Final author response

Submission “Latent Linear Adjustment Autoencoders v1.0: A novel method for estimating and emulating dynamic precipitation at high resolution”

We thank both referees for their insightful comments, helpful feedback and their positive evaluation. We address all points in detail below where we show our reply to reviewer comments in red for ease of exposition.

Review 1

Major comments:

1. I don’t understand why you train your autocoder on 1955-2070 data. There is a chance that you include some thermodynamical signal in the precipitation field when you minimise $Y - Y\hat{}_X$. I understand that you detrend the SLP EOF time series, but you don’t detrend precipitation. Why not training on 1955-1995 and potentially use more members to have the same amount of data?

Thank you for raising these important points. We agree that there is a chance to include some thermodynamical signal under a long training period (although SLP is detrended). We also agree that it is important for applications to understand the sensitivity to (i) the training period choice, (ii) the amount of training data as well as (iii) the sensitivity to different detrending approaches. Hence, we perform the following analyses:

(1) **Sensitivity test to training data amount and training period**:

We train on a shorter period from 1955-2020 (as suggested by referee 2), using the same ensemble members as in the first submission (i.e., equivalent to a ~43% reduction in the training data, but more importantly, restricting the training to a period with relatively modest precipitation change). We then reproduce the dynamical adjustment analysis with this model trained on (i) this shorter time period; and (ii) less data. We find that all performance measures (i.e., mean squared error, etc.) indicate very robust results with respect to these changes in the input data. In particular, the dynamical adjustment analysis based on the shorter training period reveals almost identical results as compared to the longer 1955-2070 training period. That is, the residual variability is much closer to the ensemble mean forced response (see plot reproduced below: Fig. 1). This sensitivity analysis thus provides support that our method is robust to (i) a shorter time period and (ii) less training data points. The detailed results can be found in the Supplement of the revised manuscript.

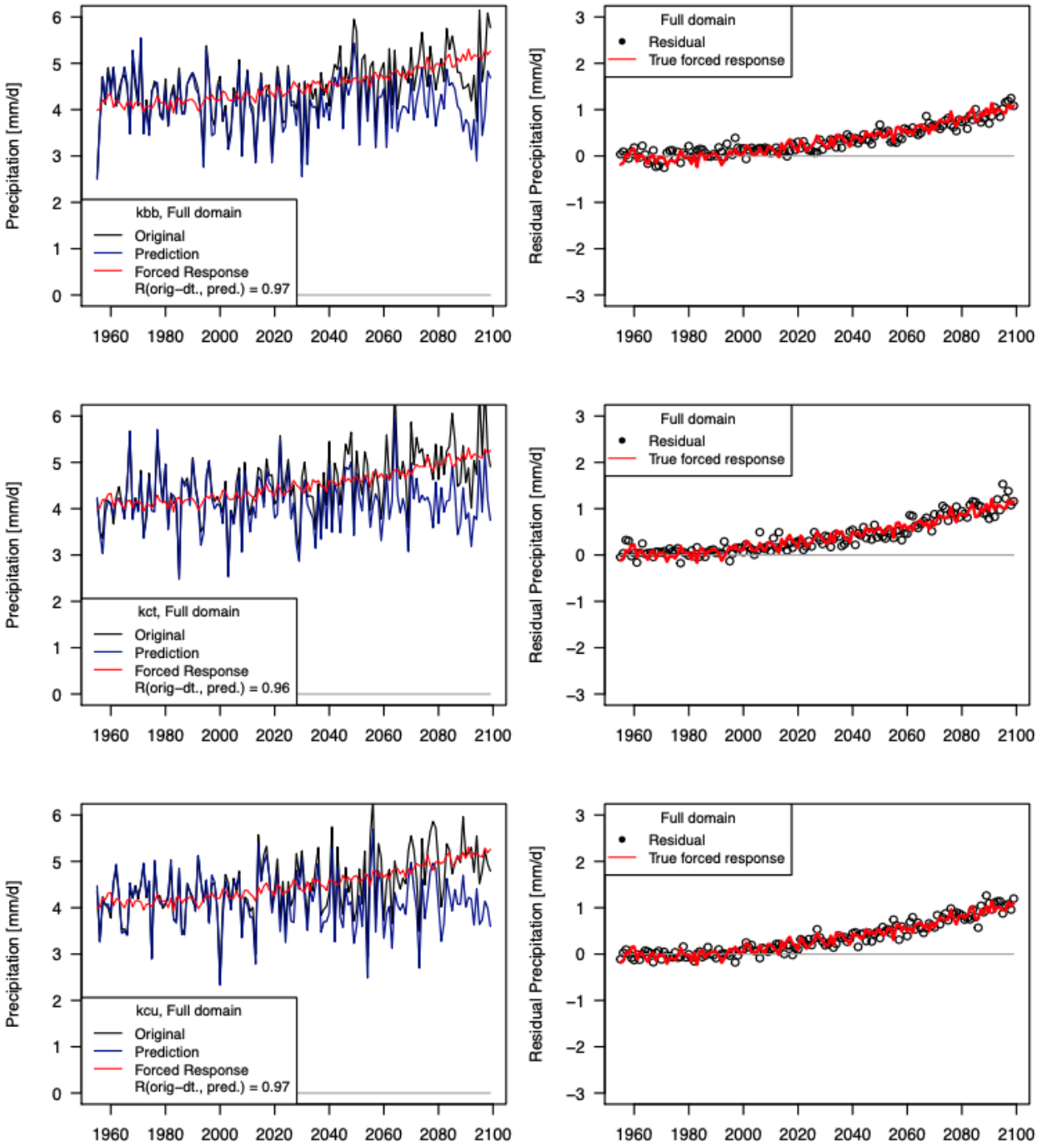

Figure 1: Dynamical adjustment analysis for the Latent Linear Adjustment Autoencoder (LLAAE) model trained only on the period 1955-2020 and thus with 43% fewer training data points. Compared to Fig. 6 in the main paper, which shows the same analysis for the LLAAE model with more training data (1955-2070), the results shown here are almost identical.

(2) **Trend removal sensitivity test**:

As correctly pointed out by the reviewer, the question of whether and how to detrend prior to dynamical adjustment is open, somewhat subjective, and often discussed as an inherent subjective choice/uncertainty in dynamical adjustment papers (see, e.g. Deser et al. 2016, or Lehner et al. 2017, and Lehner et 2018, for a discussion about trend removal). We agree that more discussion on this point is needed in the revised manuscript.

Forced changes in European winter SLP are highly uncertain, and models disagree on the sign and patterns of forced circulation change (Fereday et al. 2018) - although a northward shift in storm tracks and a dynamical extension of the subtropical dry zones is generally expected but not supported by all models (Fereday et al. 2018). Thermodynamic aspects are typically considered more robust across models (Shepherd et al. 2014; Fereday et al. 2018).

As pointed out by the reviewer, we have orthogonalized SLP EOF time series w.r.t. the **ensemble-mean SLP change** over time (i.e., a very simplistic but generic "detrending"). Our main motivation to use this somewhat simplistic detrending approach in our proof-of-concept study was to avoid that the ML method would take hypothetical dynamical changes to predict thermodynamical trends in precipitation (as the reviewer correctly pointed out). In other words, our goal was to estimate daily precipitation variability at high resolution *in absence of SLP* changes; that is introducing LLAAEs as a versatile tool for estimating a high-resolution precipitation field based on a coarse-resolution sea level pressure field.

Our analysis shows that the residuals match the ensemble mean very well (Fig. 6 in the main manuscript). Hence, if a trend signal would be included in the prediction of the precipitation field (e.g., due to hypothetical remaining trend artefacts in the pressure field), this effect is likely to be small because the residuals match the ensemble mean (forced) trend very well.

However, in addition to the results so far, we test an alternative simple detrending approach, where SLP is not detrended, but where we detrend precipitation using a simple LOESS smoother, fitted on the ensemble (seasonal) means at every location and subtracted from every day individually. (Furthermore, we here use the shorter 1955-2020 period for training the model.) For the dynamical adjustment analysis, we then compute the residuals based on the non-detrended precipitation data and our predictions (from the model trained on the detrended precipitation data; see Fig. 2). This analysis suggests that this approach to detrend precipitation is too simplistic since the residuals of the dynamical adjustment analysis underestimate forced changes (the ensemble mean) to some extent (Fig. 2). There are several possible reasons for this:
(1) Precipitation change cannot be modelled by a single additive mean change across the whole distribution. For instance, precipitation change is known to increase the variance of the precipitation distribution (Pendergrass et al. 2017). Hence, by subtracting the estimated, seasonally averaged precipitation trend we may have not fully removed the trend for wet days. Developing a more refined approach to remove the forced precipitation changes from daily data is non-trivial and beyond the scope of this work, but will be addressed in future work.

(2) There may be some dynamically-induced changes in precipitation, but it would be hard to evaluate this without any additional simulations where dynamical effects and thermodynamical effects could be separated.

Overall, we conclude that our simple SLP detrending (without detrending precipitation) is a useful approach for introducing LLAAEs as a versatile tool for dynamical adjustment, as demonstrated by the fact that the residuals of individual ensemble members after dynamical adjustment match the ensemble mean trend of precipitation very well (e.g., Fig. 6 in the main manuscript). However, we acknowledge that considerations around whether and how to detrend the data prior to dynamical adjustment are crucial, especially for real-world applications. We discuss this in the revised manuscript, and we acknowledge that more work is needed to fully understand the effect of detrending choices, but which goes beyond the scope of the present study.

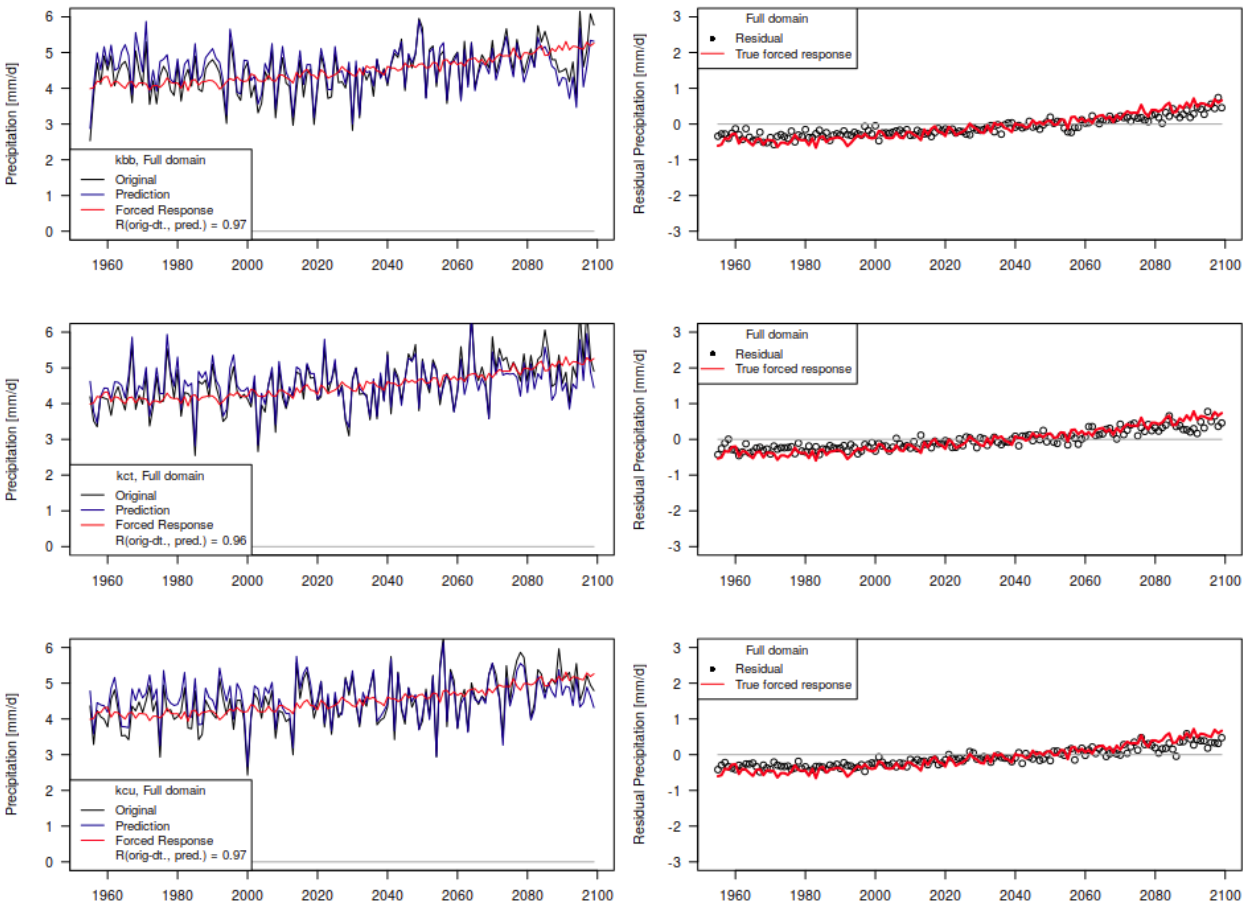

Figure 2: Dynamical adjustment analysis for LLAAE model trained only on 1955-2020 period and with detrending only precipitation using a LOESS smoother.

2. It would be good to know the minimal amount of data needed to train the algorithm. Indeed, if 1955-2070 daily data from a 9 member ensemble is needed to train the algorithm, then it would be cheaper to directly calculate the forced response from this 9-member ensemble (see comments below on Fig. 8) without dynamical adjustment. Ideally, one would like to dynamically adjust expensive simulations which cannot be run for long periods of time (e.g. a few decades).

We provide some analysis in that direction by limiting the training period to 1955-2020 (see above). This reduction in training data does not have a noticeable effect on the performance of the model (see above). In general, "the minimal amount of data" will depend on one's requirements of how to use the method. We expect the performance of the method to decrease gradually when further reducing the amount of training data.

We agree with the reviewer that, as machine learning algorithms are known to require rather large amounts of training data, "proving" the case of autoencoder dynamical adjustment in a large ensemble may not be as straightforward (as correctly pointed out by the reviewer, we used nine ensemble members for training, which we could have used instead for calculating a 9-member ensemble average). However, we anticipate the ultimate applications of autoencoder-based dynamical adjustment not on a large ensemble (where the forced response is "known" anyways, to some extent), but instead on simulations with models where only one (or very few) ensemble members may be available. Hence, our present manuscript was intended only as a proof-of-concept of the method within a large ensemble. As the next step, we envision the application to different climate models (e.g. training on a large ensemble or multiple large ensembles, and application of the dynamical adjustment to models for which only a few simulations exist), and with ultimate application of the trained autoencoders on reanalysis SLP data. This would allow us to leverage the available data from climate model simulations while applying the method in a context where a direct calculation of the (x-member) ensemble mean is not possible. We discuss and clarify this point in the revised manuscript.

On the examples of application:

1. I am convinced by the use of the new tool for dynamical adjustment on a large domain and seasonal scale (Fig. 6), this seems to be very successful, even with only 1 member. This is quite impressive. For more detailed spatial scales however, it is less successful and I guess from extrapolating Fig. 8 that using 7 or 8 members for the "traditional runs" (out of 50) outperforms the dynamical adjustment. I would like to see more discussion on this in the text and I think that Fig. 8 could be improved with a few changes:

    - extend the x axis to at least 10 members, to see when a "traditional averag- ing" outperforms the dynamical adjustment (this implies performing dynamical adjustment on more holdout members).

Thank you for these suggestions, and for pointing out that Fig. 8 was thus far a bit unclear. We will improve Fig. 8 in a revised manuscript as suggested - extending to at least 10 members (possibly to 15 or 20).

- you plot only one value and it does not correspond to the one in the text (line 255), I presume for 1 member you can have 41 different values (excluding the training set), so you can add median + inter-quartile range / sqrt(number of samples), so that one knows if the difference is statistically significant, but I presume so.

The number in Fig. 8 does not correspond with the text, because the figure illustrates RMSE's for the spatial average precipitation trend (only land grid cells), whereas the number quoted in the text refers to the RMSE averaged across all individual grid cells (i.e. the number in the text corresponds to Fig. 7, lowest middle panel). We will make this distinction very clear in the revised manuscript, i.e., clarifying the grid-cell based error analysis (Fig. 7) vs. the domain average analysis in Fig. 8. In addition, we use all 41 holdout members to better describe the variability/uncertainty of the reconstruction errors.

- Add details to the caption. I presume that it shows the RMSE of 50y trend maps calculated by averaging n members compared to 50y trends using 50 member average. It is not very clear from the caption.

Yes, exactly. We will improve the caption accordingly.

2. The tool is successful for seasonal means. Can you comment on the potential use of this tool for assessing trends in extreme precipitation, for which regional models are more trustworthy than global models? The prediction in precipitation fields seems smoothed out compared to original fields. And not taking into account thermodynamical fields as predictors may be limiting the representation of extremes, even in a present-day context.

Extreme precipitation is important and there is a demand for information about these events at spatial resolution essentially as high as possible. Our autoencoder may be able to fill an important gap in constructing extreme events in that it can reconstruct the dynamical component of extreme precipitation events (at least, the component proportional to surface pressure). Estimating the thermodynamic component is generally more straightforward than the dynamical component, and it may be possible to estimate it with other more straightforward approaches, particularly in winter. The reviewer is correct though, that the autoencoders, similar to other statistical/ML techniques, have a tendency to "smooth out" predictions (and thus probably underpredict the most extreme precipitation days). However, the technique may still be an improvement over existing alternatives: the resulting smoothing may still be less than the effective smoothing that occurs at the coarse resolution climate models. Rigorous evaluation of this application is, however, beyond the scope of this manuscript.

We agree with the reviewer that it is important to clarify these aspects related to potential extension toward extreme events more, and we will add a short discussion about future work into the Conclusion section.

3. Regarding the weather generator, I do struggle to exactly understand the novelty of your method. If I understand correctly, you are bootstrapping the time series of EOFS, but keeping each daily EOF set as it is, so you are not “creating” new pressure patterns, just shuffling them. One could do this directly by shuffling daily precipitation maps in the same way. I agree that one would need 150 years of present-day data instead of simulations with evolving greenhouse gases, but this is easily achieved these days. It is interesting that you show that shuffling 150 years of data seems as good as running several members, at least for the bulk of precipitation distribution. I wouldn’t think this is true for extremes. I think the use for dynamical adjustment has much more potential than the weather generator.

   I would suggest to reduce this section to have more space in the article for a figure to reply to my point 2 about the method.

Thank you for raising this concern. Since both referees suggested shortening this section, we have decided to follow this advice. Hence, we remove the section on the weather generator from our manuscript in order to focus on the method introduction and dynamical adjustment illustration, but we mention/discuss the possibility of weather generators here and we will expand on it elsewhere.

To answer your questions, note that the emulator generates dynamically-induced variability in the daily precipitation fields only. Hence, this cannot be achieved by using daily precipitation fields directly (additionally, note that we draw a bootstrap sample which is not equivalent to shuffling the data points).

Lastly, we agree that the performance for extremes is likely to be worse.

Minor comments:

Fig. 7, 10, 11: the scatter plots are saturated, it may be better to plot a gaussian kernel density estimation https://seaborn.pydata.org/generated/seaborn.kdeplot.html

We will improve the scatter plots in Figs. 7, 10, 11 in the revised manuscript.

Most figures with blue shading only: I find the continuous colour shading difficult, it may be best to reduce the number of colour levels used. One could also potentially use a sequential colour

map like terrain_r for precipitation fields. It will make figures more readable and may reduce the need to show square root precipitation fields, which are less intuitive.

Thank you for the suggestion. We have experimented with different color maps and different numbers of color levels but have not found alternative settings that yielded better figures.

Fig. 9: remove the numbers on it, you are not using them in the article.

This is correct, we will remove the numbers.

"As is to be expected, the emulated predictions based on the individual spatial fields are not visually distinguishable from the original predictions." Do you mean that they look “physical” with no artefacts? They are not meant to be similar to the original predictions. This is just like Fig. 3, I don’t really see the point of this figure.

Fig. 12: caption could be a bit more wordy to be self-explanatory if readers only partially read the article.

As we have decided to remove the section on the weather generator, we have also removed Fig. 12.

Typos: Line 17: are expected to remains in the Line 176: Fig. 7a -> Fig. 9a

Thank you, we have fixed the typos.

References

Deser, C., Terray, L. and Phillips, A.S., 2016. Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *Journal of Climate*, *29*(6), pp.2237-2258.

Fereday, D., Chadwick, R., Knight, J. and Scaife, A.A., 2018. Atmospheric dynamics is the largest source of uncertainty in future winter European rainfall. *Journal of Climate*, *31*(3), pp.963-977.

Lehner, F., Deser, C. and Terray, L., 2017. Toward a new estimate of “time of emergence” of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *Journal of Climate*, *30*(19), pp.7739-7756.

Lehner, F., Deser, C., Simpson, I.R. and Terray, L., 2018. Attributing the US Southwest's recent shift into drier conditions. *Geophysical Research Letters*, *45*(12), pp.6251-6261.

Shepherd, T.G., 2014. Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geoscience*, *7*(10), pp.703-708