Review of "Reconstruction of winter temperature of southwest China over the past 300 years based on a Bayesian approach"

The manuscript by Chen and Brönnimann develops a Bayesian framework to reconstruct winter temperature in Southwestern China since 1700 CE. The methodology combines multi-archive documentary evidence with prior distributions computed from climate simulation ensembles. The methodology and results are innovative, valuable, and fit very well into the scope of Climate of the Past. As detailed in my comments below, there are some methodological aspects that I ask the authors to improve. After resolving these issues, the manuscript could be suitable for publication in Climate of the Past.

General comments

- **Presentation of observations:**
  The manuscripts provides a limited overview over the number and spatio-temporal distribution of observations included in the reconstructions. Could you make a visualization of the spatio-temporal characteristics of the observations (e.g., time series with the number of observations per year, a map with all observation locations and how many observations they contain) and the number of observations available for each observation type?

- **Construction of the likelihood:**
  The advantage of the Bayesian model to naturally incorporate uncertainties in the likelihood function into the posterior is only useful if the construction of the likelihood is reproducible and can be critically assessed by peers. It is currently unclear to what extent this construction follows objective criteria and to what extent it is based on a subjective assessment of the documentary evidence. Can you clarify how Table 1, which contains a deterministic association of CWI and observed phenomena, and Table 2 are related? Table 2 seems to be the most important to understand the likelihood construction. To make an assessment by peers possible, a similar table with information for each year should be included in the supplement or data/code repository.
  It is stated that uncertainties in the documentary data such as source reliability, record quantity, and descriptive accuracy, have been incorporated in the likelihood (l. 215-216). Can you document how this is done? Potentially, the above mentioned table can clarify this if extended to all years. In general, consider moving Sect. 4.2.1 to the methods section as it clarifies several of the methodological questions I had while reading the methods section. In addition, it should appear prior to presenting the CWI reconstruction in Sect. 4.1.
  Finally, some of the observed phenomena relate to short-term (extreme) events. To what extent are winter mean conditions correlated with these events? How strongly are anomalous conditions in one subregion associated with the regional mean winter temperature?

- **Construction of the prior**
  How are the winter temperatures in the simulations split into CWI classes? Are 20 members sufficient to create a stable prior or should there be some smoothing applied to ensure that the prior has a sufficient spread (similar to a variance inflation factor in ensemble Kalman filtering)? Regarding the time-independent prior, I'm unsure about its purpose. Should it create a reconstruction with minimal influence from the temporal structure of the simulations (in this case, it would be the most convincing to take the same prior for each year, which could simply be the class probabilities across all simulated winters), or should it assess sampling uncertainties from the small ensemble size (in this case, I'd recommend to repeat the construction of CWI-ModE-Clim multiple times to quantify how results change for different realizations).

- **Computation of the posterior CWI and temperature**
  In my understanding of the Bayesian model, you first compute the posterior CWI distribution by multiplying the prior probabilities of each class with the likelihood of each class according to $p(\text{evidence}|\text{CWI} = n)$, and normalizing the resulting product. In a second step, the posterior mean of $p(T_{\text{rec}}|\text{evidence})$ is computed by computing the weighted mean of the mean temperatures of each class, where the weights are given by the probabilities of the class. If that's correct, I would rewrite Eq. 4 such that it only contains $p(\text{CWI}|\text{evidence})$ and only introduce the temperatures in the subsequent paragraph. Furthermore, the posterior distribution of T_{rec} should therefore be given by a mixture distribution

$$T_{\text{rec}}|\text{evidence} \sim \sum_{n=1}^{6} p(T_{\text{rec}}|\text{CWI} = n) \cdot p(\text{CWI} = n|\text{evidence}), \tag{1}$$

  where $p(T_{\text{rec}}|\text{CWI} = n)$ is the temperature distribution corresponding to the n-th CWI class. The posterior mean of that distribution is equal to Eq. 5 in the manuscript, but the credible intervals of that mixture distribution cannot be computed by simply computing the weighted mean of the quantiles of the temperatures of each class. In fact, this strategy will very likely underestimate the posterior uncertainty. Instead the posterior quantiles need to be computed from the posterior cumulative distribution function (CDF) $F(T_{\text{rec}}|\text{evidence})$ given by

$$F(T_{\text{rec}}|\text{evidence}) = \sum_{n=1}^{6} F(T_{\text{rec}}|\text{CWI} = n) \cdot p(\text{CWI} = n|\text{evidence}), \tag{2}$$

  where $F(T_{\text{rec}}|CWI = n)$ is the CDF of the n-th temperature class. Alternatively, it can be obtained through Monte Carlo sampling from the posterior distribution of T_{rec} (Eq. 1). In both cases, $p(T_{\text{rec}}|\text{CWI} = n)$ needs to be specified, e.g., by using its empirical distribution or a parametric approximation of it.

Specific comments:

- Title: Given the wide range of available paleoclimate data types, the title would become more meaningful if the use of documentary evidence was mentioned.
- Abstract: Using a Bayesian approach says fairly little about the underlying statistical model because almost any statistical model can be formulated in Bayesian way. Therefore, a little more information about the statistical model would be useful in the abstract, in particular how documentary evidence and climate simulations are combined in the framework.
- l. 22: To what time period does the "data-sparse" refer? In general, it would be helpful to specify the period of interest in the first paragraph.
- l. 33: Can you provide an order of magnitude for how much more data is available in eastern China compared to southwestern China?
- l. 44: To emphasize the causal structure of evidence, consider replacing "differ from reality" by "reverse the causal structure of reality".
- l. 49-50: There are plenty of applications of Bayesian frameworks in paleoclimatology. Is there a reason to highlight the study by Camenisch et al. 2022?
- l. 67/68: Can you also state the starting date of the compilations?
- l. 91: Replace "surface temperature" by "near-surface air temperature".
- l. 92: Please provide the exact definition of winter that you use in the study.

- l. 93: I wouldn't call the difference between the ensembles "bias" but rather deviation, since bias would imply that there is a known truth from which they differ.
- l. 106: What does "time-independent" mean?
- l. 118: Why is "frost" grouped into precipitation and not perception of temperature? How is frost observed?
- How are northern / middle / southern subtropical climates distinguished?
- Fig. 1: Can you comment on the reasons for the relatively large differences between DFDP and GSOD for rainy days?
- Fig. 2a: It would be more intuitive for me to use a moving window instead.
- Fig. 2b,c: Consider visualizing this as a barplot containing the observed frequencies for each phenomenon.
- Fig. 3c,d: I would suggest to not connect neighboring years by lines since the Bayesian approach models each year independently.
- Fig. 4: Please add visualizations of uncertainties in Fig. 4 (for example using a density heatmap in Fig. 4a showing the posterior probabilities of each class for each year).
- l. 320-321: Could you add uncertainties here to the numbers?
- l. 321-323: Is this a higher co-occurrence rate than expected by chance? I'm wondering because there are $\sim$50 extremely cold / cold winters (i.e., $\sim$20%) in the reconstruction. If a similar number occurs in southeastern China than a co-occurence rate of 4% or $\sim$10 common cold winters among 250 years would be expected even if the two regions were statistically independent.
- Fig. 5a: I'm surprised that the distribution of CESM-LME for 1700-1949 seems to be warmer than for the period 1950-2000. Can you comment on this?
- Table 3: Do the p-values account for autocorrelation in the data, which reduces the effective degrees of freedom?
- l. 414: Do you have any insights into which mechanisms are responsible for the warm winters? Similar to my question above, is the agreement of warm winters in CWI-ModE-Clim and ModE-Sim actually higher than expected by chance?
- Fig. 7: Can you also plot a comparison of CWI-ModE-Clim and CWI-ModE-Sim which would show the influence of the simulations on the CWI-ModE-Sim reconstructions?
- l. 419: That information would be helpful much earlier, either in the data or methods section.
- l. 439, 443-444, 446, 453: Please add references here.
- l. 459: Do you mean lower amplitude or lower temporal frequency?

Technical comments:

- l. 32-33: Please check the grammar of the sentence.
- l. 38: Please check the grammar of the sentence.
- l. 385-386: Please check the grammar of the sentence.
- l. 439: Please check the grammar of the sentence.