We thank the reviewers for their thoughtful comments, which have enabled us to improve the manuscript. We address all comments below. Our responses are in **bold.**

Reviews:

Evaluation of global gridded climate products in reproducing spatial and temporal variation in precipitation in central Panama

First reviewer: Anonymous 1

An analysis comparing the accuracy of different downscaled rainfall products over central Panama is presented. They conclude with recommendations to use CHIRPS and CHELSA2.1, although for different purposes. The analysis conducted seems competent, and the manuscript is easy to follow. Given the vast amount of research conducted in central Panama, I think this manuscript will provide value to the ecology community.

**General comments on manuscript:**

* My largest critique is the suggestion to use CHELSA2.1 - this conclusion does not seem well founded.
* **We agree and have modified our conclusions and associated discussions to recommend CHIRPS.** **We now write** **“For studies of among-site variation within central Panama, we recommend CHIRPS v2, which best captured spatial variation in total annual precipitation and dry season precipitation. For analyses requiring time series, we recommend CHIRPSv2 which best captured seasonal and interannual variation and is available in monthly intervals from 1981 to near real time. Note that CHIRPSv2 has only 0.05 degree (~5 km) spatial resolution.”**
* The title seems inappropriate to me. The title describes an analysis of global gridded climate products. There are many global gridded climate products (e.g. ERA5, JRA55, MERRA2, etc.). The study seems to focus on downscaled precipitation products, which are not at all the same thing as global climate products. Some of these are also not 'global' (CHIRPS).
  + **We modified the title to “Evaluation of downscaled gridded climate products for spatial and temporal variation in precipitation in central Panama.”**
* It might be useful to include some or one of the standard global precipitation products as a point of reference for how much these downscaled precipitation products (presumably) improve upon their accuracy. For example, GPCP v3.1 (https://disc.gsfc.nasa.gov/datasets/GPCPMON\_3.1/summary) is a well-regarded reference dataset in climatology.
  + **We appreciate the suggestion, but the coarse resolution of these products means that many study sites fall within the same pixel, so we would not be able to evaluate spatial performance in a comparable manner. Further, these products are interpolated from ground stations, and most likely are based on many of the exact same stations that we are using, which makes such an analysis somewhat circular. (The relatively low density of stations in Panama and the tropics more broadly also means that these products are expected to perform much less well in the tropics than in Europe and North America, where station density is much higher.)**
* I think the authors should consider whether the downscaling adds value. Consider how the products are downscaled. To what extent is the downscaling excessive? There is nothing to stop producers from 'downscaling' to infinitesimally fine spatial resolution - but at what point is the downscaling just fantasy? Just because it is possible to downscale precipitation, does not mean the result offers greater information than whatever it was downscaled from.
  + **We agree that downscaling is not always useful and cannot be done excessively. At the same time, we note that statistical downscaling typically relies on inputs such as topography, wind fields and clouds, which are very influential in shaping and predicting spatial variation in the precipitation. Central Panama has a very steep rainfall gradient that is underrepresented by coarser products. Our results show that the downscaled products closely approach the precipitation capture by rain gauges.**
* "TERRA" should be correctly called "TerraClimate" as it was originally published (Abatzoglou et al., 2018).
  + **Thank you, we’ve corrected this.**
* TerraClimate is also available through the end of 2022.
  + **Individual years are indeed available through 2022, and we use individual years for the interannual variation comparison 1979-2016. However, for the climatologies we use the TerraClimate 1981-2010 climatology.**
* To my understanding, "R" is short for the coefficient of determination, whereas 'r' is Pearson correlation coefficient.
* **Thank you, we’ve made the correction.**
* The recommendation to use CHELSA 2.1 for anything seems ill founded. It has effectively the same Pearson correlation as CHIRPS for total annual precip but is far worse in the other metrics. CHIRPS seems to offer the best tradeoff both spatially, seasonally, and interannually. I would strongly urge the authors to reconsider how ecologists are most likely to use a precipitation data product. It is unlikely they would use one to represent spatial variation, and another to represent temporal variation. I suggest the concluding recommendation to be streamlined - CHIRPS does well across all comparisons, whereas the others are a mixed bag with typically far worse results.
  + **We agree with the reviewer and have revised the conclusions. Our previous recommendation of CHELSA for capturing spatial variation was influenced by its higher spatial resolution and was partly a holdover from before we added CHIRPS v2 (which we did only at the very end of the analysis). On further consideration, stimulated by the reviewer comments, we agree that CHIRPS is the best product to recommend for precipitation, as noted above.**
* The authors might consider the consequences of the forcing dataset. Reanalysis products tend to do more poorly than satellite products over the tropics (Alexander et al., 2020).
  + **This is a good point, and we now address this in the discussion. We write “**Some of our in-situ stations are included in the observational datasets that inform WorldClim, TERRA, and CHPclim v1.0, and thus we expect these products to exhibit higher performance relative to these in situ datasets, even though they are likely to perform more poorly at independent validation stations.”
* I think the authors should look to the climate research community's preferred products. My impression is that GPM IMERG is currently regarded as the best rainfall products (Tang et al., 2020). I was sort of surprised not to see it included. The resolution is only slightly coarser than the products compared here, but the authors might consider including it.
* Alexander, L. V., Bador, M., Roca, R., Contractor, S., Donat, M. G., & Nguyen, P. L. (2020). Intercomparison of annual precipitation indices and extremes over global land areas from in situ, space-based and reanalysis products. Environmental Research Letters, 15(5), 055002.
* Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., & Hong, Y. (2020). Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets. Remote sensing of environment, 240, 111697.
  + **We considered including one or more of the GPM IMERG products (V06, V07, and/or TRMM), and reconsidered including all three of these in response to this reviewer comment (previously we had evaluated only TRMM). We now include all three of these products in the supplemental table of products considered for inclusion. One issue is that all three have a coarser resolution than the 5 km maximum we set for our analysis: TRMM has a spatial resolution of ~30 km per pixel, and V06 and V07 have a resolution of ~10 km.**
  + **Nonetheless, we have now looked more closely at these data in relation to the rain gauge data. Our analyses show that these datasets have high errors; rainfall data quality indexes are generally too low after 2008. Many of these data have quality flags for our region (low Quality Index values), which appear to be related to the relative dearth of ground data for calibrating the satellite data. Given these issues, including only the higher-quality portion of the dataset (2000-2008) would yield insufficient coverage for the full-time extent of our study, and the product's resolution would still fall outside the scope of our analysis. For these reasons, we did not include any GPM IMERG products in our analysis.**

**Figure and table comments:**

* Figure 1: ● What are the units? ● Suggest adding a North arrow ● Suggest increasing the font size of the stations
  + **The figure has been modified by adding a north arrow, including hillshade instead of elevation and increasing the font size.**
* Figure 2: The default color palette in the raster package is not perceptually uniform. I suggest using a perceptually uniform palette that is colorblind accessible. The authors might consider viridis or scico: https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html https://github.com/thomasp85/scico It would be better to put the units on the color bar, and not in the title. ● A background color that is different from the color palette would be good. I suggest a dark gray.
  + **Thank you for the specific suggestion; we have revised the figure accordingly.**
  + Figure 3: ● Bias is a continuous variable. I suggest using a continuous color bar instead of discretized bias classes in a legend. ● The usage of green and red is often not differentiable to colorblind people. I suggest using a colorblind safe, diverging palette: https://colorbrewer2.org/#type=diverging&scheme=BrBG&n=5 https://github.com/thomasp85/scico **We made use of scico palette and cite it in the reference of the online book at https://vasquezvicente.com/ClimatePanama/references.html**
* Figure 4: Suggest capitalizing Jan-April
  + **Done.**
* Figure 6: Suggest allowing the scale of the y-axis to vary by site.
  + **We prefer to maintain the same absolute scale across sites.**
* Table 2: 1981 marked the beginning of the satellite era for many products. It may be worth reconsidering the analysis range to be from 1981-2016 instead of 1979.
  + In the case of previously published climatologies, we cannot adjust the date range ourselves - we are constrained to what is provided. We thus left the date ranges unchanged.
* Table 2: I suggest grouping the climatology products together, maybe even in a different table. The comparison of the climatology products is a bit challenged by the different time periods encompassed by the products.
  + **We agree that ideally all the products would be available for all the years of interest and compared for those same years, but we are constrained by what is available. All the included climatologies are within the period of 1979 to 2016. The differences in time extent were addressed by subsetting the ground data to only include values collected in the same time period for each of the datasets.**

**Line comments:**

* 75: The reader is unlikely to know who Steve Paton is.
  + **We changed this to include a citation.**
* 84: How were they resampled? Nearest neighbor, cubic convolution, bilinear …?
  + **We used nearest neighbor resampling and have now clarified this in our methods text.**
* 87: Is Table S2 supposed to be Table 1? There is no Table S1 referenced.
  + **Thank you for pointing out this error; we have corrected the reference.**
* 90: Kind of odd to compare dry season, but not wet season. It would be useful to justify this choice further.
  + **We chose to evaluate dry season precipitation in addition to annual precipitation because dry season precipitation is especially important for the biota and ecosystem function, thus of special interest to ecologists. We also ran analyses of wet season precipitation initially, but those results paralleled those for annual precipitation, and we expect annual precipitation results to be of more general interest. Hence our choice to include both dry season and annual precipitation results. We note that we also addressed differences between seasons in our seasonality plots.**
* 97-102: The gap filling approach would be clearer if the equations of the model were included. Also, it would be important to know how much of the data was gap-filled.
  + **We have now included the equation and a citation to the lmr4 package that we used for the gap-filling. About 10% of the ground data is gap-filled as shown in the full code provided in the associated code repository, which is currently publicly available in GitHub and https://vasquezvicente.com/ClimatePanama/ground\_data\_wrangling.html (**<https://doi.org/10.5281/ZENODO.13992333>**r). Intermediate tables can be reproduced with the code provided.**
* 118: Readers might not know what the "lab clearing" is.
  + **Good point, addressed.**
* 190: I am unclear about how the authors concluded to recommend CHELSA 2.1 The correlation is effectively equivalent to CHIRPS, whereas CHIRPS has notably less error. Thus I really disagree about the recommendation of CHELSA 2.1. Consider the likelihood that a user is just going to use a rainfall product to characterize spatial variation within a defined region - they will not. They will likely use the same product again for different purposes (e.g. temporal change, etc.)
  + **We agree and are grateful to the reviewer for these comments.**

Reviewer 2

General comments:

The manuscript shows results from a direct comparison of precipitation data from ground-stations in Panama (with a special focus on BCI), comparing long monitoring records to model/reanalysis datasets. This is a timely and important set of results, especially considering that the evaluation uses records from tropical ecosystems, often poorly represented in gridded data products. Precipitation is a particularly hard quantity to model and create gridded products, further increasing the value of making these comparisons.

**Thanks!**

The results from the comparisons are very direct, without exploring several aspects that would be of interest and possible with the datasets used here. For instance, the exact grid cell where the location might not always be the most representative cell when comparing to a ground-station, especially for cells close to large water bodies. Such an analysis is surely out of scope for the current manuscript, but a potential follow up study could consider a buffer of cells around the central cell where the ground-station is located (e.g., a 9x9, or 16x16 cells evaluation, identifying the closest match to modeled results).Another potential improvement (future analysis?) would be to evaluate the effects of gaps in the ground-station data, e.g., by artificially introducing gaps to the measured data and evaluating the impact on the comparisons.

**These are good points, and we also hope to see future such analyses, building on the current work.**

The authors mention methodological biases on the sensor data, it seems important to understand the uncertainty stemming from the instruments and data processing methods. Also, what was the rationale for focusing only on annual and dry season precipitation? Although dry season patterns are more impactful ecologically, differences across seasons might show important similarities/differences to models. Figure 5 shows monthly level data, showing considerable differences closer to the end of the year for the estimates (which is the wet season). Adding these to the paper would be valuable, and if not, an explanation as to why not would be beneficial.

**A good question, which we addressed in response to reviewer 1.**

On Line 67: the following statement could be rephrased or expanded for precision, since the study area for the paper goes beyond BCI, even if centered on it. “The focal region is centered on Barro Colorado Island and the Panama Canal (Figure 1).”

**Good point, we have revised the text to read “**The focal region of this study is an area in Central Panama defined by coordinates between -80.2 and -79.4 degrees longitude and 8.8 to 9.5 degrees latitude. This region includes the Panama Canal watershed and centers around Barro Colorado Island, as shown in Figure 1. **”.**

The methods mention 79 stations (6 from STRI), but Figure 1 seems to show fewer (only 2 for STRI). This apparent mismatch should be clarified in the methods or figure caption, e.g., by more precisely describing the stations used (Lines 77-79). A table listing the sites and their coordinates would be ideal to represent this information, especially considering reproducibility purposes.

**We have revised the methods text to clarify that we include only 30 stations in the analysis, as these are the ones that remain after filtering by data availability and quality. The supplemental material include complete listings of the stations considered and their coordinates. All the data and code are published, our revised manuscript includes the relevant citations to the data and code packages.**

The text should contain references for descriptions for the data and code availability, as well as resources for long-term preservation of the ground-station datasets used in this work. In the methods section, it might be interesting to mention the “terra” by name and, which functions within the package were used.

**We agree data and code availability are important. All data and code our now published, and we have modified the manuscript to include the appropriate data and code citations. We cite the terra package in the main text, and the code shows exactly which functions are used.**

(Line 84) The recommendations for use-cases based on strengths and weaknesses of each dataset are very helpful and a good contribution from this work.

**Thank you.**

Reviewer 3- Steve Paton

* You might also mention that different rain gauges report different values - we have two electronic gauges on BCI that differ by 4%
  + **We consider this recommendation valuable. We performed variance analysis between the BCI records. The supplemental material and available code posted on** [**https://vasquezvicente.com/ClimatePanama/ground\_data\_wrangling.html**](https://vasquezvicente.com/ClimatePanama/ground_data_wrangling.html) **describes the differences between station BCIELECT(STRI), BCI(ACP) and BCICLEAR(STRI) which vary by about 4%. We present our rationale for including the 3 BCI records.**
* This would probably be phrased as "can underestimate precipitation by…" Pollock 2018 only tested three gauges. There are some very sophisticated sensors that may have less of a problem. Furthermore, wind and evaporation are big contributors and there would certainly be places where they are not nearly as important. I have unpublished data from BCI that show that our tipping buckets underestimate by 14%
  + **This is a valuable insight for our manuscript. We modify the text and now write “). Rain gauges also systematically can underestimate precipitation by ~9-23% due to wind effects and evaporation (Pollock et al., 2018)”**
* The way this sentence is phrased, it could be interpreted is referring to both ACP and STRI stations. ACP stations are electronic only. I can confirm that the ACP has been using (at least for the last couple of decades) tipping bucket-type gauges. They have changed the type of gauge once during this period. It would be better to refer to these as electronic sensors. On BCI we have a secondary electronic gauge that does not employ a tipping mechanism. STRI rainfall data were collected both manually and electronically only at BCI and Punta Culebra

**Thank you for the insights. We now write “Rainfall data were collected using electronic tipping buckets for ACP data, and both manual and electronic sensors for STRI data (Paton, 2022).”. Additionally, the mentioned BCI record is included in our analysis as BCIELECT station.**

* There is also the possible complication of sensor changes through time. STRI has used almost the same sensor for the last 30 years, but there may have been different devices used before then. I have reached out to the ACP to try to find out more information about what devices have been used - including if they are using tipping-type sensors.
  + **Thank you for requesting such valuable information. We wish we could list the types of sensor used in each of the records, however it is mostly incomplete. We would like to include the sensor type as part of the supplemental material.**
* You could reference Paton and Stallard here. We have a map showing the rainfall gradient in the Canal watershed. The map shows very steep gradients.
  + **The citation is relevant for the manuscript, we have added the following: “. Annual precipitation varies more than twofold due to a steep rainfall gradient from the drier Pacific to the wetter Caribbean side of the isthmus, as well as elevational rainfall variation (Muller-Landau et al., 2023; Paton and Stallard, 2023).”**
* 'lab clearing' is never defined above
  + **The correction has been included as follows: “BCICLEAR is the STRI meteorological record from the lab clearing area, which combines data from an automated tipping bucket with a manual rain gauge (an average of two manual gauges, with missing days filled by prorating the rainfall recorded at the end of the gap based on electronic sensor data).”**
* which is based on the average of two, professional-grade, manual rainfall gauges. Rainfall at these gauges are recorded an average of 3-4 times per week. Missing days are gap-filled based by prorating the rainfall recorded at the end of a gap using the electronic sensor data during the same period; assuming a standard manual gauge recording time of 9am, local time.
  + **See the response above.**
* Which ACP sites? Not possible to re-do this analysis without knowing which stations were included.
  + **We have added Supplemental table 2. Which contains all downloaded sites, source, coordinates and the analysis that were done.**
* There is another important factor relevant to the temporal and spatial variability - rainfall in this area can be very localized with storms frequently dropping significant amounts of rainfall (50mm or more) that completely miss locations only a few kilometers away.
  + **We strongly agree with this comment, we now state the following: “The steep precipitation gradient of Panama, and the narrow isthmus requires high resolution products. Products with coarser than 0.05 degrees resolution miss most of the local variability.”**
* This is assumed to be true only for electronically measured rainfall. Manually collected data are usually assumed to be 'true'.
  + **Thank you for the conceptual clarification.**
* I would also suggest that more research needs to be done on the problem of electronic sensor underestimates - differences between gauge models, and how climate might affect underestimates.
  + **We appreciate the suggestion.**