Review of "Weather Type Reconstruction using Machine Learning Approaches" by Pfister et al.

General comments

The authors carry out a study for reconstructing Weather Types daily series back to the 1700s using a known WT classification with the use of several site measures and applying different Machine Learning Approaches to assess which one is more fit for the job.

The paper is well presented and written, however there are points that should be addressed:

1. The paper is based on an assumption that is not stated explicitly and cannot be taken for granted: Weather Types stemming from atmospheric variables can be assumed to remain the same across centuries. (See specific comment L.30-31). If we believe this hypothesis to hold, the authors made little effort to characterize temporal trends in the occurrence of the WTs and assess whether, from their reconstructed series, there have been shifts in occurrences from one season to another. From a climatic stand point I think these are relevant features of your very long (200+ years) classification.
   Thank you for bringing up this important point. We will state and discuss the stationarity assumption and its effects on the interpretation of our results in the revised manuscript. Furthermore, we will discuss seasonal shifts and trends of WT occurrence in more detail in Sect. 3.3. Further details are given in our response to the specific comment on L. 30-31.

2. The authors used CAP9 as classification but comment that two of the WTs can be considered similar/redundant and that is why CAP7 was preferred by a previous study which is often cited for comparison. It is unclear to me why CAP7 was not preferred over CAP9 provided that throughout the manuscript there are indications that having 9 WTs makes identifiability of WTs more complicated and prone to error.
   Thank you for this comment. The CAP9 WTs can be understood as the original classification used in operational weather forecasting and climatic analyses by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss, see Weusthoff, 2011), whereas CAP7 is a simplification made in accordance with the limitations of the methods used by Schwander et al. (2017). Therefore, and also to maintain comparability with future studies using the CAP9 classification by MeteoSwiss, we chose CAP9 as the target of our reconstructions, whereas the CAP7 series was needed to assess the performance in the method inter-comparison. In the revised manuscript we will use CAP7 only in Sect. 3.1 and use CAP9 throughout the rest of our analyses to avoid confusion for the reader (see also reviewer's comment #3).

3. Evaluation metrics in the summer season are systematically lower, making one doubt if the Weather Type classification is suffering from an under-representation of the atmospheric variable amplitude which is typically low in the summer season and high in the winter season (i.e., PCA input data has not been normalized by the seasonal cycle standard deviation). This aspect is important and should be clarified (see specific comment L.306-307).

   This is an important point. The calculation of the CAP9 WT classification accounts for

seasonal differences in pressure variation amplitude (see Weusthoff, 2011). Regarding our reconstructions we assessed several approaches to address this issue. We examined a seasonally dependent standardization of pressure observations (L. 171 f), as well as training individual models for each season (L. 459 ff). Both ideas had to be omitted as no consistent improvements in WT attribution could be determined, but on the contrary results at least partially deteriorated. We will try to emphasize this topic more in the revised manuscript. Further details are given in our response to the specific comment on L. 306-307.

Specific comments (section addressing individual scientific questions/issues)

L.12-13 – "In Europe, where daily weather is mainly governed by transient high and low pressure systems driven by the westerly jet stream". It seems a bit simplistic, especially because there are differences between the north-northwest part of the domain, influenced by the Atlantic, and the south-southeast part of the domain, where that influence is smaller and the Mediterranean acts as a frontier between the warm south and cold north.

Thank you for this comment. We agree that our statement omits many processes influencing European weather; we wanted to emphasize the important role of mentioned pressure systems. We will change this (in accordance with our response to the reviewer's comment #3) to „In regions such as Europe, where daily weather is largely governed by transient high and low pressure systems, such classifications […]".

L.30-31 – "in order to study long-term changes in atmospheric circulation patterns and associated surface effects, long-term series of WT are needed". I think this statement is debatable and not justified in the manuscript. There is an assumption behind it: Weather Type classification is an adequate way to analyze long-term changes in atmospheric circulation, and, more importantly, Weather Types are stationary, meaning that the same Weather Types are there in 1800 as well as in the 2000 – in other words, let's hypothesize that a reanalysis existed in 1700, applying a principal component analysis to e.g. geopotential height at 500 hPa to the period 1750-1800 and repeating the same to the period 1950-2000 would yield the same or similar EOFs and in turn describe the same patterns. These two are assumptions on which your paper is based upon, which deserve attention, cannot be taken for granted and should be clearly stated before carrying out your study.

Thank you for this suggestion. We agree that it is important to mention this stationarity assumption when analyzing our findings. The characteristics of typical synoptic situations may have changed over the course of the last centuries and deriving e.g. the CAP9 weather type classification with data from the early 1700s may lead to different weather types than for our reference period in the 20th century. However, for deriving typical WT classifications for periods back in the past by analyzing EOFs of past synoptic patterns and in consequence to detect such changes in the characteristics of governing WTs, the scarce station data available is insufficient. For this reason, our approach of reconstructing a (stationary) set of defined „modern" weather types for the past constitutes the only way possible to extend WT classification that far back in the past and allows to gain important information even despite the limitations that such a stationarity assumption may have. Whereas we cannot exclude slight changes of typical circulation patterns with this approach, the fact that average detection probabilities of reconstructed WTs are high (Fig. 6) throughout the last 300 years points to the validity of the stationarity assumption, as strong changes in the characteristics of governing WTs would lead to a decreased detection probability further back in the past. Also a lack

of trends in the reconstructed WT series can be interpreted as supporting the assumption in the way that changes in the governing WT characteristics over the last centuries are – if at all – small. Furthermore, by analyzing changes of a consistent, „stationary“ set of WTs in the past, changes in their occurrence frequency certainly hold important information on past variability or even long term shifts with respect to large-scale atmospheric circulation. In the revised manuscript, we will elaborate on this point in Sect. 2.1 and 3.3 and change the statement in L. 30f to: „By creating long–term time-series of WT classifications, important information may be gained to study long–term changes (i.e. over multiple decades or even centuries) in atmospheric circulation patterns and associated surface effects.“

L.40 – When you discuss the limitations of station-based reconstructions you could also mention that weather types generally describe atmospheric circulation over relatively large areas, so going beyond measures from a single point. Also: have these series been detrended?

Thank you for this suggestion. We will include it in the revised manuscript. As described in Sect. 2.2, temperature series have been detrended. Furthermore, all series have been bias-corrected with respect to the monthly EKF400v2 dataset, thus eliminating artificial trends or break points (also described in Sect. 2.2).

L.63 – You wrote that you use CAP7 for the study but then at the end of the section you write that you reconstruct WTs extending the CAP9. Please clarify.

Thanks for this comment. In accordance with our response to your initial comment No. 2, as well as to similar comments in review #3, we will simplify and clarify the use of the CAP7 and CAP9 WT classifications in the revised manuscript, including the line mentioned here.

L.65 – “It does not suffer from subjective WT classes”. WT classifications suffer from subjectivity because the choice of the number of classes is subjective unless there is a metric that helps choosing that number (e.g. BIC, Bayesian Inference Criterion, in Falkena et al. 2020).

Thank you for this comment. Whereas the choice of the number of classes in the CAP classification is subjective (see Ekstroem et al., 2002 for details), the term „subjective“ in this context refers to WT classification based on expert judgement, i.e. based on visual analyses of hand-drawn weather maps involving personal „subjective“ decisions. Those areare distinguished from „objective“ classifications (automated, based on statistical approaches using quantitative information) after Philipp et al., (2010). This terminology was introduced in L. 22 ff. We will try to clarify this in the revised manuscript.

L.76-78 – I don’t understand the purpose of these lines and perhaps it could be introduced if the authors (or other studies) had assessed the added value of wind direction on periods where this type of record is available.

Thanks for this comment. We refer to the ability of machine learning approaches to include also qualitative information (see L. 50), which could complement quantitative data and – especially in the earlier period – is more often recorded (but not digitized) than quantitative measurements. We will make this clearer in the revised manuscript.

To our knowledge, no other study has yet assessed the added value of wind direction. Neither have we been able to assess this for the reason given in L. 78.

L.81-82 – Why base the study on two classifications CAP7 and CAP9? And not one of the two?

Thank you for this question. We would like to refer to our response to your 2nd introductory remark, as well as your comment to L. 63.

L.103 – How are the WT classified into advective and convective? Please explain.

Thanks for this question. This classification is taken from Weusthoff (2011) which identified the dominating process (advective or convective) in a given WT class. We will indicate this in the revised manuscript.

L.104 – I understand that the WT are computed all year round implying that the larger amplitude of the variations of the atmospheric variables in the winter will potentially bias the WT towards winter patterns. Is there some sort of normalization of this amplitude throughout the year? Please clarify.

Thank you for this comment. Indeed, WTs are computed all year round. As tests with a seasonal normalization (see L. 170 ff), as well as with training individual models for each season (see L. 459 ff.) did not yield any clear benefits, we chose not to include a seasonal dependence, but to train the model on the full set of data. An idea behind this was that machine learning algorithms can eventually also detect seasonal differences in the distribution of input variables and thus correctly attribute WTs for individual seasons. Directly retracing this capability through the numerous connections within a neural network unfortunately is not possible. However, whereas small seasonal differences are apparent in Tables 2 and 3, a consistent bias towards winter patterns cannot be detected in Fig. 7. This supports the previous assumption and our methodical choice not to determine a different treatment for individual seasons.

L.107 – If some of the WT are hard to distinguish from one another as you write, why didn't you use CAP7 also for training your machine learning models?

Thank you for this question. As mentioned in our response to the reviewer's initial comment No. 2, CAP9 represents the „original" WT classification, whereas CAP7 is a simplification of the former made by Schwander et al. (2017), as their method struggled with distinguishing said WTs. We targeted reconstructions to extend the original CAP9 dataset back into the past to ensure comparability with future studies based on this more common (see Weusthoff, 2011) WT classification (and as first tests with ML approaches showed good results with respect to the similar WT pairs). CAP7 was thus merely used for the model intercomparison.

L.191-194 – "Increasing number of covariates can lead to overfitting of the model", I guess this characteristic is valid not only for this method. Also, could you clarify the choice of 4 as threshold for the VIF?

Thank you for this insightful comment. We agree that overfitting can occur in any statistical model with multiple predictors, particularly when linear combinations of predictors (e.g., in logistic regression) or functions (e.g., Generalized Additive Models) are used. In this sentence, we were specifically referring to overfitting caused by multicollinearity. If two predictors are strongly correlated, failing to account for multicollinearity can distort predictor estimates, undermine the statistical significance of features, inflate variances, and increase standard errors. This may render a parameter "useless," contributing to the curse of dimensionality without adding any benefits (e.g., better class separation). Overfitting may also be caused simply by adding too many predictors, even if they are not correlated, leading to good training performance but poor generalization in validation or testing datasets. While this type of overfitting can also happen in machine learning models, we addressed this by using two loops of 10/8 datasets to optimize model performance across all data

splits and did independent testing, effectively limiting overfitting.

Regarding multicollinearity, many machine learning algorithms are more robust. Algorithms that internally perform feature selection or use regularization techniques (as done in all the ML models tested in our study) are generally less vulnerable to multicollinearity due to their non-parametric nature.

To make this clearer we will change L. 191-194 accordingly.

On the choice of a VIF threshold of 4, we selected this as it is a commonly accepted conservative threshold in statistical modeling. In our testing, higher thresholds (e.g., 5 or 10) led to diminished performance in the validation datasets.

L.221-222- This characteristic is crucial: “As circulation patterns can persist several days”, as WT must persist a few days on average. The average persistence in days of each CAP7 (or CAP9) should be added to the manuscript along with a contingency table with weather patterns in rows and columns with shares (or counts) of transitions (e.g., see table 1 / 2 in Robertson et al. 2020). I wonder if transitions and preferential paths of transitions among WTs should be fed to the different machine learning approaches. Please comment.

Thank you; persistence and preferential transition paths of WTs are both important points, which we shall include in the supplement to the revised manuscript in the form of the suggested contingency tables. Regarding the suggestion of feeding transitions / preferential paths to the ML models: as both time-dependent models (RNN / CNN, taking into account information from three consecutive days) did not yield consistent benefits compared to feedforward NNs, adding time-dependent information such as probabilities or preferential paths of transitions to the model input (in addition to the fact that the WTs of the previous day would have to be known) is unlikely to improve model performance.

L.243-245 – As noted above, choice of CAP7 over CAP9. To make things comparable with Schwander et al. 2017 things are adjusted between the two classifications (cap7, cap9) in a way that it seems like a single one would have been more convenient.

Thanks for this suggestion. We will simplify and clarify the use of CAP7 and CAP9 in the revised manuscript in order to avoid confusion (see our response to the initial comment No. 2, as well as to suggestons in the reviewer's comment #3).

L.267-273 – In light of the error in the model set up found in Schwander et al. 2017 I wonder if it is worth following their footsteps so closely. I acknowledge the importance of having a reference study to compare to, but perhaps the authors could have been more brave in overcoming that study.

Thank you for this comment. Whereas the error in the model set up found in Schwander et al. (2017) represents a strong limitation of their approach, we feel that it serves as an excellent baseline also as the operational attribution of WTs (see e.g. Weusthoff, 2011) follows similar approaches. In the revised manuscript, we will, however, move comparisons with CAP7 in Sect. 3.3 to the supplement in order to follow a clearer structure and enhance our independence of that study.

L.296 – Table 2 shows that, of the four methods proposed, it seems like NN and RNN outperform the other methods, with RF always behind regardless of the number of stations.

Thank you for this observation. Yes, RF shows slightly lower accuracies (by 2 – 3 %) than the neural networks. We will mention this explicitly in the revised manuscript.

L.303-305 – I think this statement is not sufficiently supported. Are there works that have carried out similar analysis with statistical approaches that do not involve machine learning?

Thank you for this comment. To our knowledge, the studies by Schwander et al. (2017) and Delaygue et al. (2019) provide the only station-based European WT reconstructions and do not involve machine learning; both studies, however, do not go into detail about such methodological limitations. Nevertheless, a distance measure (even a statistical distance) is more rigid with respect to complex, non-linear patterns than machine learning approaches. We thus think that from a theoretical point of view, this statement can be considered valid. We will specify „capture details in the data and non–linear effects“ to „better fit non-linear relationships and interactions in the data“ and emphasize the hypothetical nature of the statement in the revised manuscript.

L.306-307 – Well, as noted above, this does not come as a surprise if no normalization of the atmospheric field was carried out prior to the computation of the regimes (the standard deviation during the year varies considerably with very low in the summer compared to the winter months e.g. figure 1d in Lee et al. 2023). This aspect is important and should be clarified.

Thank you for this comment. We agree that the seasonal differences in pressure variations should receive a more prominent position, which we will introduce in Sect. 2.2 in the revised manuscript. Whereas these differences have been accounted for in the CAP9 WT classification (Weusthoff, 2011), tests with seasonally dependent standardization of pressure observations (L. 171 f), as well as seasonal model training (L. 459 ff) unfortunately did not yield better results compared to using the full set of raw pressure observations. However, as seasonal variation of accuracy and HSSs in Tables 2 and 3 (especially the latter) are small compared to the baseline approach, ML methods appear to be able to tackle this issue at least to a certain extent, as argued in L. 306 ff.

L.309 – In light of the drops in accuracy in the summer this statement is perhaps optimistic and limited to the comparison to Schwander et al. 2017. “Our models are better capable of coping with seasonal differences”.

Thank you. We will change this sentence to „[…] models are better capable of coping with seasonal differences although some seasonal patterns in the accuracy remain.“ to put the statement in a clearer context.

L.350 – increased accuracy in fall and winter, otherwise for summer months – another clue in the direction of lacking summer information (normalization of atmospheric field)? Or the fact that wet days are more frequent in fall and winter as opposed to summer months (regardless of the type of precipitation – large scale in winter vs. convective in summer as noted at L. 355)?

Thank you for these considerations. The evaluation for the station sets of 1728 and 1864 did not show a uniform pattern with worse performance for summer months and better performance for winter months. We therefore cannot make such a conclusion regarding the use of wet days.

L.367 – I struggle with the term “accuracy” which, in e.g. operational forecasts, relies on the evaluation of simulated vs. actual variable values. In this case the actual values cannot be used as they can only be reconstructed. Therefore, is it appropriate to use this term? I suggest the authors clarify this point at the beginning of the paper, either in the Introduction or in the Data and Methods sections.

Thank you for this comment. The accuracy in this case refers to the station sets for the historical period with respect to a modern reference period 1957–2020. In accordance with a similar comment

made by reviewer #2, we will change the phrasing in L. 367 ff in order to be transparent about this issue: „The achieved accuracy using the smallest station set (stations available from 01.01.1728 to 31.12.1737) is already remarkably high […]. Adding more stations […]. Note that validation metrics shown in Table 3 only provide values with respect to the reference period 1957-2020. The actual values for the past periods may be lower due to larger uncertainties and errors in the data, but unfortunately cannot be determined due to the lack of a historical reference WT series."

L.370 – Summer months lower accuracies AND L. 399 – False WT predictions in summer seems to originate from other sources - related to year-round WT classification?

Thank you for this remark. The year-round WT classification (i.e. indifference to the seasonality of pressure variance) is certainly linked to the results shown in Table 3 (L. 370), as well as in Fig. 5b (L. 399), which is stated in L. 400. As for other passages in the manuscript, we will clarify this link in the revised version.

L.402 – "Weather types might change over the course of one day". Are you sure this characteristic is relevant in errors assigning WTs? Aren't WTs on average lasting 2+ days?

Thanks for this question. The daily CAP9 classification as described in Weusthoff (2011) tries to attribute instantaneous (i.e. daily) synoptic situations rather than patterns persisting over several days (see e.g. Mittermeier et al., 2022). Consequently, CAP9 does not include an undefined/transitional WT category. Whereas some WTs (e.g. WT 8) on average persist for two or more days, WTs occurring only for one day make up approximately one third of all situations and are prone to false predictions (as indicated in L. 405 f). As such changing conditions leave mixed imprints on daily averages of measured variables, the issue seems relevant for the reconstruction of daily WTs.

L.408 – Please clarify what you mean by "transient WTs".

Thanks for this remark. We'll change „transient WTs" to „transient situations" (see response to RC #3), for which a definition is given in L. 406.

L.415 – It is of great value that the NN attributes a probability of occurrence to all WTs, and I think this feature should be discussed further in the assessment of the good/bad WT daily classifications. I would expect that the WTs with highest probability isn't always with values of 0.8 or above and that days in which probabilities are more evenly distributed among the 9 classes exist. E.g. WT1 0.1, WT2 0.1, WT3 0.1, WT4 0.1, WT5 0.1, WT6 0.1, WT7 0.1, WT8 0.14, WT9 0.16, In this case what is the chosen WT, the WT9? One could argue that "no regime" class would be a more suitable choice. Have you counted how many times the probability of the winning WT is not crystal clear (probability much larger than the remaining WTs)?

Thank you for this comment. In Fig. 6b, we show the probability of the winning WT with respect to correct and false classifications. Whereas the model is confident for the correctly assigned WT (probabilities > 0.7), for false detections probabilities as low as 0.35 are apparent. Although not as extreme as the example given by the reviewer (in which case – as a side note – WT 9 would be chosen), the probability of winning categories especially for wrongly assigned WTs is not always crystal clear. As the CAP9 series does not have a transitional / neutral „no regime" class, we did not want to introduce such a class for our reconstructions, even though it might be a suitable choice. However, as probabilities are provided together with the respective WTs in the published series, users may choose to introduce such a class for uncertain cases.

L.431 /Figure 7 – Biases are visibly low for WT 8 and WT 9, do the authors have an explanation for this? From Figure 9 it seems that these two occur very little in the summer.

Thanks for this question. First, the small biases are directly linked to the low occurrence frequency, as in Fig. 7, the percentage of the biases is shown with respect to the number of days in a year. We will make this clearer in the revised manuscript. Second, WTs 8 and 9 show among the highest accuracies (see Fig. 3) for the individual WTs. Low bias values thus are to be expected.

L.450 – On the absence of artificial discontinuities: it makes no sense to comment on discontinuities using the eye over a plot with smoothed lines (10yrs running mean). Why don't the authors apply a statistical test for discontinuities/change-points on the non-smoothed series?

Thank you for this excellent suggestion. We will include a statistical test for discontinuities / change-points applied on the reconstructed WT series in order to support our statement in the revised manuscript.

L.457 – "artificial trends can be dectected" – have you found significant trends through the application of a statistical test? It would be interesting to know if/which WTs have become more or less frequent throughout the period of analysis and if WTs occurrences have shifted in season.

Thanks for mentioning this important point. We tested the yearly WT occurrence, as well as 10-year running averages for linear trends (linear regression with t-test) over the full 300 year reconstructions. Considering $\alpha = 0.05$, no significant trends could be detected for any WT. We did, however, find significant trends for individual seasons. An examination of our new reconstructions with respect to trends in WT occurrence is still ongoing and was thus not included in the presented manuscript. We will include a sentence in the revised manuscript, that statistical tests have been applied to detect trends.

L.470 – use either "thus" or "indeed".

Thanks. We will opt for „indeed" in the revised manuscript.

L.474-475 – I found no description of the detection method for trends and discontinuities in the manuscript.

Thank you for this remark. As mentioned in our response to the comments on L. 450 and L. 457, we will add results from our trend analysis and a yet-to-implement break-point detection method applied to the reconstructed WT series in the revised manuscript.

L.484 – "WTs with low occurrence and strong seasonality can pose a challenge for reconstructing WTs", this is why I wonder why CAP9 was preferred over CAP7 (fewer WTs).

Thank you for this comment. We would like to refer to our response to your initial comment No. 2 which treats this issue.

L.488-490 – "Transient WTs make the distinction on a daily resolution difficult,… issue might be solved with the use of subdaily data". I consider this option inadequate for the very nature of reconstructing WTs back to 1700s, it is already a miracle if you get a daily value, imagine subdaily, utter wishful thinking! Also, as far as transient WTs are there and may hinder daily classification, the degree to which the knowledge of sub-daily WTs would help such classification is far from demonstrated. WTs are, by design, approximation of reality at at daily time scale, it is to be expected that in some days a good match with the archetypal WT is lacking, it's part of the game.

Thank you for this remark. First, we would like to emphasize that subdaily station records reaching back to the 1800s and even 1700s are by far not as much wishful thinking as the reviewer suggests. Efforts to gather historical meteorological data such as the International Surface Pressure Databank (ISPD; Compo et al., 2019) have made available multiple sub-daily pressure time series back to the early 19th century. Also, some long and homogenized sub-daily series reaching back even to the 17th century have been recently introduced (Cornes et al., 2023). Nevertheless, more digitization and homogenization efforts focusing on sub-daily data would be needed to provide a robust basis for WT reconstruction. We totally agree with the reviewer that synoptic situations sometimes do not match well with archetypal WTs; the issue of transient WTs could nevertheless be improved by using sub-daily information.

References:

Falkena SK, de Wiljes J, Weisheimer A, Shepherd TG. Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector. Q J R Meteorol Soc. 2020; 146: 2801–2814. https://doi.org/10.1002/qj.3818

Robertson, A. W., N. Vigaud, J. Yuan, and M. K. Tippett, 2020: Toward Identifying Subseasonal Forecasts of Opportunity Using North American Weather Regimes. Mon. Wea. Rev., 148, 1861–1875, https://doi.org/10.1175/MWR-D-19-0285.1.

Lee, S. H., M. K. Tippett, and L. M. Polvani, 2023: A New Year-Round Weather Regime Classification for North America. J. Climate, 36, 7091–7108, https://doi.org/10.1175/JCLI-D-23-0214.1.