LegacyClimate 1.0: A dataset of pollen-based climate reconstructions from 2594 Northern Hemisphere sites covering the late Quaternary

Response to comments of Anonymous Referee #1

1. General comments

Reviewer comment: (1) *The authors provide temperature and precipitation reconstructions based on pollen assemblage time series. They provide three different types of reconstructions and provide a clear description of the methods. The dataset is highly valuable and the manuscript is clearly written and the figures are of high quality (if sometimes a bit small).*

Response: Thank you for this encouraging comment.

Reviewer comment: (2) *The manuscript seems to be part of a set of articles (a trilogy?): a manuscript describing the raw pollen data, a manuscript dedicated exclusively to the chronology and the present manuscript about the pollen-derived climate reconstructions. I can to some degree follow the rationale of the sequence, but I think this (last?) article would benefit from a closer integration with the article describing the chronology. The chronology, and importantly its uncertainty, is an integral part of the climate reconstruction that the authors present here.*

Response: Thank you for this comment. We revised the method part and not more clearly indicate the rationale of the three manuscripts.

2. Major issues

2.1 Integration with chronology

Reviewer comment: (1) *This manuscript focuses entirely on the reconstruction of temperature and precipitation, yet the time series also have a chronology with associated uncertainty. By separating these two aspects into two manuscripts it becomes unclear how the full uncertainty of the paleoclimate time series can be derived. Looking at the data (on pangaea.de) it seems that the provided error only accounts for the reconstruction, not for the chronology. This is not the full story and the manuscript would be tremendously improved if the authors made this third manuscript of the sequence a true integration of the papers on the chronology and the climate reconstruction. In L341-343 the authors even touch on this possibility, but they refrain from taking the logical next step that would make the data product more useful for other researchers.*

Response: We thank the reviewer for this suggestion. While it is possible to obtain the ensemble of age-models from the Bacon modelling in the chronology paper from Li et al. (2022, LegacyAge 1.0, https://doi.org/10.5194/essd-14-1331-2022), it would clearly simplify the task for users and foster better practices related to age-uncertainty if we also included it in the current manuscript along with the reconstructions data.

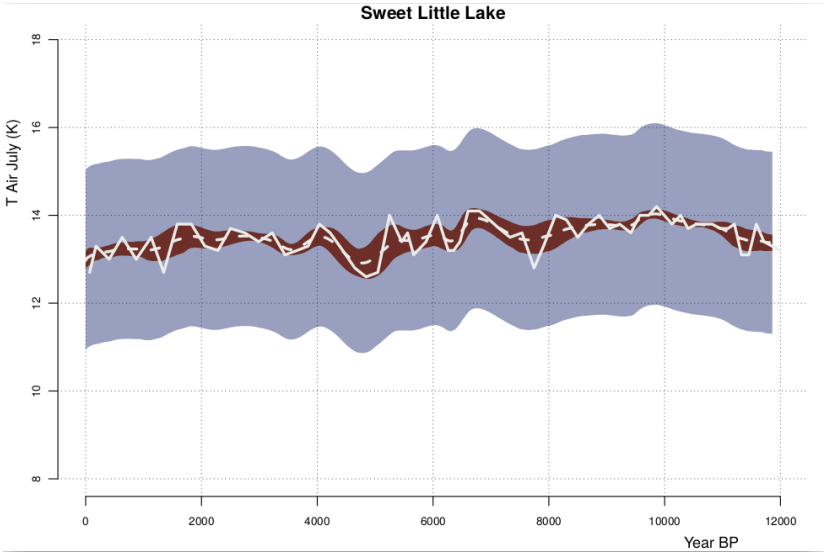
Therefore, we added to the database of reconstructions an ensemble of 1000 realizations of the age-models for each record based on the Bacon age modelling performed by Li et al. (2022, LegacyAge 1.0).

Reviewer comment: (2) *This means that the first order analysis of the time series as shown in figures 5 and 6 should include some combined error resulting from the reconstruction and the chronology and a clear description of the methodology to combine these errors. The provided data sets should also contain uncertainties that reflect both the chronological and the reconstructions errors. This is not a complicated step, but would massively improve the value of the data product.*

Response: Using these 1000-member ensembles, we recalculated the maps for figure 5 taking each time the average value over the ensemble of the interpolated values at 1 ka BP and 6 ka BP. The spread (standard deviation) of the ensemble is then used as a measure of the error related to age-uncertainty, or the chronological error. The standard reconstruction error (based on the RMSE from the transfer function) is likewise interpolated at 1ka BP and 6ka BP and the mean over the ensemble is taken as the reconstruction error for the given age. As the chronological and reconstruction errors are independent, they can be added in quadrature to obtain the combined error. This information was added to the source data for Figure 5 and the methodology described.

In the case of Figure 6 (now Figure 10), the figure is provided to show the overall spatio-temporal availability and variability of the timeseries and adding errors on the figure would make it illegible. The data product does include however all the necessary information to perform an analysis, namely the reconstruction error for every sample, and the ensemble of 1000 realizations of the age-model for each record. With this information, users can easily produce curves with all relevant uncertainties such as for example the following figure (title “Sweet Little Lake" on which is shown the reconstruction error (shaded blue) and the chronological error (shaded red) around the reconstruction smoothed by the time-uncertainty (i.e. when we interpolate at regular timesteps for the 1000 realizations and average over the ensemble, dashed white). As the reconstruction and chronological errors are independent, we could also show the combined error by adding them in quadrature, although it will be almost indistinguishable from the reconstruction error given it is much bigger. The original reconstruction with the median ages is also shown for comparison (solid

white); this underlines that averaging over the age models only preserves the low-frequencies but (unrealistically) smooths out the high-frequencies. We see that showing all this information on Figure 6 for all the reconstructions would not be possible.



2.2 Meaning of reconstruction differences

Reviewer comment: (1) *The authors also mention other reconstruction methods (L372), which begs the question why MAT and WA-PLS were chosen. Only because they are widely used, or because they yield superior results?*

Response: Thank you for this comment. By providing the fully harmonized modern and fossil datasets as well as the climate data it is rather simple to adjust the R code to run customized reconstruction using further reconstruction methods. Given that many reconstruction methods were proposed in the last 20 years, we decided to provide here only the two most generally used methods. We scanned the literature before running the reconstructions to select the most commonly used methods and selected MAT and WA-PLS. To provide reconstructions from more methods would be beyond the focus of our manuscript. However, we added further discussion on the potential to use the framework for further reconstruction methods.

Reviewer comment: (2) *In addition, the authors provide three different reconstructions for each time series. What I miss is a discussion of how these different reconstructions can be used. Does the difference between them represent additional uncertainty on the reconstruction? How should the user include or use this information? Are certain reconstruction methods better than others? If so, which is to be preferred? If not, how can the (information from the) reconstructions be combined?*

Response: We revised the discussion in '**5.3 Reconstruction method and LegacyClimate 1.0 quality'** and addressed the questions raised in detail.

2.3 Reconstruction quality

Reviewer comment: *The CCA suggests that only some part of the variance in the training sets is explained by T and precip and the significance testing indicates that a shocking 60-70 % of the reconstructions are basically noise. Whilst the authors go some way and filter out the time series that do not pass the significance test, I feel that the authors hardly mention this, let alone discuss. I also realize that this manuscript should not analyze the data, but perhaps some discussion in place and the different ways in which (pollen) assemblages could be used in paleoclimate science, including forward modeling, could be highlighted.*

Response: We did not filter the dataset as we here only provide a dataset that could be used for climate analyses. However, we provide quality measures for each fossil pollen site including measures for the quality of the modern training set (e.g. CCA), for the transfer function (e.g RMSEP) and for the reconstruction (significance test) etc.. In particular, the significance test should rather be taken as additional information than as an exclusion criterion. The significance test tests whether the variation in the pollen data can be significantly explained by the reconstructed climate variable. If the reconstruction does not pass a significance test it indicates that either 1) the climate did not or only marginally change and hence variation in the pollen signal is small and the reconstructed climate variable does not explain a significant amount; or 2) the climate change signal in the pollen data was too small compared to non-climate related changes (e.g. taphonomic changes) or, 3) the changes in the pollen signal are not depicted by reconstructed variables e.g. because the modern data set is not appropriate. Only cases 2) and 3) indicate a failure of the reconstruction method.

We now highlight at the end of the discussion that further assessments and a more comprehensive uncertainty analyses would improve the quality of the dataset.

New text in the discussion part: Our assessments of the modern dataset (e.g. CCA), the transfer function (e.g. RMSEP) and the reconstruction (e.g. the significance test) revealed also the potential biases in the pollen-based reconstruction and pointed to limitations. Further validation and assessments of the results and a more comprehensive uncertainty analyses e.g. by applying forward modelling approaches (Izumi & Bartlein, 2016; Parnell et al., 2016) would be highly valuable.

2.4 Land use issues/human influence

Reviewer comment: *Some of the time series must bear an imprint of human influence. Can the authors briefly discuss to what degree and if and how this influences the reconstructions?*

Response: We added plots of typical land use indicators (as far as available from the harmonized pollen data LegacyPollen 1.0, Herzschuh et al., 2022).

New text in the methods part: We used *Plantaginaceae* (mostly representing *Plantago lanceolata*-type in Europe) and *Rumex*-type to assess human influence as an indicator for intense herding (Behre, 1988). In addition, we calculated the correlation between the WA-PLS reconstruction of T_{July}, T_{ann} and P_{ann} and the pollen percentages of *Plantaginaceae* and *Rumex* for 9000, 3000 and 1000 years BP.

New text in the result part: We used the abundance of *Plantaginaceae* and *Rumex* as indicators of grazing and such intense animal husbandry. Overall weak human impact is inferred for North America and Northern Asia. The indicators *indicate strong human impact* only in single records at 9000 years BP in China and the Mediterranean region (Fig. 7). The percentage values of *Plantaginaceae* and *Rumex* were high especially in Europe for 3000 year and 1000 years BP which indicates growing human impact on that region. High *Plantaginaceae* correlate with low T_{July} in Central Europe indicating potential biases in the temperature reconstructions i.e. too low temperatures become reconstructed (Fig. 8).

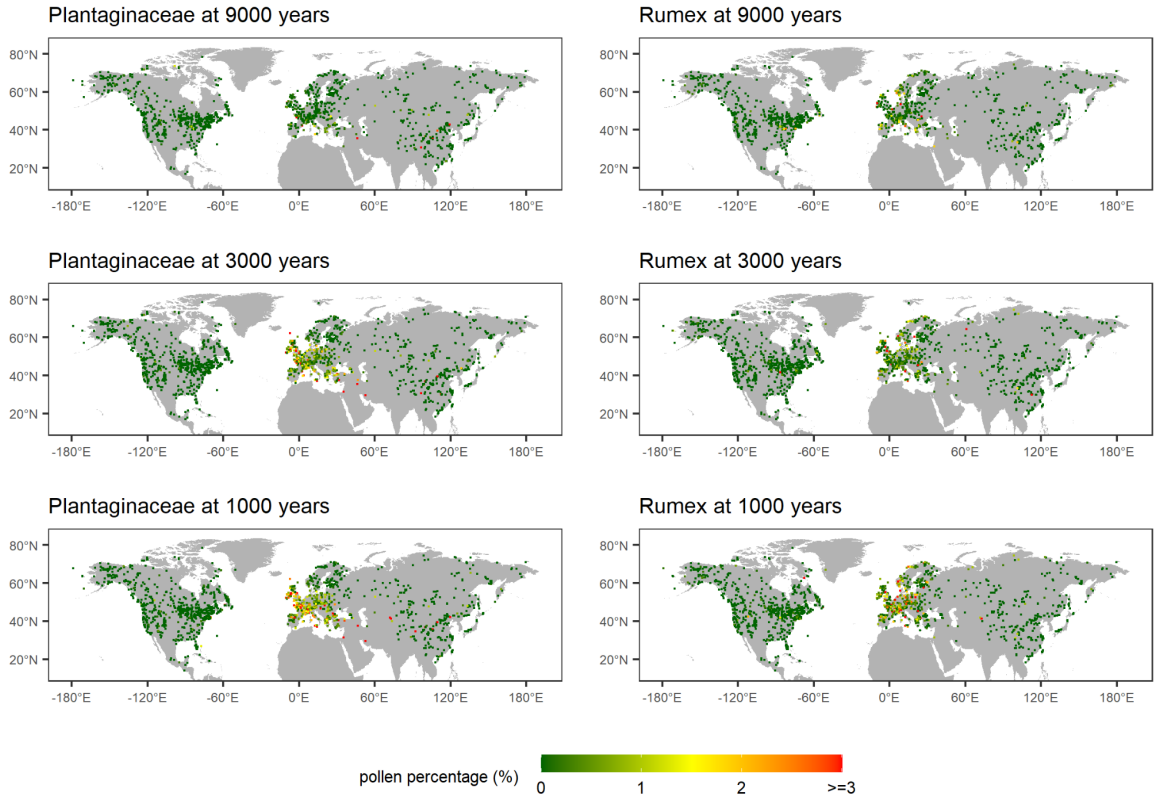


Figure 7. Abundance of *Plantaginaceae* (left) and *Rumex* (right) at 9000, 3000 and 1000 years BP. Colors indicate percentage values.

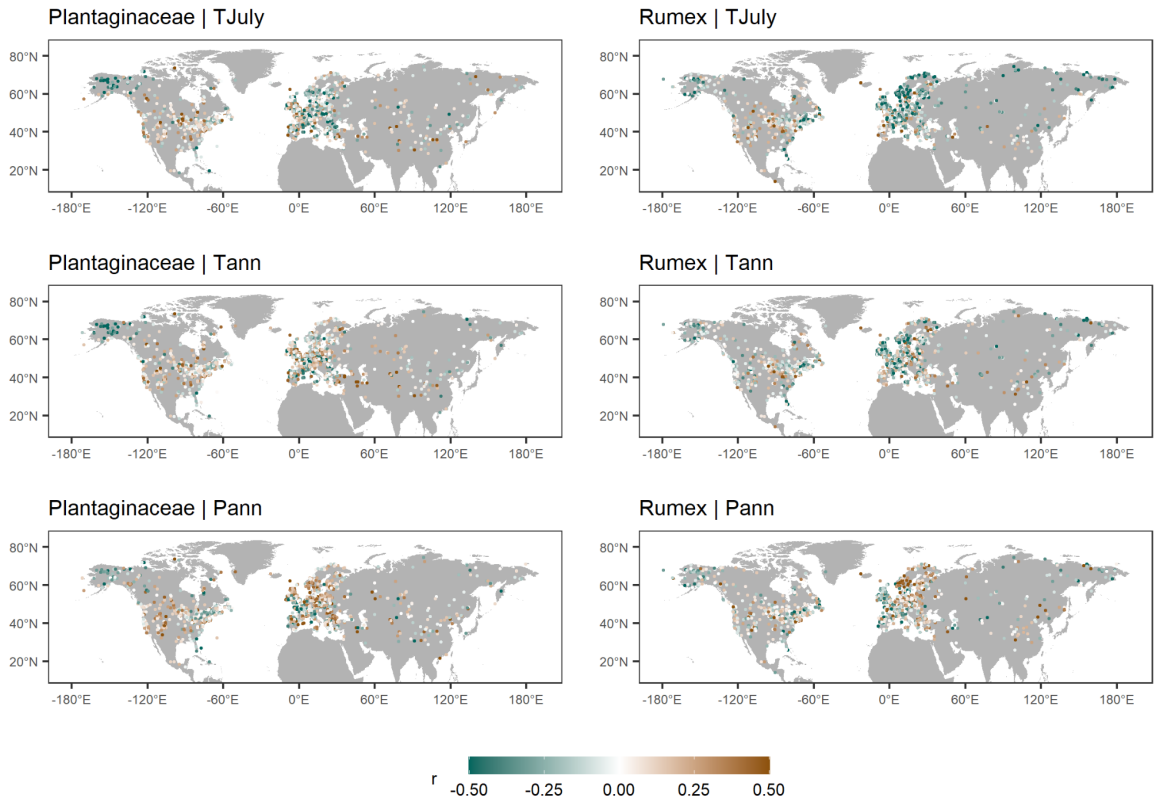


Figure 8. Correlation between the percentage of *Plantaginaceae* (left) and *Rumex* (right) and reconstructed T_{July}, T_{ann} and P_{ann} with WA-PLS.

2.5 Insufficient explanation and detail in the methods

Reviewer comment: (1) *2,000 km radius for training set. Please explain why this was done and why the distance is (globally) appropriate.*

Response: As part of this study we did not perform specific investigations to assess the optimal size of the modern training-set; which would go beyond the focus of this study. In a study from Eastern Asia 1000-1500 km was considered optimal (Cao et al., 2017). However, due to the low number of modern samples in some areas (e.g Northern Asia) we fixed the radius to 2000 km as a good compromise.

New text in the methods part: We fixed the radius to 2000 km, instead of 1500 km as suggested from a study in Eastern Asia by Cao et al. (2017), because the modern dataset density is rather low in Northern Asia.

Reviewer comment: (2) *Why were seven analogues used for MAT? Are the reconstructions weighted to analogue quality, or simply the arithmetic mean of the seven closest analogues?*

Response: We made some tests in advance of the analyses. We found that the results are not very sensitive to the number of analogues (i.e. we tested whether more or less records would pass the significance test). However, we refrained from a systematic study which would be computationally very expensive and go beyond the focus of this study. Accordingly, we decided to stay with the default parameters of the rioja R package used which is 7 analogues.

Reviewer comment: (3) *How is the calibration error determined? Was spatial autocorrelation taken into account? From the code it seems that this is not the case, why?*

Response: The calibration error was determined using the default leave-one-out cross-validation of the rioja package. We report the RMSEP from cross-validation for the models and RMSE for all samples. The *rioja* R-package is one of the most commonly used packages for climate reconstruction using proxy data.

Reviewer comment: (4) *What is the sample-specific error based on? Why is this provided and not the calibration error?*

Response: We provide the full model RMSE as well as the RMSEP derived from leave-one-out cross-validation.

Reviewer comment: (5) *If I am correct, the tailoring approach serves the purpose of reducing the effect of co-variation between T and P. Please mention this earlier in the methods. I understand the point and that this goes some way to alleviating the problem. But what is done in cases where the correlation is not reduced? After all, there still is a large proportion of the sites for which there is a marked correlation in the training set. Some discussion would be appropriate here.*

Response: We now provide more explanation on the rationale. We assume that information about temperature and precipitation cannot be separated from each other if all samples are almost located along a linear line in a temperature vs precipitation space i.e. if they are highly correlated.

New text in the methods part: In addition to the classic WA-PLS reconstruction, we also propose WA-PLS_tailored. This approach addresses the problem that co-variation of climate variables today in space is transferred to the reconstruction even if the past temporal relationship among the climate variables mechanistically differs. In fact, this approach aims to make use of the full climate space covered by the modern pollen samples avoiding those samples in the calibration set that cause the spatial covariation. This approach is based on the assumption that several climate variables can be reflected in one and the same pollen assemblage because different plant taxa have different optima

in temperature and precipitation ranges and might therefore occur with different co-occurrence and abundance pattern.

Reviewer comment: (6) *Please provide more detail on the significance test. How were the random environmental fields generated? Simple permutation, or taking spatial correlation into account. Why?*
Response: The significance test is described in Telford and Birks (2011, https://doi.org/10.1016/j.quascirev.2011.03.002). We extended the text by a little more information.
New text in the methods part: A statistical significance test (Telford and Birks, 2011) was applied using the *randomTF* function in the *palaeoSig* R-package (version 2.0-3, Telford, 2019). In this test, the proportion of variance in the fossil pollen data explained by the reconstructed environmental variable is estimated from redundancy analysis (RDA) and tested against a null distribution generated from a total of 999 randomly generated environmental variables from the training data. A reconstruction is considered statistically significant if the reconstructed variable explains more of the variance than 95% of the random reconstructions (Telford and Birks, 2011). The reconstructed climate parameters were tested as introducing the environmental variable as a single variable in a run, as well as with partialling out the explained variance in the pollen data by the respective other variable.

Reviewer comment: (7) *Why were the tailoring and the significance testing not applied to the MAT reconstructions?*
Response: Significance testing is currently also applied to the MAT and summary will be reported in Table 2. From some tests we can see that the percentages of sites that pass the significance test are in the similar order of magnitude as for the WA_PLS. However, running this test is extremely computational time-consuming; accordingly we can provide the results only with the next review round. Tailoring would not make sense with MAT as here the same analogues for temperature and precipitation are used.

Reviewer comment: (8) *The CCA seems to be the first step in the development of the transfer function model to demonstrate that T and Precip really explain the variance in the assemblages. Would it not be better placed earlier in the description? And why are the implications barely discussed?*
Response: We agree and present the CCA now in the beginning of the results part and added some discussion text.

New text in the discussion part: We a priori selected T_{July}, T_{ann} and P_{ann} as target variables for our reconstructions. However, we provide λ_1/λ_2 (i.e. explained variance of the climate variable in the modern pollen data set relative to the variance explained by the unconstrained first axis; ter Braak, 1988), a commonly used proxy for the assessment of reconstructions. The higher λ_1/λ_2 in the spatial modern dataset desto higher the chance that this target climate variable has also impacted vegetation over time and is thus reflected in the variation of the fossil pollen dataset. As a rule of thumb a ratio of 1 is considered to indicate reliable reconstructions (Juggins, 2012) though useful reconstruction may also be obtained from datasets with lower values. .As expected, maps of RMSEPs reveal similar spatial pattern as the results of constrained ordination. Our results indicate that in particular calibration sets from Europe have low ratios and a high RMSEP for all climate variables (despite we have a high number of modern samples), likely related to the human impact on the modern and fossil data. Some areas that are known for its sensitivity to precipitation e.g. Eastern Asia show low RMSEPs as expected for P_{ann} but on the other hand show a low sensitivity to T_{ann} and T_{July}.

Reviewer comment: (9) *How are poor analogues treated? Do they occur at all after the lumping? There is some discussion in L327-332, but it is unclear what the user of the data can do with this information.*

Response: We now calculated the analogue quality of the all samples and the thresholds (1%, 2.5%, 5% of modern calibration set) for single calibration sets. Results are presented in the manuscript now. However, we did not exclude reconstruction analogues without analogues because almost all samples had more than 7 analogues <5%.

New text in the methods part: To infer the analogue quality as an indicator of no-analogue situations we calculated the minimum dissimilarity (squared chord distance) between modern pollen assemblages and fossil pollen assemblages with probability thresholds of 1%, 2.5% and 5% using the *minDC* function from the *analogue* package (version 0.17-6, Simpson et al., 2021).

New text in the discussion part: We report the analogue distance for each sample to help identify such situations. From our assessments we revealed that analogues quality is overall rather good at least for the Holocene and except for Western Europe in particularly the British Isles (Fig. 4).

3. Minor issues

Reviewer comment: (1) *L3: reconsider the use of "late quaternary" in the title. The meaning is actually rather vague and something along the lines of 30,000 years would be more informative.*

Response: Done.

Reviewer comment: (2) *L108: not sure what the policy is to refer to submitted manuscripts.*

Response: Meanwhile, the manuscript about LegacyPollen 1.0 and LegacyAge 1.0 became accepted.

Reviewer comment: (3) *L131: please provide a bit more detail on WorldClim 2. For instance, what are the data based on, over what period are the data integrated, etc.*

Response: We added more detailed information about the WorldClim 2 dataset compilation to the text.

New text in the methods part: The site specific T_{ann}, T_{July}, P_{ann} were derived from WorldClim 2 version 2.1 (spatial resolution of 30 seconds (~1 km^2), https://www.worldclim.org, Fick and Hijmans, 2017) by extracting the climate data at the location of the modern sample sites using the *raster* package in R (version 3.5-11, Hijmans et al., 2021; R Core Team, 2020). The WorldClim 2 dataset provides spatially interpolated gridded climate data aggregated from weather stations as temporal averages between 1970-2000 (Fick and Hijmans, 2017). We used monthly average temperature data to extract the mean T_{July} and the "bioclimatic variables" bio1 (T_{ann}) and bio12 (P_{ann}).

Reviewer comment: (4) *L385: crucially, this manuscript does not describe a fossil pollen data set, but a data set of temperature and precip*

Response: We clarified this in the text.

Reviewer comment: (5) *L402-404: this seems a somewhat dangerous statement. Are the two reconstructions really independent?*

Response: We now refer to our tailoring approach where we target on the independent reconstruction to temperature and precipitation.

Reviewer comment: (6) *Why is the x axis of figure 6 on a log scale?*

Response: We now provide reconstruction on normal time-scale for the last 30 ka.

Reviewer comment: (7) *Whilst glancing through the code I missed the significance testing and the CCA. (But thumbs up for sharing the code.)*

Response: We used standard packages in climate reconstruction and reconstruction assessment. We decided to provide the code for the reconstruction in particular to show how the tailoring-approach is implemented which is methodologically new.