The author develops a complete daily gridded dataset of precipitation and temperature for Mexico at a very high resolution considering the extension of the spatial domain. The research is mostly correct; however, I have some concerns about the method used to develop the daily grid and the validation process.

I thank R2 for these positive comments and for the time taken to review this manuscript.

My answers to R2's suggestions/observations are written in blue.

The choice of an interpolation method is not easy, especially for precipitation, since it can yield very different results depending on the method and the parameters. The KED is a trustworthy option when used for a single timestep, for example to show the daily precipitation/temperature in one day/event (as shown in the examples of extreme events). However, when you apply the same method for a long-term climate time series without further corrections, you are introducing temporal biases that can lead to unwanted inhomogeneities both in temporal and spatial dimensions. This is not new and it is a basic assumption in gridded datasets creation as shown in the wide (not cited) scientific literature (e.g. https://doi.org/10.1002/wat2.1555, https://doi.org/10.1002/joc.1322,https://doi.org/10.1029/2008JD010100, https://doi.org/10.1559/152304085783914686). I recommend the author to make a deeper review on the requirements for creating reliable grids, starting by some of the datasets that are cited in the article. The actual problem with creating a single grid for each day, independently from the previous and following, is that the number and location of neighboring observations change with time, and that lead to biases that have a significant impact in, for example the analysis of trends or even in the aggregation at coarser temporal scales. In this regard, my second main concern is the validation approach.

Thanks for the recommendation; I will improve this section using the provided references.

It is ok to check the errors by series with the proposed statistical tests, however, you are comparing the complete series of observations with their corresponding predictions without considering the number of missing values, the differences by elevation ranges, by months, seasons, etc. It is difficult to see biases that are useful to interpret how reliable are the results. In addition, the kriging process usually provides a variance dataset, measuring the uncertainty associated with the prediction at every specific location. It would be useful to see an associated gridded dataset with the error/uncertainty for each day and variable, as done in many other datasets, to account for the reliability of each prediction and let the user decide how to use the information.

I plan to use data from automated meteorological stations operated by Mexico's Meteorological Service which are located at different elevations. However, there are not even

100 of these stations throughout Mexico, and their temporal coverage is limited (some stations were deployed in 2010, others in 2012 and so on).

Here are the minor and specific comments, line by line:

Introduction:

L30: “along the migratory route of the Monarch butterfly” Maybe this is too specific.

My idea is to show that climate data is used in studies that are not related to water resources and the Monarch butterfly is probably the most well known species of butterfly in North America.

L44: Terraclimate is regularly updated and now it is available until 2024.

Thanks for pointing this out; I have modified it.

L66: For CONUS, I think that PRISM deserves to be mentioned since it was one of the first and still one of the more reliable gridded datasets (https://prism.oregonstate.edu)

Thanks for your suggestion; I will include it.

Methods:

L101-102: how many stations was the final number?

The current version of the manuscript reads (L101-103) “… and once the climate records were in PostgreSQL only those stations with more than 10 years of registered data were selected; accordingly, not all available stations were used, and the number of stations varied across the 1951‑2021 period, as shown in Fig. 2.”

Also, the caption of Fig. 2 reads “Number of weather stations used for daily interpolations of (a) temperature, (b) precipitation”.

L108: regarding the outliers, wasn't any additional quality control performed? There's a lot of scientific literature on this.

For Mexico, the L15 dataset used those stations with > 50 days of data.

L108-109: how many stations were discarded?

The total number of stations with data is 5467; I will add this to the manuscript.

L110: As mentioned before about the use of KED independently for each day, it can generate further problems in long-term trends and temporal aggregations. Also, why 30 nearest stations and a 140km radius? The number of observations can greatly vary under these conditions. Lastly, was the internal coherence of temperature (TMAX>TMIN) checked after the interpolation, for each day? The above problems are especially important if a proper quality control was not performed to control the spatial coherence of the data (and I read nothing in this regard).

Regarding the selection of 30 stations and 180 km to apply KED_l, I will add the following lines:

"These values were recommended by Carrera-Hernandez et al., (2024) after a detailed comparison of different Kriging variants at both national and stratified domains showed that KED_l using elevation as a secondary variable provided the best representation of yearly precipitation in Mexico." The manuscript that details these analyses is currently under review (Hydrological Sciences Journal).

The internal coherence of Tmax>Tmin was verified for each station before undertaking the interpolation. I will include the Tmax>Tmin coherence of the interpolated datasets on the revised manuscript.

L120-121: Is the code publicly available?

The code is not publicly available, but I can add an example to the new version of the manuscript.

Validation:

I expected a more complete validation since here only daily data considering the whole series was checked. For example, how the interpolation worked at different elevation ranges? or in different months? Did the method correctly predict the number of dry/wet days? Are monthly (or other) averages and standard deviation fit between predicted and observed values? These are the basic checks for any gridded dataset.

On the revised manuscript I plan to use independent data from automated meteorological stations located at different elevations. However, there are not even 100 of these stations throughout Mexico, and their temporal coverage is limited (some stations were deployed in 2010, others in 2012 and so on).

Figure 5: I am not sure how to interpret these graphics since, for example, R^2 needs complete series of predictions and observations to be compared but here you have one value per day/month/year

Indeed, Figure 5 shows the daily values of the coefficient of determination (R^2), the Coefficient of Efficiency (COE) and the Index of Agreement (IOA). These values were obtained through daily leave-one-out cross-validation. I will ad a flow diagram to show how the leave-one-out cross-validation was performed, along with how the semivariogram was modelled daily to perform the daily interpolations.

L242: why not comparing monthly or annual aggregates? or even trends? that would be more useful than comparing extreme events, which are not common (by definition) and the users may need a more regular use of the dataset.

I decided to use these extreme events because these events are important to study areas where flash floods occur and because floods (in general) are natural disasters in Mexico that need to be analyzed, In addition, these extreme events also cause landslides in Mexico´s mountainous regions.

L274-277: This comparison is not fair since you're comparing predictions with observations but only in the case of your dataset you know that the observation does not participate in the interpolation, but not in the rest of datasets. Furthermore, not all of them were built with the same observations, so it is hard to justify better results on your dataset.

I do not agree with this point; DAymet and L15 used some of the observations that I used to develop the MexHiResClimDB. In fact, I explain this issue on lines 277-281:

These performance statistics are shown in Fig. 7 along with their respective scattergrams of differences; however, it should be kept in mind that the performance metrics shown in the aforementioned figure for the MexHiResClimDB can not be directly compared with the metrics of Daymet, L15 or CHIRPS, because for the latter three cases some of the weather stations used to compute the metrics were used to develop the datasets - in summary, the performance metrics obtained through cross-validation are expected to be lower.

L298: what this function does?

I will describe what this function does in the revised manuscript. The r.univar command computes univariate statistics from the non-null cells of a raster map; these statistics are number of cells, minimum and maximum cell values, range, arithmetic mean, variance, standard deviation, coefficient of variation and sum.

L300: what was the threshold for considering a dry day? 0 mm / 0.1 mm / 0.001 mm?

Accoding to L300: "With this procedure, the ten wettest and driest days, months and years were obtained and summarized in Table 2 … ". The values of Table 2 are in 1×10^9 m^3 per day, month and year. I wanted to keep three decimals for all the columns and that is why the minimum daily values of precipitation shown in Table 2 do not change much. These values

are ordered in ascending order, and they were obtained by creating a table using the values obtained with the `r.univar` command.

L309: I dont see tendency in that table

I agree and in fact that is what can be read on line 309: ".. that no wettness or dryness tendency is easily seen on the values shown in Table 2"

L326-327: again, this is not a validation, just a comparison with other datasets that were not constructed with the same procedure. The only validation must be with the observations.

These lines refer to the performance metrics obtained with leave-one-out cross-validation that are shown on Fig. 8.

Lines 325-327 state the following: "To validate the interpolated temperature maps, the maximum and minimum values of Tmax (1998-6-15, 1967-1-10) and Tmin (2020-8-31, 1962-1-12) were selected to report their validation in detail. The results obtained with the leave-one-out crossvalidation are shown on Fig. 8, …"

L332-335: a visual comparison does not guarantee a correct validation.

I agree and will rephrase these lines accordingly.

L347: Fig 10 shows absolute values, but this is not trends. If you want to show trends you should calculate some statistics (Mann Kendall, Sen's slope) with their corresponding reliability value (p-value).

L350-351: this is not an acceptable way to indicate that there is a trend. Without statistical validation, this complete section must be removed.

I have modified Figure 10 and added a five-year moving average of the monthly and yearly anomalies in order to improve what is written on lines 347-351. The modified figure is shown as a separate file.