RC3: Review of „Weather Type Reconstruction using Machine Learning Approaches“ by Pfister et al.

Summary

The authors assess various machine learning approaches for weather type reconstruction and then use the best-performing model to extend the time series of weather types over central Europe back to 1728. I consider the study a worthwhile addition to the effort of reconstructing past weather, but the authors should first address some questions that I raise below.

We'd like to thank the anonymous reviewer for his thorough assessment and his valuable comments to improve our manuscript.

Major comments

The authors try to make their results comparable with the study by Schwander et al. (2017), which has a marked negative impact on the readability of the text (e.g., in 81, 108, 150, 247, 268, etc.). Repeated efforts to explain differences in the dataset and methods, especially the constant switching between the CAP7 and CAP9 methods, lead to unnecessary confusion for the reader. Additionally, if I understand correctly, different classifications were used in validation and reconstruction, which could be considered questionable, to put it mildly. I suggest moving the discussion/comparison of the authors' results with the other study to one paragraph in Section 3, including Schwander's method among the methods they validate, and consistently using either the CAP9 or the CAP7 method. Schwander's model was already recalculated by the authors (and an error in the original calculation was found), which means that the change will be relatively easy.

Thank you for this suggestion. We agree that switching between the two closely related weather type (WT) classifications (CAP7 and CAP9) throughout the paper might lead to a certain amount of confusion for the reader, as was also pointed out in the reviewer's comment #1. We will therefore simplify the use of the two different WT classifications in the revised manuscript and restrict analyses using CAP7 to Sect. 3.1, while CAP9 is used in the rest of the manuscript.

As you remarked correctly, in Sect. 3.1, the machine learning approaches were trained on the CAP9 WTs, whereas the baseline approach (Schwander et al., 2017) was calculated for the CAP7 WT classification. Results of a validation with CAP9 (ML approaches only, not shown) revealed similar patterns to the validation with CAP7. In order to be able to compare all approaches, we show the validation results (Table 2) only for the CAP7 weather types, meaning that the 9 weather types predicted by the machine learning approaches are reduced to 7 WTs accordingly. We deem this comparison and the accompanied difference of the WT classification series used for model training and validation suitable for the following reasons:

- the CAP7 classification has been derived from the original CAP9 WT classification and the weather types of both classifications exactly correspond to each other with the exception that WT pairs 5 and 8, as well as 7 and 9 are merged. Implications with respect to training on CAP7 vs. CAP9 can thus be considered minor (i.e. a slight underestimation of accuracy).
- Our aim was to accurately reconstruct the CAP9 weather type classification using machine learning which is among other applications used in operational weather forecasting and climatological analyses (Weusthoff, 2011). The reduced CAP7 classification has been introduced by Schwander et al. (2017) as they found 9 WTs hard to distinguish with their approach. Using CAP7 would possibly omit relevant information on circulation types.
- As shown in Schwander et al. (2017), the baseline approach is not suitable to predict the 9 weather types of CAP9. Recalculating this approach for CAP9 would thus exceed its already

known limitations and thus not be meaningful for a comparison. Therefore, we used CAP7 for model comparison in Sect. 3.1.

In the revised manuscript, we will limit the comparison with CAP7 to Sect. 3.1 and move all such comparisons from Sect. 3.3 to the supplement, as we feel that this information might be interesting for readers working with the CAP7 reconstructions. The main part of the article will then consistently use the original CAP9 weather types.

In Conclusions, the authors should mention that best performing method differs depending on season, used data, and validation metric, instead of the over-generalizing "The feedforward neural network slightly outperformed the other ML approaches..." and that other method(s) lead to comparable results in a shorter time. I would also consider worth stating here that the ML approaches may be less sensitive to the quality of the classification (see my suggestion bellow, l. 306).

Thank you for this comment. We will specify this in more detail in the conclusions following your suggestion.

Minor comments

4 „in the accuracy of the used methods" is rather vague
Thank you. Indeed, the wording is not really precise about the nature of the restrictions of WT reconstructions. We will change this to „by methodical limitations".

5 CAP9 abbreviation is used but not defined
Thank you. We will introduce the abbreviation in the revised abstract.

11 WT abbreviation is defined here but not used consistently in the paper
Thank you for noticing. We will harmonize its use in the revised manuscript.

12 Can you refer to a global-scale evaluation of the skill of classifications that would confirm your claim that „In Europe classifications prove particularly useful to describe the prevailing atmospheric conditions?"
Thank you for this question. Our statement here seems to have missed its point. Our aim was to state that particularly in regions where weather is largely governed by large-scale circulation (such as Europe), WT classifications are particularly useful, rather than emphasize the role of Europe with respect to other regions in the world. We'll rephrase this in the revised manuscript (see also our response to reviewer's comment #1): „In regions such as Europe, where daily weather is largely governed by transient high and low pressure systems, such classifications […]"

19 Given your references here, probably "model outputs" would be more appropriate and clearer than "weather forecasting model simulations"
Thanks for this suggestion. As it is important to state the type of model, we will opt for „weather forecast model outputs".

26 "introduced, that" –> "introduced that"
Thank you for hinting at this. We will adjust in the revised manuscript.

32 "With the newest generation of reanalysis datasets, many WT records could already be extended back to the 19th century (Philipp et al., 2010; Jones et al., 2014)." How do these two rather old references support your claim regarding the newest reanalyses?

Thank you for this observation. Whereas the newest generation of reanalyses (i.e. particularly 20CRv3) would allow to extend WT reconstructions as far back as the early 19th century, the cited papers used older versions of reanalysis datasets. We will adjust the phrasing in the revised manuscript accordingly.

50 Maybe “machine learning” would be better than „artificial intelligence“?
We absolutely agree with this suggestion; thank you.

50-55 First, you write that “artificial intelligence is commonly used for classification...in climatological research” and then “In the context of WT classifications, ML is still a rather novel approach.” Please clarify. Furthermore, cluster analysis is commonly accepted as a machine learning approach and some clustering methods are among the oldest WT classification methods. Therefore, I would oppose your latter claim.
Thank you for these comments.
Regarding your first point: whereas artificial intelligence is used for classification and pattern recognition tasks related to the examples given in L. 50-55, we found only few applications of AI to WT classification specifically. We'll try to make this clearer in the revised manuscript.
Regarding the second point: the distinction between what are common statistical approaches and what is machine learning may be an issue of discussion. ML in this context was meant to refer to more advanced approaches like random forests or neural networks. Clustering approaches certainly are among the oldest and most prominent approaches for WT classification, but we would not necessarily count them as machine learning in this context.

59 It is not clear what “this pioneering work” refers to
Thank you. We refer to the three aforementioned references which used modern ML approaches for WT reconstruction. We will try to make this clearer in the revised manuscript.

94 Please add a little bit more detail to the decomposition; for example, what similarity matrix is decomposed?
Thank you for this excellent suggestion. We will include more details on the PCA step in the revised manuscript.

94 I’d suggest “the first step... the second step” rather than “a first...a second”
Thanks. We'll change the wording accordingly.

Fig. 1 Some stations in the figure (e.g., Cadiz) are not visible. I suggest moving the symbols in front of the coastal lines and increasing the size a little
Thank you for this suggestion. We'll adjust the figure accordingly in the revised manuscript.

102 ERA5 goes back to 1940, why was only 1957–2020 used?
Thank you for this question. We chose the shorter period as it corresponds to our reference period for which the CAP9 WT series is available.

112 Instead of “are well distributed across most parts of Europe” I would suggest a considerably more accurate “are relatively well distributed across central Europe”, or similar.
Thanks for this suggestion. We'll state the geographical distribution more precisely in the revised manuscript.

114 Probably “SLP” instead of “pressure” would be clearer and more accurate. This occurs several times throughout the paper
Thank you. We will change this to „sea level pressure“ in the revised manuscript.

126 It is not clear what “the latter study” refers to
Thanks. It should refer to Schwander et al. (2017). We will adjust the phrasing to make this clear.

148 Is not “The station sets used for the reconstruction of CAP9 WTs (Sect. 3.3.1) are summarised in Fig. 2” at odds with “Figure 2. Station sets of a) pressure and b) temperature used for the model comparison”, considering that CAP7 was used for comparison?
Thank you for this detailed observation. Indeed, Fig. 2 shows station sets used for both, comparison and reconstruction. We will adjust text and figure caption accordingly.

167 “Thus, temperature data were corrected for their seasonality by fitting the first two harmonics to each temperature record, which was then subtracted from the data.” should be made clearer; your sentence suggests that the temperature record was removed. I suggest "Thus, temperature data were corrected for seasonality by fitting the first two harmonics to each temperature record and then subtracting these harmonics from the data."
Thanks for this excellent suggestion. We'll adopt the suggested phrasing in the revised manuscript.

169 “their contribution” is not clear
Thank you for mentioning this point. We will rephrase it in the revised manuscript: „[…] trend or seasonality, which contribute only a negligible part to the total variability of these variables.

246 “they are trained” is not clear
Thank you. We will change this to „the former“.

246 “trained on the same station data” is surprising to me, since above you explained considerable differences between your dataset and that used by Schwander et al.
Thank you. In fact, we state in L. 120 ff. that the station data used for the model comparison is the same as in Schwander et al. (2017). The described differences refer to additional station series used for our CAP9 reconstructions presented in Sect. 3.2 and 3.3.

253 “both, hyperparameter” –> “both hyperparameter”
Thanks.

Table 2 Is it necessary to repeat “Acc =” and “HSS=” for all values? Did you consider another type of graphical output, which may be more readable?
Thank you for your suggestion. Given the large amount of combinations of station sets and models and the sometimes very small differences between the validation metrics, we deemed it more suitable to present the detailed quantitative results in a table instead of a graphical output. To make Tables 2 and 3 better readable in the revised manuscript, we will use different font styles for the two validation metrics instead of writing „Acc =“ and „HSS =“ and add a corresponding explanation in the table captions.

293 “This best–performing model” is not clear
Thank you. We'll change this to „The best-performing MLG model [...]“.

306-309 A useful interpretation could be that ML models may be less sensitive to the used classification compared with the simple baseline model. One may argue that projecting summer circulation on 7 WTs trained for an annual time series is far from ideal, because the classification lacks the necessary detail to explain the relatively weak and specific summer circulation. One reason for this is the dominant WT1: if a stand-alone classification was trained by cluster analysis only for summer days, it would find relevant summer patterns that would be more or less equally populated instead of the snowballed WT1.

Thank you for this remark. We agree with your interpretation that ML models are less sensitive to the used classification and especially to seasonal differences compared to our baseline model. However, CAP9 were originally determined from seasonality-corrected pressure data (see Weusthoff, 2011). The dominance of WT1 therein may thus not be interpreted as an inability to capture the specific characteristics of summer circulation due to the definition from annual time series. Regarding the training of our reconstruction models on annual time series, we examined a seasonally dependent standardization (see L. 170 ff), as well as training the models for each season (L. 459 ff), both of which did not yield satisfactory results.

319 “the overall atmospheric signal seen in a combination of information” is rather vague and unclear
Thank you for this comment. We will change this to „the information on an atmospheric state over a larger region“.

320 I would also consider adding “western Europe” because (south)westerly advection is an important feature of circulation over central Europe
Thank you for this suggestion. Whereas we do have the stations of London, Paris and Cadiz in western Europe, we agree that due to the important role of (south)westerly advection, more stations in this region may yield benefits. We will add this region to the list in the revised manuscript.

330 I think that “The model comparison revealed the feedforward neural network (NN) to exhibit the highest accuracy and HSS estimates” is an exaggeration and at least “on average” should be added into the main clause. However, to me it seems that the NN, RNN and CNN lead to nearly identical results and Table 2 shows that the best performing method is sensitive to the choice of season, data and validation metric.
Thank you for this point. We will rephrase this sentence in the revised manuscript.

377-9 “For a true WT 8 most false predictions show WT 5, and for WT 9 most false predictions show WT 7. Already Schwander et al. (2017) found these two pairs hard to distinguish, leading them to reduce the number of WTs accordingly.” I do not think that the results shown in Table 3 support Schwander’s claim. First, accuracy for WTs 8 and 9 are not worse than that for the other WTs. Second, these two WTs are outliers – if you imagine the 9 WTs ordered such that their position (for instance in a 1D or 2D Sammon map, the first PC plain, etc.) respects their position in the high-dimensional space, the two outlying WTs would simply neighbour with fewer WTs. Consequently, most (all) of the false hits will be linked to the one “closest” WT (which has a strongly correlated but weaker circulation pattern). I would not even consider this an artefact of the methodology but rather a geometric necessity.
Thank you for this remark. We absolutely agree that false predictions for the extreme WTs 8 and 9 will most likely be attributed to the „closest“ WTs, in this case WTs 5 and 7, independent of the methodology, and as you correctly state, our accuracies for these extreme WTs are not lower than for other WTs (whereas Schwander et al. (2017) seem to have had large difficulties to distinguish between the extreme and „closest“ WTs). We did not mean to support Schwander's claim with our interpretation of Fig. 3, but to emphasize the „false detection pattern“ for the extreme WTs and the general capability of ML models to correctly predict them. We will change this section to:
„For the „extreme“ WTs 8 and 9, most false predictions – as expected – identified WTs 5 and 7, which show the most similar patterns to the correct WTs 8 and 9, respectively (compare Fig. 1). Whereas Schwander et al. (2017) found these two WT pairs hard to distinguish and reduced the number of WTs accordingly, the NN model accuracies for WTs 8 and 9 are comparable to the other WTs. The NN model is thus capable of correctly distinguishing between these „extreme“ (i.e. with respect to the intensity and extent of high/low pressure systems) WTs and their less extreme counterparts.“

Figure 3: Consider specifying what are the “reference period” and “reference CAP9 series” in this case. In the validation phase, it is clear what accuracy means. However, this is not the case for the cross-validation periods
Thanks for this suggestion. We will include such an explanation in L. 375 in the revised manuscript.

Paragraph starting at 380+Figure 4 This could use some additional introduction and explanation – you lost me here
Thank you for this comment. We will add the following explanation in L. 384 in order to help the reader understand Fig. 4: „Deviations of the red and blue circles at individual/all observation points indicate regional/overall discrepancies in the observed pressure distribution as reason for false detection. Coinciding red and blue circles would mean that observation patterns of true and false predictions are identical and that the reason for false predictions are not explainable from the observations.“

Figure 5 Showing true and false maps as deviations from obs would probably support your interpretation more clearly
Thank you for this suggestion. We originally considered showing deviation maps with respect to observations. However, this blurs information on the position of low and high pressure systems in the false prediction maps. For the reader, it would be harder to see whether wrongly predicted patterns are just weaker (e.g. WT8 in Fig. 5a) or whether inverse pressure systems are apparent in certain regions (WT9).

408+409+488 Consider “transient situations” or similar instead of “transient WTs”. Transient WTs are WTs with low persistence, but you did not show that.
Thanks for this suggestion. We'll adopt it in our revised manuscript.

409 “The chosen WT for these cases might be arbitrary depending on slightly stronger patterns (i.e. dominating by a small margin)” Please reword, I do not follow
We'll rephrase this sentence in the revised manuscript: „The chosen WT for these cases typically is the one with the strongest imprint on the daily average station observations and not necessarily the one persisting throughout most of the day. Furthermore, a dominating WT might be chosen by a very small margin.“

411 “or by calculating WTs for a specific time of the day” I do not believe that this would have any effect on the presence of transient/boundary circulation fields in your data. Circulation fields form a continuum of patterns and I do not know of any reason to expect that instantaneous and averaged fields differ in this respect. Ditto 489
Thank you for this comment. Whereas daily averages of station observations may have blurred/mixed information from two or even three WTs in a transient situation, measurements for a specific time (e.g. 12 UTC) would provide a sharper pattern more likely to be attributable to one specific weather type. Hence, we would certainly expect some improvement with respect to using daily observations.

Figure 6b boxplots add y-axis labels (%?) and explanation of shown percentiles
Thanks for this suggestion. Figure 6b shares its y-axis with Fig. 6a; we'll adjust the figure to make this clearer in the revised manuscript and add an explanation of shown percentiles in the figure captions.

Figure 8 and especially 9: Consider using an even longer filter, the lines are still very noisy
Thank you for this suggestion. Our aim was to show the large year-to-year variability of WT occurrence. However, to make the figures better readable, we will remove CAP7 in the revised manuscript (see response to RC1).

450 “The results presented in Fig. 8 suggest that the reconstructed CAP9 time series do not show any apparent artificial discontinuities that go beyond natural variability.” How was this tested? You present this in Conclusions as one the main results but you did not provide any testing
Thank you for this important point. In accordance with the reviewer's comment #1 we will include results from statistical tests with respect to trends and discontinuities in the revised manuscript.

476 “Our results emphasize the importance of constantly improving WT classification methods…” One may argue that you did not test or enhance the CAP9 methodology itself, therefore rewording the sentence may be advised
Thank you for this suggestion. We'll change this to „Our results emphasise the importance of continuously improving methods of WT reconstruction...“

484 “occurrences” or similar may be better than “samples”
Thank you. We'll change „number of samples“ to „occurrence frequency“ in the revised manuscript.

486 “Adding seasons as additional predictors or training different models per season could solve this issue, although the sample size of rare WTs might be too small.” If season-specific classifications are trained (and a suitable classification method utilized that does not tend to identify marginal WTs), the issue of WTs with marginal occurrences will disappear. Additionally, season-specific classification could have fewer WTs, which would (I suppose) decrease the computation cost
Thank you for this comment. Whereas season-specific classifications (with an aptly trained classification model, of course) may considerably improve the issue of WTs with a strong seasonality or marginal occurrence, we wanted to point out some crucial issues related to that: smaller training datasets typically deteriorate the robustness of machine learning models. Especially a smaller number of samples of marginal WTs leads to an under- or overrepresentation of these WTs in the training dataset, exacerbating model training and possibly leading to an under- or overrepresentation of marginal WTs in the reconstructions. Furthermore, „hard-coding“ seasons into the reconstruction model like this may ignore seasonal shifts in WT occurrence. Regarding computational costs, training multiple models (i.e. for each season) with smaller input datasets does not hold much benefits with respect to training one model on a larger dataset.

The availability of monthly gridded datasets is mentioned in the paper. I was wondering whether including monthly SLP patterns as one the predictors could improve the models.
Thank you for this suggestion. Despite the idea being attractive for reasons of data availability, using monthly SLP averages directly as a predictor for the ML models is unlikely to improve results, as an artificial tendency towards an „average monthly WT“, as well as discontinuities in WT occurrence at the turn of each month would be introduced.

I would also welcome a note on the possible sensitivity of the models to the chosen classification and its parameters. Classification method is one of the major factors of any synoptic-climatological study and I would expect a significant sensitivity of WT reconstructions (and validation metrics) to changes in the classification methodology.
Thank you for this remark. Our article focuses on a single pre-defined weather type classification (CAP9) in order to compare the skills of different machine learning approaches to a baseline approach provided by Schwander et al. (2017), providing information on the sensitivity of reconstructions with respect to different models and input data (station sets) for this WT classification (Tables 2 and 3). While model performance is certainly sensitive to the chosen WT classification (e.g. dependent on the number of WT classes and their relation to the available station observations used as input, see e.g. the validation in Mittermeier et al., 2022), we did not perform a sensitivity analysis for the ML approaches with respect to other WT classifications (e.g. GWT or

Lamb weather types), as this – although certainly interesting – would go beyond the scope of this paper and thus has to be left for future research.