regularized sliding-window* approach shows only little improvements for the plain and foreland, and an overall performance loss for alpine stations. By contrast, the *sliding-window plus* and the *smooth model* approaches show distinct improvements over the classical *sliding-window* approach with largest values for alpine sites.

[12] removed: Apparently, the EPS temperature forecasts have a higher information content during summerwhich yields

[13] removed: for this period.

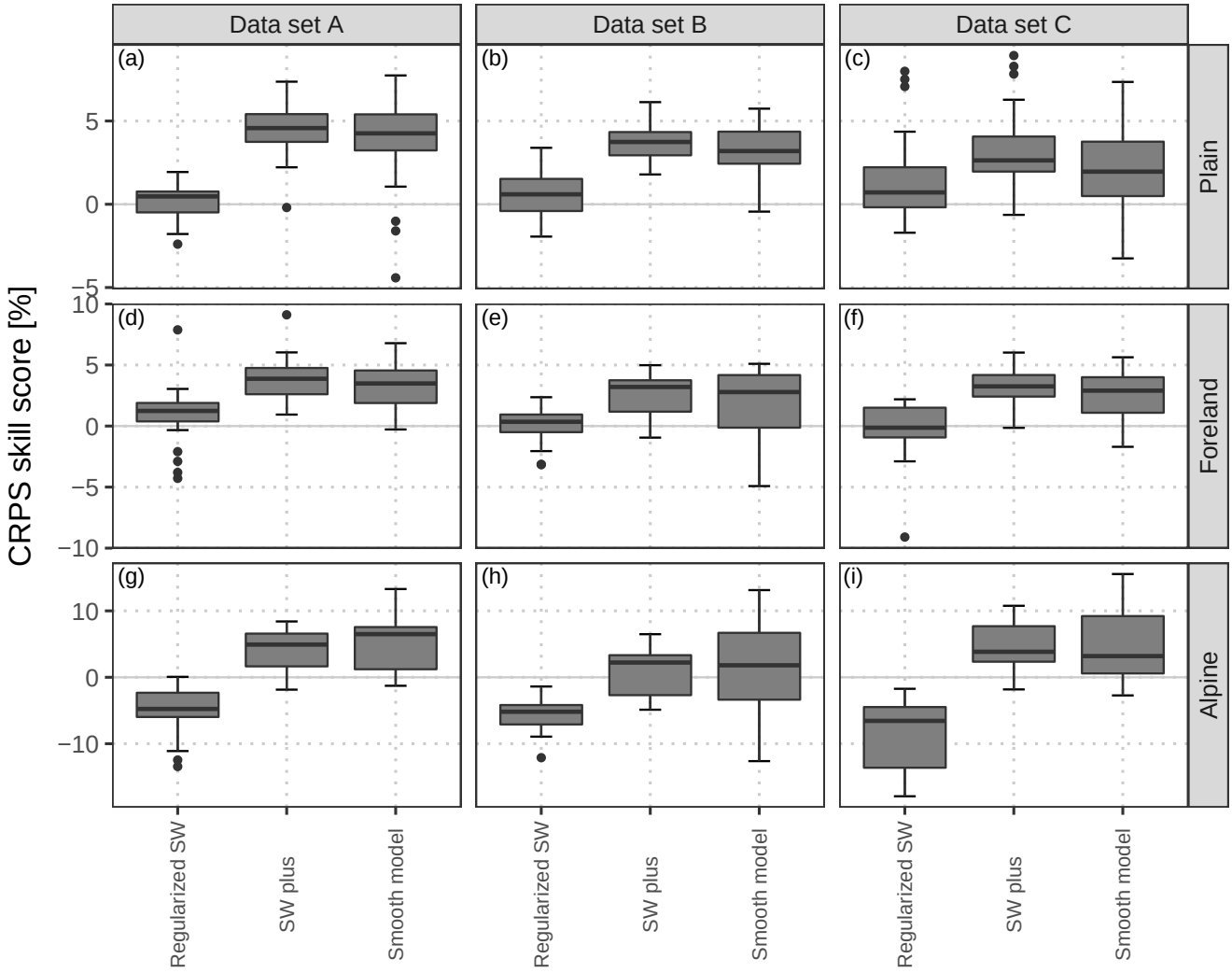

Figure 5. CRPS skill scores clustered into groups of stations located in the plain, in the mountain foreland near the Alps, and within mountainous terrain and for the out-of-sample validation periods according to the different data sets: Data set A without [..[14]]the change in the horizontal resolution of the EPS, data set B with [..[15]]the EPS change in between the training and the validation data sets, and data set C with [..[16]]the EPS change withing training data (Fig. 2). Compared are the different time-adaptive training schemes specified in Sect. 2.2 with the classical *sliding-window* approach as a reference; note that 'sliding-window' is abbreviated as SW in the figure. Each box-whisker contains aggregated skill scores over the forecast steps from +12 h to +72 h on a 12 hourly temporal resolution and over five respective weather stations (Fig. 1). Skill scores are in percent, positive values indicate improvements over the reference.

– For data set B at stations in the plains and foreland, the mean predictive skill behaves similarly to data set A, except that the *smooth model* shows a slightly larger variance. For alpine stations, the *regularized sliding-window* approach

performs slightly worse than in data set A, while the two approaches using training data over multiple years do no longer outperform the reference.

- For data set C at stations in the plains and foreland, the predictive skill is again similar to data set A with slight performance losses. For alpine stations, the *regularized sliding-window* approach shows even less skill as in data set B, while the two other approaches again outperform the *sliding-window* approach and are on a similar level as in data set A.

The validation of the different time-adaptive training schemes shows that the *sliding-window plus* approach and the *smooth model* perform overall similar and are clearly superior for all station types compared to the classical *sliding-window* approach. However, the *smooth model* shows the highest variance in the predictive performance in case of a change in the EPS, especially in mountainous terrain (data sets B and C); this is likely due to its reduced ability to adapt to temporal changes in the data. Furthermore, the validation shows that the *regularized sliding-window* approach seems to have difficulties in mountainous terrain and yields only minor improvements for plain and foreland sites.

4 Summary and conclusion

Non-homogeneous regression (NR) is a widely used method to statistically post-process ensemble weather forecasts. In its original version it was used for temperature forecasts employing a Gaussian response distribution, but over the last decade various statistical model extensions have been proposed for other quantities employing different response distributions or to enhance its predictive performance. When estimating NR models there is always a trade-off between large enough training data sets to get stable estimates and still allowing the statistical model to adjust to temporal changes in the statistical error characteristics of the data. Therefore, different training schemes with specific advantages and drawbacks have been developed as presented in this paper. To show a wide spectrum of possible approaches in a unified setup, we consider typical basic applications of the training schemes and refrain from more elaborate tuning or combinations.

The classical *sliding-window* approach has the advantage that no extensive training data set is required which allows the statistical model to adjust itself rapidly to changing forecast biases, for example in case of changes in the EPS. On the other hand, statistical models trained on a small training data set have typically large variance in the estimation of the regression coefficients, which can yield unstable wiggly coefficient paths. Additional regularization allows one to stabilize the evolution of the regression coefficients without losing the simplicity of the classical *sliding-window* approach. However, inappropriate settings of the optimizer as, e.g., unrealistic starting values or insufficient update steps, can quickly lead to incorrect coefficients. The alternative *sliding-window plus* strategy foregoes regularization but stabilizes the coefficients by using an extended training data set which includes data from the same season over several years. Compared to the classical approach the method requires historical data and partially loses its ability to rapidly adjust to changes in the error characteristics. The last approach presented in this paper can be seen as a generalization of the *sliding-window plus* approach. Rather than using a training data set centered around the date of interest, the *smooth model* makes use of all historical data in combination with cyclic regression splines which allows the coefficients to smoothly evolve over the year.

The differences between the methods presented can be seen in the coefficient paths shown in Fig. 3 and 4. The coefficients of the classical *sliding-window* approach show strong fluctuations and pronounced peaks throughout the year. Regularization allows to stabilize the evolution, however, strong step-wise changes in the coefficient paths still occur. The two methods using data from multiple years perform comparably similar with stable coefficient paths over the year. Figure 5 confirms that more stable estimates have a positive impact on the predictive performance. The *sliding-window plus* approach and the *smooth model* show an overall improvement of about 3–5 % (in median) over the classical *sliding-window* approach, while the *regularized sliding-window* only partially outperforms the *sliding-window* training scheme. [..[17]]Even in case of the model change chosen to demonstrate the effect of non-seasonal long-term changes on the coefficient estimates, the training schemes using multiple years of [..[18]]data are still superior to the ones using the most recent days only, even if they technically allow to adjust to the EPS change more rapidly.

To conclude, all four training schemes shown in this paper have their advantages in particular applications. If only short periods of training data are available ($<$ 1 year), the classical *sliding-window* approach may already provide sufficiently good estimates. However, as soon as one has access to longer historical data sets, the approaches using data from multiple years become superior due to a more stable coefficients' evolution over time which yields an overall improved performance. This even holds in case of the EPS change considered in this study, but may be different for other changes or EPSs. While the *sliding-window plus* is a natural extension of the classical *sliding-window* approach and, therefore, can be estimated by the same software, the *smooth model* approach can be seen as a generalization and only a single model has to be estimated for all seasons using all available data. The *smooth model* yields, by definition, the smoothest and most stable coefficient paths but with the lowest ability to adjust itself to a new error characteristic.

[17] removed: In case of a change in the EPS the approaches

[18] removed: training

Code availability. All computations are performed in R 3.6.1 (R Core Team, 2019): The statistical models using a sliding-window approach are based on the R package **crch** (Messner et al., 2016) employing a frequentistic maximum likelihood approach. The statistical models using a time-adaptive training scheme by fitting cyclic smooth functions are fitted with the R package **bamlss** (Umlauf et al., 2018). The package provides a flexible toolbox for distribution regression models in a Bayesian framework; introductory material can be found on http://BayesR.R-Forge.R-project.org/. The computation of the CRPS is based on the R package **scoringRules** (Jordan et al., 2019).

Author contributions. This study is based on the PhD work of MNL under supervision of GJM and AZ. The majority of the work for this study was performed by MNL with the support of RS. All the authors worked closely together in discussing the results and commenting on the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project was partly funded by the Austrian Research Promotion Agency (FFG, grant no. 858537) and by the Austrian Science Fund (FWF, grant no. P31836). Sebastian Lerch gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) through SFB/TRR 165 "Waves to Weather". We also thank the Zentralanstalt für Meteorologie und Geodynamik (ZAMG) for providing access to the data.

References

Baran, S. and Möller, A.: Bivariate Ensemble Model Output Statistics Approach for Joint Forecasting of Wind Speed and Temperature, Meteorology and Atmospheric Physics, 129, 99–112, https://doi.org/10.1007/s00703-016-0467-8, 2017.

Barnes, C., Brierley, C. M., and Chandler, R. E.: New approaches to postprocessing of multi-model ensemble forecasts, Quarterly Journal of the Royal Meteorological Society, https://doi.org/10.1002/qj.3632, accepted, 2019.

Demaeyer, J. and Vannitsem, S.: Correcting for Model Changes in Statistical Post-Processing – An approach based on Response Theory, Nonlinear Processes in Geophysics Discussions, 2019, 1–27, https://doi.org/10.5194/npg-2019-57, 2019.

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A.: Estimation Methods for Nonhomogeneous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood, Monthly Weather Review, 146, 4323–4338, https://doi.org/10.1175/MWR-D-17-0364.1, 2018.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annual Review of Statistics and Its Application, 1, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Hamill, T. M.: Practical Aspects of Statistical Postprocessing, in: Statistical Postprocessing of Ensemble Forecasts, edited by Vannitsem, S., Wilks, D. S., and Messner, J. W., pp. 187–217, Elsevier, https://doi.org/10.1016/C2016-0-03244-8, 2018.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation, Monthly Weather Review, 136, 2620–2632, https://doi.org/10.1175/2007MWR2411.1, 2008.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Statistical Science, 1, 297–310, https://www.jstor.org/stable/2245459, 1986.

Hemri, S., Haiden, T., and Pappenberger, F.: Discrete Postprocessing of Total Cloud Cover Ensemble Forecasts, Monthly Weather Review, 144, 2565–2577, https://doi.org/10.1175/mwr-d-15-0426.1, 2016.

Henzi, A., Ziegel, J. F., and Gneiting, T.: Isotonic Distributional Regression, arXiv 1909.03725, arXiv.org E-Print Archive, http://arxiv.org/abs/1909.03725, 2019.

Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with scoringRules, Journal of Statistical Software, 90, 1–37, https://doi.org/10.18637/jss.v090.i12, 2019.

Junk, C., Monache, L. D., and Alessandrini, S.: Analog-Based Ensemble Model Output Statistics, Monthly Weather Review, 143, 2909–2917, https://doi.org/10.1175/mwr-d-15-0095.1, 2015.

Klein, N., Kneib, T., Klasen, S., and Lang, S.: Bayesian Structured Additive Distributional Regression for Multivariate Responses, Journal of the Royal Statistical Society: Series C (Applied Statistics), 64, 569–591, https://doi.org/10.1111/rssc.12090, 2014.

Lang, M. N., Mayr, G. J., Stauffer, R., and Zeileis, A.: Bivariate Gaussian Models for Wind Vectors in a Distributional Regression Framework, Advances in Statistical Climatology, Meteorology and Oceanography, 5, 115–132, https://doi.org/10.5194/ascmo-5-115-2019, 2019.

Lerch, S. and Baran, S.: Similarity-based semilocal estimation of post-processing models, Journal of the Royal Statistical Society: Series C (Applied Statistics), 66, 29–51, https://doi.org/10.1111/rssc.12153, 2017.

Messner, J. W., Mayr, G. J., and Zeileis, A.: Heteroscedastic Censored and Truncated Regression with crch, The R Journal, 8, 173–181, https://doi.org/10.32614/RJ-2016-012, 2016.

Möller, A., Spazzini, L., Kraus, D., Nagler, T., and Czado, C.: Vine Copula Based Post-Processing of Ensemble Forecasts for Temperature, arXiv 1811.02255, arXiv.org E-Print Archive, http://arxiv.org/abs/1811.02255, 2018.

NASA JPL: NASA Shuttle Radar Topography Mission Global 30 Arc Second [Data Set], NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL30.002, 2013.

Palmer, T. N.: The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades, Quarterly Journal of the Royal Meteorological Society, 128, 747–774, https://doi.org/10.1256/0035900021643593, 2002.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U.: Forecasting Wind Gusts in Winter Storms Using a Calibrated Convection-Permitting Ensemble, Quarterly Journal of the Royal Meteorological Society, 144, 1864–1881, https://doi.org/10.1002/qj.3380, 2018.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2019.

Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, Monthly Weather Review, 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., and Wernli, H.: Flow-dependent reliability: A path to more skillful ensemble forecasts, Bulletin of the American Meteorological Society, 99, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1, 2018.

Scheuerer, M.: Probabilistic Quantitative Precipitation Forecasting Using Ensemble Model Output Statistics, Quarterly Journal of the Royal Meteorological Society, 140, 1086–1096, https://doi.org/10.1002/qj.2183, 2014.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A.: Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain, The Annals of Applied Statistics, 13, 1564–1589, https://doi.org/10.1214/19-AOAS1247, 2019.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, Monthly Weather Review, 144, 2375–2393, https://doi.org/10.1175/mwr-d-15-0260.1, 2016.

Umlauf, N., Klein, N., and Zeileis, A.: BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond), Journal of Computational and Graphical Statistics, 27, 612–627, https://doi.org/10.1080/10618600.2017.1407325, 2018.

Vogel, P., Knippertz, P., Fink, A. H., Schlueter, A., and Gneiting, T.: Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa, Weather and Forecasting, 33, 369–388, https://doi.org/10.1175/waf-d-17-0127.1, 2018.

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R.: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging, Monthly Weather Review, 135, 1364–1385, https://doi.org/10.1175/MWR3347.1, 2007.

Wood, S. N.: Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, https://doi.org/10.1201/9781315370279, 2017.

Supplement A: Post-processing of daily precipitation sums

Moritz N. Lang Universität Innsbruck
Sebastian Lerch Karlsruher Institut für Technologie
Georg J. Mayr Universität Innsbruck
Thorsten Simon Universität Innsbruck
Reto Stauffer Universität Innsbruck
Achim Zeileis Universität Innsbruck

Abstract

The study presented in the manuscript "Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression" compares three widely-used training approaches with the classical sliding-window model for the application of post-processing near-surface air temperature forecasts across Central Europe. While the normal distribution is typically employed for post-processing air temperatures, this supplement extends the study by post-processing daily precipitation sums using a zero left-censored Gaussian distribution. Despite the different characteristics of daily precipitation sums and the alternative response distribution, the results are very similar to the ones for 2 m temperature forecasts and, hence, nicely support the conclusions given in the paper.

Keywords: Non-homogeneous regression, training data, sliding training window, post-processing, regression splines, ensemble forecasts, daily precipitation sums.

1. Introduction

This supplement extends the comparison study presented in the main manuscript by performing the same evaluation for daily precipitation sums employing an alternative response distribution.

As in the main manuscript the temporal evolution of the estimated coefficients is shown for two stations with different site characteristics followed by the analysis of the predictive performance. The results of all training schemes is evaluated in terms of the continuous ranked probability score (CRPS) conditional on the three data sets with and without the change in the horizontal resolution of the ensemble prediction system (EPS) on March, 8, 2016, as well as grouped for stations classified as topographically plain, mountain foreland, and alpine site. As a reminder, the different training-schemes are briefly summarized as follows:

- *Sliding-window:* The classical *sliding-window* approach as introduced by Gneiting, Raftery, Westveld III, and Goldman (2005) uses solely the n most recent days prior to the day of interest as training data to estimate the statistical models.
- *Regularized sliding-window:* The regularized adaption of the classical *sliding-window*

approach stabilizes the estimation based on early stopping in statistical learning. In this study, it is applied in its original version where the coefficients of the previous day are used as starting values and the optimizer is stopped after a single iteration (Scheuerer 2014).

- *Sliding-window plus:* In order to stabilize the coefficient estimates and to address seasonal effects, data from previous years are additionally included in the training data set. In contrast to the classical *sliding-window* approach, the most recent n days prior to estimation and a respective $(2n + 1)$ days interval centered around the day of interest over the previous years available are used to estimate the coefficients (Vogel, Knippertz, Fink, Schlueter, and Gneiting 2018; Möller, Spazzini, Kraus, Nagler, and Czado 2018).

- *Smooth model:* Rather than adapting the training data set, the *smooth model* makes use of all historical data in combination with cyclic regression splines which allows the coefficients to smoothly evolve over the year.

In comparison to the main manuscript, to account for potentially long periods without precipitation at a specific site, we use a training length of $n = 80$ days in the sliding-window approaches for post-processing daily precipitation sums. This is in the order of common choices in the literature (e.g., Baran and Nemoda 2016) and exactly two times the length employed for post-processing 2 m temperature forecasts as presented in the main manuscript.

2. Methodology and data

To account for non-negative values, the large fraction of zero observations, and the heavy-tailed distribution of precipitation, we proceed as proposed by Stauffer, Mayr, Messner, Umlauf, and Zeileis (2017a): We power-transform observed and modeled daily precipitation sums (member-by-member) with an ad-hoc chosen power parameter of 2. This transformed precipitation y can be assumed to follow a zero left-censored Gaussian distribution $\mathcal{N}_0$,

$$y \sim \mathcal{N}_0(\mu, \sigma). \tag{1}$$

As in the manuscript, the two distribution parameters location μ and scale σ are expressed by the ensemble mean m and ensemble variance or standard deviation s, respectively:

$$\mu = \eta_\mu = \beta_0 + \beta_1 \cdot m, \tag{2}$$

$$\log(\sigma) = \eta_\sigma = \gamma_0 + \gamma_1 \cdot s, \tag{3}$$

with $\beta_\bullet$ and $\gamma_\bullet$ being the corresponding intercept and slope coefficients.

As covariates, we employ the ensemble mean m and the ensemble standard deviation s of bilinearly interpolated power-transformed daily precipitation sum forecasts issued by the global 50-member EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF). Ensemble forecasts and corresponding observations are considered at 15 measurement sites located across Austria, Germany, and Switzerland. The data comprises three groups of five stations located either in plains, mountain foreland, and within mountainous terrain. An overview of the study area is provided in the main manuscript in Fig. 1. For training and validation, we assess forecast steps from +24 h to +72 h ahead on a 24 hourly temporal resolution for the EPS run initialized at 0000 UTC and use the same data period employed in the manuscript, from March 8, 2010 to March 7, 2019.

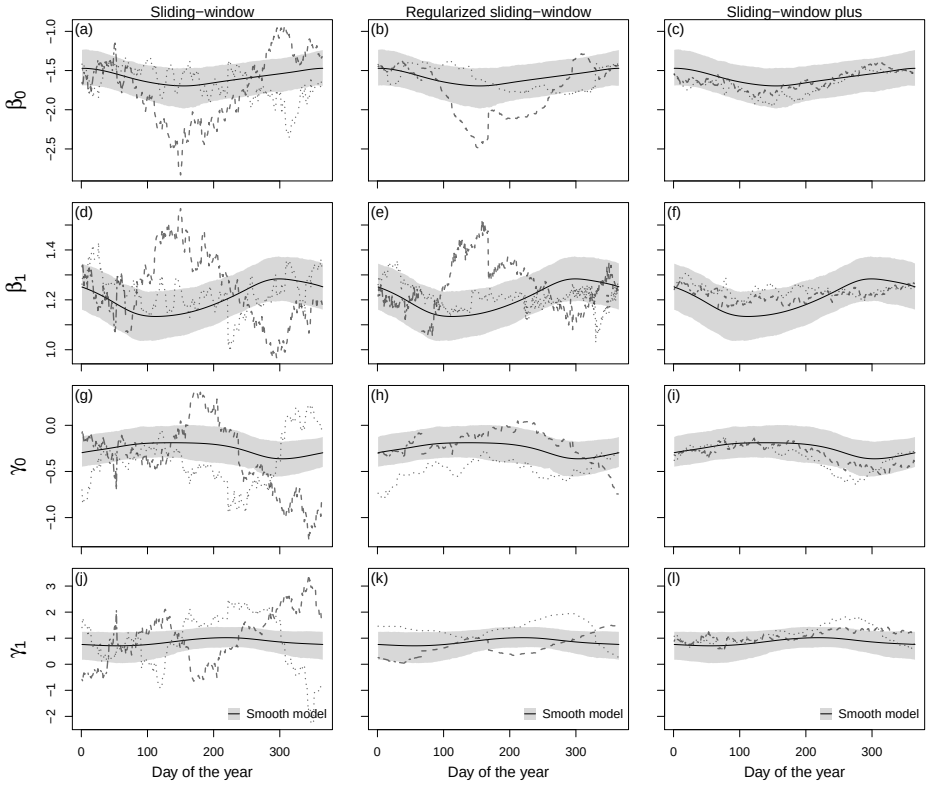

Figure 1: Temporal evolution of regression coefficients for the validation period in data set A, cf. Fig. 2 in the main manuscript, for Innsbruck at forecast step +24 h (valid at 0000 UTC). The coefficient paths are shown for the coefficients β_0 (a–c) and β_1 (d–f) in the location parameter μ, and for the coefficients γ_0 (g–i) and γ_1 (j–l) in the scale parameter σ based on the *sliding-window*, *regularized sliding-window*, and *sliding-window plus* approach (dashed, from left to right) compared to the *smooth model* approach (solid line). The coefficient paths are plotted for the consecutive calendar years 2014, 2015, and 2016 as dashed, dotted, and two-dashed line, respectively. The grey shading represents the 95% credible intervals of the coefficients in the *smooth model* based on MCMC sampling.

3. Results

In comparison to the main manuscript, the presented results compare the performance of the different time-adaptive training schemes for post-processing daily precipitation sums. Figure 1 and Fig. 2 illustrate the temporal evolution of the estimated coefficients shown for two stations representative for one measurement site in the plains (Hamburg) and one in mountainous terrains (Innsbruck). Figure 3 shows the predictive performance of the training schemes evaluated for three groups of stations with different site characteristics and in terms of the CRPS conditional on the three data sets with and without the change in the horizontal resolution of the EPS, as defined in Sect. 2.3.2 of the main manuscript. Due to the employed power-transformation, CRPS values are computed by quantile sampling with $n = 1000$; for a more detailed description compare Stauffer, Umlauf, Messner, Mayr, and Zeileis (2017b).

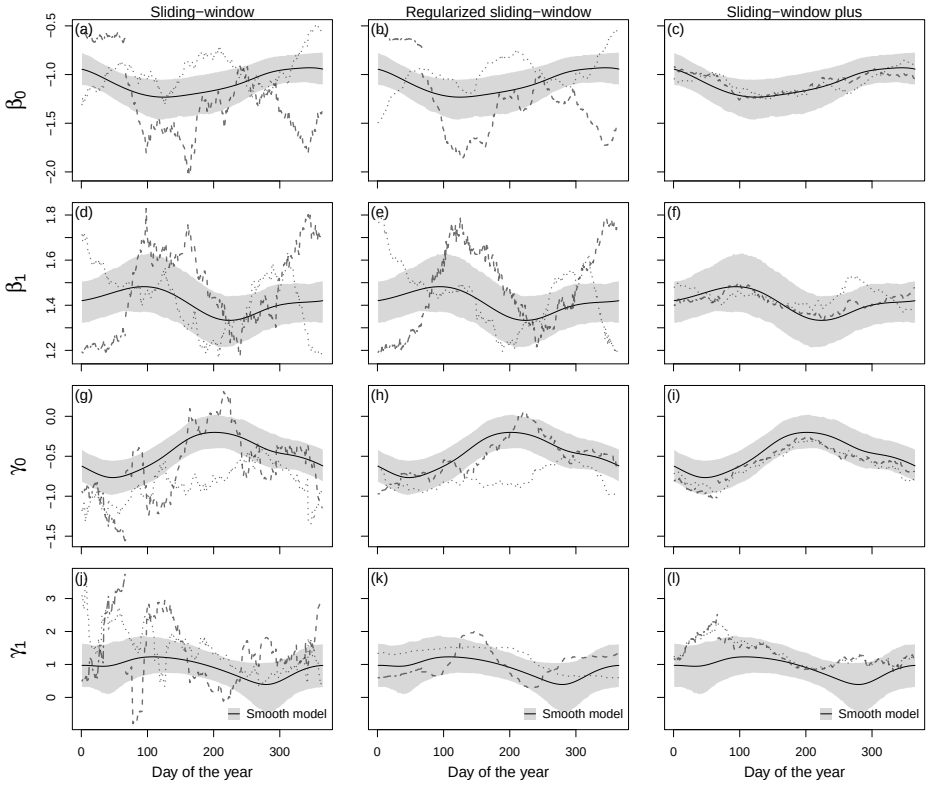

Figure 2: As Fig. 1, but for Hamburg at forecast step +24 h (valid at 0000 UTC).

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).

4. Conclusion

This supplement provides a comparison evaluating the different time-adaptive training schemes for post-processing daily precipitation sums. To account for non-negative values, the large

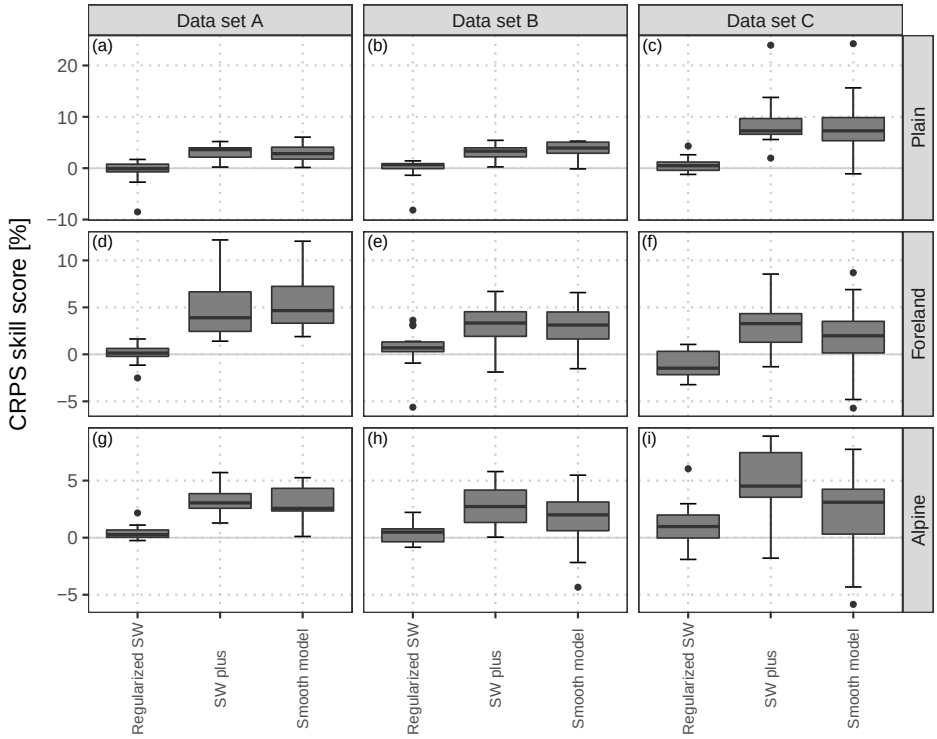

Figure 3: CRPS skill scores clustered into groups of stations located in the plain, in the mountain foreland near the Alps, and within mountainous terrain and for the out-of-sample validation periods according to the different data sets: Data set A without the change in the horizontal resolution of the EPS, data set B with the EPS change in between the training and the validation data sets, and data set C with the EPS change withing training data (cf. Fig. 2 in the main manuscript). Compared are the different time-adaptive training schemes specified in Sect. 2 with the classical *sliding-window* approach as a reference; note that 'sliding-window' is abbreviated as SW in the figure. Each box-whisker contains aggregated skill scores over the forecast steps from +24 h to +72 h on a 24 hourly temporal resolution and over five respective weather stations (cf. Fig. 1 in the main manuscript). Skill scores are in percent, positive values indicate improvements over the reference.

fraction of zero observations, and the strongly positively skewed characteristics of daily precipitation sums, we employ a power-transformation to both the observations and to each ensemble member, and use the zero left-censored Gaussian distribution in the framework of non-homogeneous regression (Stauffer *et al.* 2017a).

Despite the different characteristics of daily precipitation sums and the alternative response distribution, the results are very similar to the ones for 2 m temperature forecasts presented in the main manuscript. This shows that the findings for 2 m temperature can also be transferred to other quantities using different model assumptions and, hence, nicely support the conclusions given in the paper.

References

Baran S, Nemoda D (2016). "Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting." *Environmetrics*, **27**(5), 280–292. doi:10.1002/env.2391.

Gneiting T, Raftery AE, Westveld III AH, Goldman T (2005). "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation." *Monthly Weather Review*, **133**(5), 1098–1118. doi:10.1175/MWR2904.1.

Möller A, Spazzini L, Kraus D, Nagler T, Czado C (2018). "Vine Copula Based Post-Processing of Ensemble Forecasts for Temperature." *arXiv 1811.02255*, arXiv.org E-Print Archive. URL http://arxiv.org/abs/1811.02255.

Scheuerer M (2014). "Probabilistic Quantitative Precipitation Forecasting Using Ensemble Model Output Statistics." *Quarterly Journal of the Royal Meteorological Society*, **140**(680), 1086–1096. doi:10.1002/qj.2183.

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. doi: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. doi:10.1175/MWR-D-16-0260.1.

Vogel P, Knippertz P, Fink AH, Schlueter A, Gneiting T (2018). "Skill of Global Raw and Postprocessed Ensemble Predictions of Rainfall over Northern Tropical Africa." *Weather and Forecasting*, **33**(2), 369–388. doi:10.1175/waf-d-17-0127.1.

Affiliation:

Moritz N. Lang
Department of Statistics
Department of Atmospheric and Cryospheric Sciences
Universität Innsbruck
6020 Innsbruck, Austria
E-mail: moritz.lang@uibk.ac.at