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.

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.

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).

Minor comments

4 „in the accuracy of the used methods" is rather vague

5 CAP9 abbreviation is used but not defined

11 WT abbreviation is defined here but not used consistently in the paper

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?"

19 Given your references here, probably "model outputs" would be more appropriate and clearer than "weather forecasting model simulations"

26 "introduced, that" –> "introduced that"

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?

50 Maybe “machine learning” would be better than „artificial intelligence“?

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.

59 It is not clear what “this pioneering work” refers to

94 Please add a little bit more detail to the decomposition; for example, what similarity matrix is decomposed?

94 I’d suggest “the first step… the second step” rather than “a first…a second”

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

102 ERA5 goes back to 1940, why was only 1957–2020 used?

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.

114 Probably “SLP” instead of “pressure” would be clearer and more accurate. This occurs several times throughout the paper

126 It is not clear what “the latter study” refers to

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?

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."

169 “their contribution” is not clear

246 “they are trained” is not clear

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.

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

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?

293 “This best–performing model” is not clear

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.

319 “the overall atmospheric signal seen in a combination of information” is rather vague and unclear

320 I would also consider adding “western Europe” because (south)westerly advection is an important feature of circulation over central Europe

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.

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.

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

Paragraph starting at 380+Figure 4 This could use some additional introduction and explanation – you lost me here

Figure 5 Showing true and false maps as deviations from obs would probably support your interpretation more clearly

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.

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

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

Figure 6b boxplots add y-axis labels (%?) and explanation of shown percentiles

Figure 8 and especially 9: Consider using an even longer filter, the lines are still very noisy

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

476 “Our results emphasise 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

484 “occurrences” or similar may be better than “samples”

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

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.

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.