Reviews:

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

**First reviewer: Anonymous 1**

**Summary by reviewer:**

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.
  + **Response:**

**The manuscript discussion has been modified to give a more comprehensive recommendation supported by the evidence provided in the results section.**

* 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).
  + **The title has been modified to: Evaluation of downscaled gridded climate products in reproducing 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.
  + **Response: Due to the coarse resolution of this products and the narrow area of the study sites many sites will fall within the same pixel and the ultimate purpose of capturing high local variation would be far from achievable for this product. It would also add an extra amount of work to download and include the product into the paper, modifying figures and the manuscript itself.**
* 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.
  + **Response: statistical downscaling does include valuable inputs such as topography and wind fields and clouds, which are very influential in the precipitation spatial distribution. Panama in particularly has a very steep rainfall gradient that is underrepresented by coarser products. Our results show that this downscaled products closely approach the precipitation capture by rain gauges. We did perform resampling and reprojection of products at ~5km to match the 1 km resolution, in that case there is no value added since we preserve the values from the coarser resolution by doing nearest neighbor.**
* "TERRA" should be correctly called "TerraClimate" as it was originally published (Abatzoglou et al., 2018).
  + **Response: Correct, the correction was made.**
* TerraClimate is also available through the end of 2022.
  + **Response: individual years are available through 2022. 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.
  + **Response: This is fixed**
* 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.
  + **Response: I agree with the reviewer, we added chirps v2 at the very end of the analysis and it does perform well all around. It just has a coarser resolution than CHELSA and for that reason I think we recommend it. Also, it is not only precipitation, but we are also talking about dozens of variables with similar quality. On the other hand, CHIRPS is near real time whereas CHELSA has an end date.**
  + **Modifications to the text has been included since the time of this review.**
* 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).
  + **Is Response: we are aware of the effect of forcing datasets in the results depicted in our study. Some of the products used in this study do share forcing datasets and thus they show very similar results with little variation among them. Variation is large between different forcing datasets. Our results do show similar trends as shown by alexander et al 2020. In fact, the only satellite products we use is the one that has the best performance, CHIRPS V2.**
* 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.
  + **Response:**

We initially considered GPM IMERG (V06, V07, TRMM), but only cited TRMM in the supplemental materials. We found that the spatial resolution and quality were not suitable for this study.

Although the GPM TRMM temporal coverage starts in 2000 (like some datasets we used, like CHELSA EarthEnv 2003-2016), the product's resolution of approximately 30 km per pixel exceeds the 5 km maximum we set for our analysis. Additionally, we found notable quality issues in the Panama region. Specifically, the "gaugeRelativeWeighting" in the TRMM dataset (1998-2019) indicated a lower weighting of rain gauges for Panama compared to regions like Colombia and the United States, which have denser and more evenly distributed gauge networks.

When examining the GPM IMERG V06 and V07 versions, we found that the Quality Index monthly (QIm) values for Panama were high (around 30) between 2000 and 2008 but dropped significantly to around 5 thereafter. In version V07, QIm values between 2-10 are flagged as yellow which indicates that the bias corrections are reasonable but rely heavily on interpolation between stations (<https://gpm.nasa.gov/sites/default/files/2023-07/IMERG_TechnicalDocumentation_final_230713.pdf> in page 56). This inconsistency in the QI across the time series would likely result in inaccurate climatology, especially when averaged over a long period. In version V06, values above 4 are flagged as good (<https://gpm.nasa.gov/sites/default/files/2020-02/IMERGV06_QI_0.pdf>), however this stoplights are relatively new for 0.1 degrees.

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. Therefore, we believe that excluding GPM IMERG was necessary to maintain the scope of our analysis.

* + **To do: Find the product and demonstrate the bad quality of the data.**

**Figure and table comments:**

* Figure 1: ● What are the units? ● Suggest adding a North arrow ● Suggest increasing the font size of the stations
* 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.
* 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
* Figure 4: Suggest capitalizing Jan-April
* Figure 6: Suggest allowing the scale of the y-axis to vary by site.
* 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.
  + **Response: The comparisons between rain gauges and gridded product subsets for each of the given time extents. This ensures that our comparisons are consistent.**
* 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.
  + **Response: Overall, all included climatology are within the period of 1979 to 2016. The differences in time extent were addressed by subseting 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.
* 84: How were they resampled? Nearest neighbor, cubic convolution, bilinear … ?
  + **Response: nearest neighbor, this was incorporated into the manuscript.**
* 87: Is Table S2 supposed to be Table 1? There is no Table S1 referenced.
  + **Response: granted S2 should be called S1 or supplemental 1 which is the extended version of table 1.**
* 90: Kind of odd to compare dry season, but not wet season. It would be useful to justify this choice further.
  + **Response: Dry season length and precipitation is especially relevant to ecologist. The precipitation values of the whole year are parallel to the wet season precipitation. Furthermore, we reported on the 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.
  + **Response: we provide with citation to lmr4 package for that equation. About 10% of the ground data is gap filled as shown in the full code provided in my GitHub page. Intermediate tables can be reproduced with the code provided. Formula has been added to the text**
* 118: Readers might not know what the "lab clearing" is.
  + **Response: name of the station available in figshare**
* 190: I am unclear about how the authors came to the conclusion 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.)
  + **Response: I agree with this.**

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.

**Response: 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.

**Response: no revision needed.**

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.

**Response: Indeed, wet season is not included in the manuscript. Partially, because 8/12 months are wet season and the annual precipitation variables is as it is very similar to that of the wet seasons results.**

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).”

**Response: this is true, but I guess we want to highlight the barro Colorado island as an important site for our study**

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.

**Response: a supplemental table with the sites considered and the sites used should help with this.**

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.

**Response: code availability remains a major to do item.**

(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.

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%
* 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%
* The way this sentence is phrased, it could be interpreted that it 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 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.

* 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.
* 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.
* 'lab clearing' is never defined above
* 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.
* Which ACP sites? Not possible to re-do this analysis without knowing which stations were included.
* 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.
* This is assumed to be true only for electronically measured rainfall. Manually collected data are usually assumed to be 'true'.
* 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.