Technical Note: Evaluation and bias correction of an observations-based global runoff dataset using streamflow observations from small tropical catchments in the Philippines

Daniel E. Ibarra[1,2], Carlos Primo C. David[3], Pamela Louise M. Tolentino[3]

[1] Department of Earth and Planetary Science, University of California, Berkeley, California 94720 USA

[2] Institute at Brown for Environment and Society and the Department of Earth, Environmental and Planetary Science, Brown University, Providence, Rhode Island 02912 USA

[3] National Institute of Geological Sciences, University of the Philippines, Diliman, Quezon City, Philippines 1101

Correspondence to: Daniel E. Ibarra (dibarra@berkeley.edu) and Carlos Primo C. David (cpdavid@nigs.upd.edu.ph)

Abstract

In relatively wet tropical regions, seasonal fluctuations in the water cycle complicate the consistent and reliable supply of water for urban, industrial and agricultural uses. Importantly, historic streamflow monitoring datasets are crucial in assessing our ability to model and subsequently plan for future hydrologic changes. In this technical note we evaluate a new global product of monthly runoff (GRUN; Ghiggi et al., 2019) for small tropical catchments in the Philippines. In particular, we evaluate the observations-based monthly runoff product using archived monthly streamflow data from 55 catchments with at least 10 years of data, extending back to 1946 in some cases. Since GRUN didn't use discharge data from the Philippines to train/calibrate their models, the data presented in this study, 11,915 monthly data points, provide an independent evaluation of this product. We demonstrate across all observations significant but weak correlation ($r^2 = 0.372$) between the GRUN predicted values and observed river discharge, and somewhat skilful prediction (Volumetric Efficiency = 0.363 and log(Nash-Sutcliff Efficiency) = 0.453). GRUN performs best among catchments located in Climate Types III (no pronounced maximum rainfall with short dry season) and IV (evenly distributed rainfall, no dry season). We also found a weak negative correlation between volumetric efficiency and catchment area, and a positive correlation between volumetric efficiency and mean observed runoff. Further, analysis for individual rivers demonstrates systematic biases (over and under) in baseflow during the dry season, and under-prediction of peak flow during some wet months for most catchments. Finally, to correct for underprediction during wet months we perform a log-transform bias correction which greatly improves the nationwide Root Mean Square Error between GRUN and the observations by an order of magnitude (2.648 vs. 0.292 mm/day). This technical note demonstrates the importance of performing such corrections when accounting for the proportional contribution of smaller catchments in tropical land such as the Philippines to global tabulations of discharge. These results demonstrate the potential use of GRUN and future data products of this nature after consideration and

correction of systematic biases to: 1) assess trends in regional scale runoff over the past century, 2) validate hydrologic models for un-monitored catchments in the Philippines, and 3) assess the impact of hydrometeorological phenomenon to seasonal water supply in this wet but drought prone archipelago.

1 Introduction

The global water crisis is considered one of the three biggest global issues that we need to contend with; it affects an estimated two-thirds of the world's population (Kummu et al., 2016; WEF, 2018). Among the sources of freshwater, the most important compartment in terms of use is surface water. It is the primary resource for irrigation, industrial use and provides the bulk of water supply for many large cities. Long term streamflow datasets are useful for resource management and infrastructure planning (e.g., Evaristo and McDonnell, 2019). Such data is even more critical in areas that rely on run-of-the-river supply and do not use storage structures such as dams and impoundments. Further, a robust, long term dataset is crucial in the face of increased variability in stream discharge due to land use change, increased occurrence of mesoscale disturbances and climate change (e.g., Abon et al., 2016; David, et al., 2017; Kumar et al., 2018). In the absence of long-term streamflow datasets for most locations in the world, several researchers have compiled datasets worldwide which are used to extrapolate streamflow in non-gauged areas (Maybeck et al., 2013, Gudmundsson et al., 2018, Do et al., 2018; Alfieri et al., 2020; Harrrigan et al., 2020). Several global hydrological models have also been created to project variations in streamflow and extend present-day measurements to the future (Hagemann et al., 2011; Davie et al., 2013; Winsemius et al., 2016). The latest contribution to modelled global runoff products is the Global Runoff Reconstruction (GRUN) (Ghiggi et al., 2019). GRUN is a global gridded reconstruction of monthly runoff for the period 1902-2014 at 0.5 degree (~50km by 50km) spatial resolution. It used global streamflow data from 7,264 river basins that to train a machine learning algorithm which learn the runoff generation processes from precipitation and temperature data.

There is a disparity in the availability of long-term gauged rivers datasets between continental areas and smaller island nations, which have more dynamic hydrometeorologic system owing to the size of the catchments and proximity to the ocean (e.g., Abon et al., 2011; Paronda et al., 2019). However, the impact of climate change on the hydrological cycle can be observed the most for tropical island nations (Nurse et al., 2014). The Philippines offers a unique example where manual stream gauging programs have started in 1904 and, while spotty at times, have continued to today. This island nation on the western side of the Pacific Ocean is characterized by a very dynamic hydrologic system because it is affected by tropical cyclones, seasonal monsoon rains, sub-decadal cycles such as the El Nino Southern Oscillation (ENSO) and climate change (Abon et al., 2016; David, et al., 2017; Kumar et al., 2018).

This technical note evaluates the accuracy of the GRUN dataset (GRUN_v1) as applied to the hydrodynamically-active smaller river basins in the Philippines. Additionally, it explores the possible hydrologic parameters that may need to be considered and/or optimized such that global datasets are able to predict runoff in smaller, ungauged basins more accurately.

2 Dataset and Methods

2.1 GRUN observations-based global gridded (0.5°x0.5°) runoff dataset

GRUN is a recently published global reconstruction of monthly runoff time series for the period 1902 to 2014. It was created using a machine learning algorithm based on temperature and precipitation data from the Global Soil Wetness Project Phase 3 (GSWP3; Kim et al., 2017; http://hydro.iis.u-tokyo.ac.jp/GSWP3/index.html) using the Global Streamflow Indices and Metadata Archive (GISM) (Ghiggi et al., 2019). In this contribution we analysed GRUN v1 (https://figshare.com/articles/GRUN_Global_Runoff_Reconstruction/9228176; accessed September 9th 2019) which was trained on a selection of catchments with area between 10 and 2500 km^2 GSIM (Do et al., 2018; Gudmundsson et al., 2018) and validated using 379 large (>50,000 km^2) monthly river discharge datasets from the Global Runoff Data Centre (GRDC) Reference Dataset (https://www.bafg.de/GRDC/EN/04_spcldtbss/43_GRfN/refDataset_node.html). Additionally, due to the criteria for training data, GRUN's calibration is biased towards the northern hemisphere mid-latitudes because data are available for only a few sites in the tropics in Africa and southeast Asia (Mulligan, 2013). Ghiggi et al. (2019) discuss that because of the dataset training technique, uncertainty scales with the magnitude of runoff. GRUN is likely to have high prediction uncertainty in regions with less dense runoff observations such as in tropical southeast Asia. However, they show for southeast Asia an increase in runoff and a strong correlation of runoff with ENSO for the period of analysis (1902 to 2014). We refer the reader to Ghiggi et al. (2019) for more information but note that because of the catchment size filtering criteria, none of the GISM and GRDC data from the Philippines were used in calibration or evaluation (*Personal Communications*, G. Ghiggi, 2019). As such, we view our analysis as a completely independent test of the GRUN runoff prediction for small tropical catchments.

2.2 Historical streamflow observations

In this contribution we analyse monthly observations of discharge from 55 manually observed streamflow stations from three Philippine datasets. The observations span the period between 1946 to 2016, although only data through 2014 are used due to the time period included in GRUN. All datasets needed to include at least 10 years of data. The location of all streamflow stations is shown on Figure 1 and listed in Table 1.

2.2.1 Bureau of Research and Standards (BRS) Dataset

The discharge data was originally acquired from the Bureau of Research Standards (BRS) under the Department of Public Works and Highways (DPWH). The records keeping was transferred to the Bureau of Design, also under DPWH, which continues to record gage data from some rivers up to today. A majority of the reprocessed BRS data used in this analysis come from Tolentino et al. (2016), however, some of the datasets were subsequently updated using data available from the Department of Public Works and Highways. A discussion of the accuracy of this data based on comparison to manual daily discharge measurements can be found in Tolentino et al. (2016).

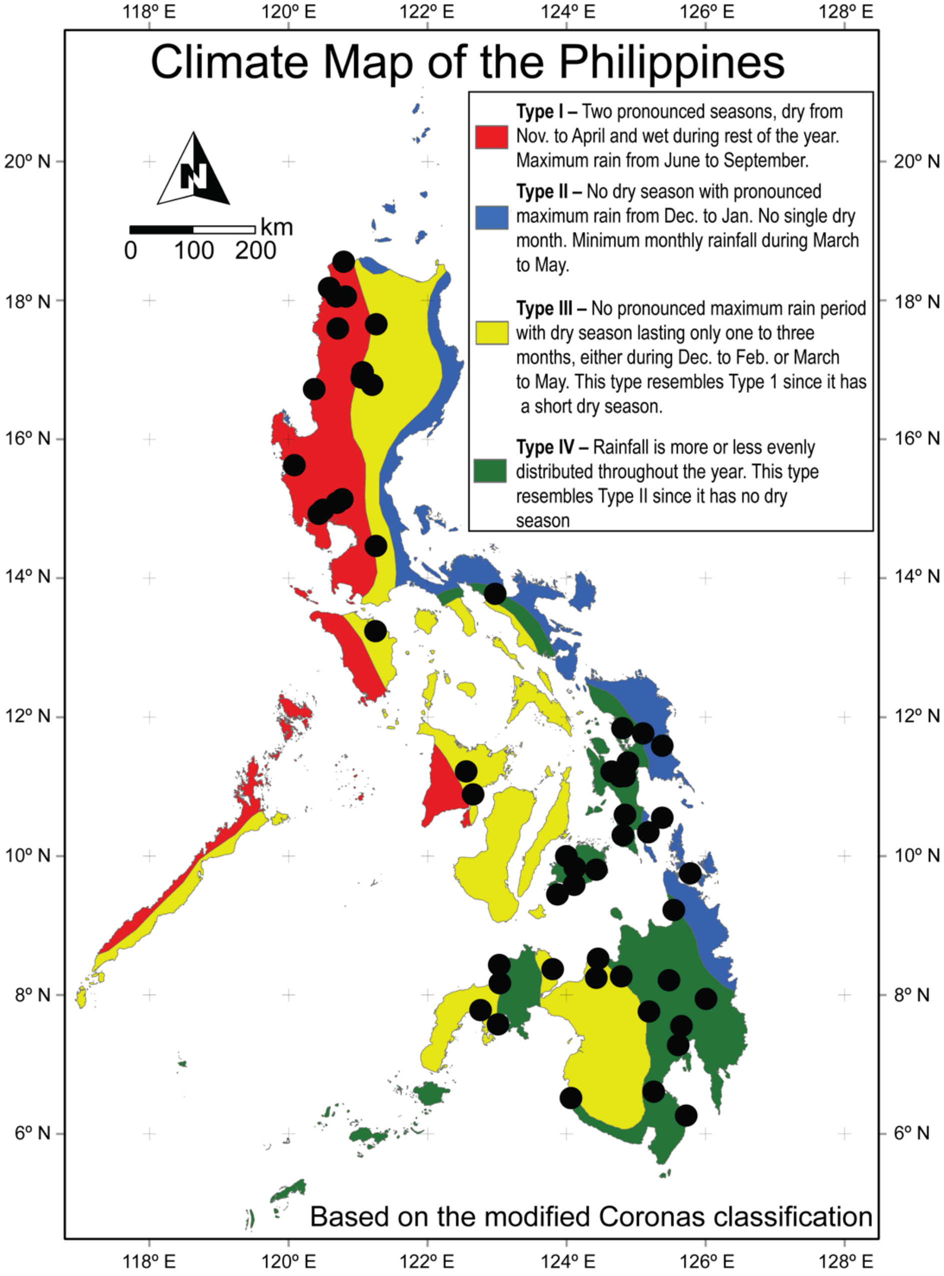
Climate Map of the Philippines
Type I – Two pronounced seasons, dry from Nov. to April and wet during rest of the year. Maximum rain from June to September.
Type II – No dry season with pronounced maximum rain from Dec. to Jan. No single dry month. Minimum monthly rainfall during March to May.
Type III – No pronounced maximum rain period with dry season lasting only one to three months, either during Dec. to Feb. or March to May. This type resembles Type 1 since it has a short dry season.
Type IV – Rainfall is more or less evenly distributed throughout the year. This type resembles Type II since it has no dry season
N
0 100 200 km
118° E
120° E
122° E
124° E
126° E
128° E
20° N
18° N
16° N
14° N
12° N
10° N
8° N
6° N
Based on the modified Coronas classification

Figure 1: Map of Philippines with the location of streamflow stations used in this analysis overlaid on climatic type (as in Tolentino et al., 2016; Kintanar, 1984; Jose and Cruz, 1999). Note that no publicly available long-term stations are available for Palawan.

2.2.2 Global Runoff Data Centre (GRDC) Reference Dataset

Ten catchments from the GRDC Reference Dataset (https://www.bafg.de/GRDC/EN/04_spcldtbss/43_GRfN/refDataset_node.html; requested July 2019) were analysed. Over 45 sites from the Philippines are available in the GRDC data; however, almost all do not fulfil our criteria of having over 10 years of data. Four of these catchments match or extended the BRS datasets, and one extends a GSIM dataset (see below). Notably four of the time series available from GRDC are available back to the 1940s (Table 1).

2.2.3 Global Runoff Data Centre (GSIM) Reference Dataset

Only two time series of the available five GSIM time series (Gudmundsson et al., 2016, Do et al., 2018) contain more than 10 years of data.

2.3 Criteria for Inclusion of Datasets

All catchment areas were verified using the digital elevation model from the 2013 Interferometric Synthetic Aperture Radar (IfSAR) data. All runoff datasets were normalized (mm/yr), i.e., ‘specific discharge’. We only considered streamflow stations where the published and verified areas agreed. Catchment areas span 4 orders of magnitude (8.93 to 6,487 km^2) and cover the majority of the Philippines excluding Palawan (see Figure 1). The location of catchments was paired to GRUN grid cells (0.5° by 0.5°) for the analysis. Instead of computing the weighted area runoff over the catchment, we employed the nearest neighbour interpolation between the catchment outlet location and the GRUN gridded product (0.5° by 0.5° resolution). All but one catchment is smaller than the area of the GRUN grid cells (~2,500 km^2), thus, we view this pairing as sufficient for validation purposes. This assumption was tested by interpolating the GRUN grid to the gauging location as well as the watershed centroids, but this did not lead to a significant difference in correlation.

2.4 Statistical Performance Metrics and Tools

To assess the performance of GRUN we use a suite of metrics commonly used to assess model performance in hydrologic studies. These metrics are calculated for each individual catchment (n=55) and in aggregate for each climate type (n=4; see below) shown in Figure 1.

Firstly, we use the commonly used square coefficient of determination (r^2 or r-squared). This bivariate correlation metric measures the linear correlation between two variables. In this case the predicted monthly values from GRUN and the observed monthly values from the streamflow datasets. It varies from 0 (no linear correlation) to 1 (perfect correlation). The use of r-squared does not account for systematic over- or under-prediction in runoff because it only accounts for correlation among the observed and predicted values (see Krause et al. (2005) for further discussion of the use of r-squared in hydrological model assessment).

The second metric used here is the Volumetric Efficiency (VE) (Criss and Winston, 2008), utilized previously by Tolentino et al. (2016) on a subset of the BRS catchments analysed here. VE is defined as:

$$VE = 1 - \frac{\sum Q_P - Q_O}{\sum Q_O} \tag{1}$$

where P is the modelled/predicted values and O is the observed runoff values. A value of 1 indicates a perfect score. Because we are interested in the performance of GRUN over the period of each streamflow record, we calculate VE using all paired monthly observed and simulated values rather than the monthly medians that were used in Tolentino et al. (2016). This results in lower VE scores than those reported by Tolentino et al. (2016).

Further, we use both the linear and logged Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe (1970):

$$NSE = 1 - \frac{\sum (Q_P - Q_O)^2}{\sum (Q_O - mean(Q_O))^2} \tag{2}$$

The NSE can vary between -∞ and 1 (perfect fit). NSE values are useful because values less than zero indicate that the model is no better than using the mean value of the observed data as a predictor. NSE is also calculated using logarithmic values of runoff to reduce the influence of a mismatch during peak flow and to increase the influence of low flow values (see further discussion in Krause et al., 2005).

To evaluate a possible strategy for performing a bias correction of the GRUN simulated values at a countrywide scale, we use the Root Mean Square Error (RMSE) in units of runoff (i.e., mm/day). The RMSE is applied to the raw GRUN simulated values and the observation based bias corrected GRUN values at the country, climate type (see below) and individual catchment level.

Finally, to evaluate distributions of flow duration using flow duration curves by catchment and aggregated by climate type we use the ‘fdc’ function in the R package ‘hydroTSM’ (Zambrano-Bigiarini, 2020), and include in our comparison GRUN predicted values only for months which observations are available.

2.5 Climate Types

The Philippines has four Climate Types (see also Abon et al., 2016; Tolentino et al., 2017; Figure 1): Type I Climate on the western seaboard of the Philippines is characterized by distinct wet (May to October) and dry (November to April) seasons; Type II Climate on the eastern seaboard has no distinct dry period with maximum rainfall occurring from November to February; Type III inland climate experiences less annual rainfall with a short dry season (December to May) and a less pronounced wet season (June to November); and, Type IV southeast inland climate experiencing depressed rainfall and is characterized by an evenly distributed rainfall pattern throughout the year.

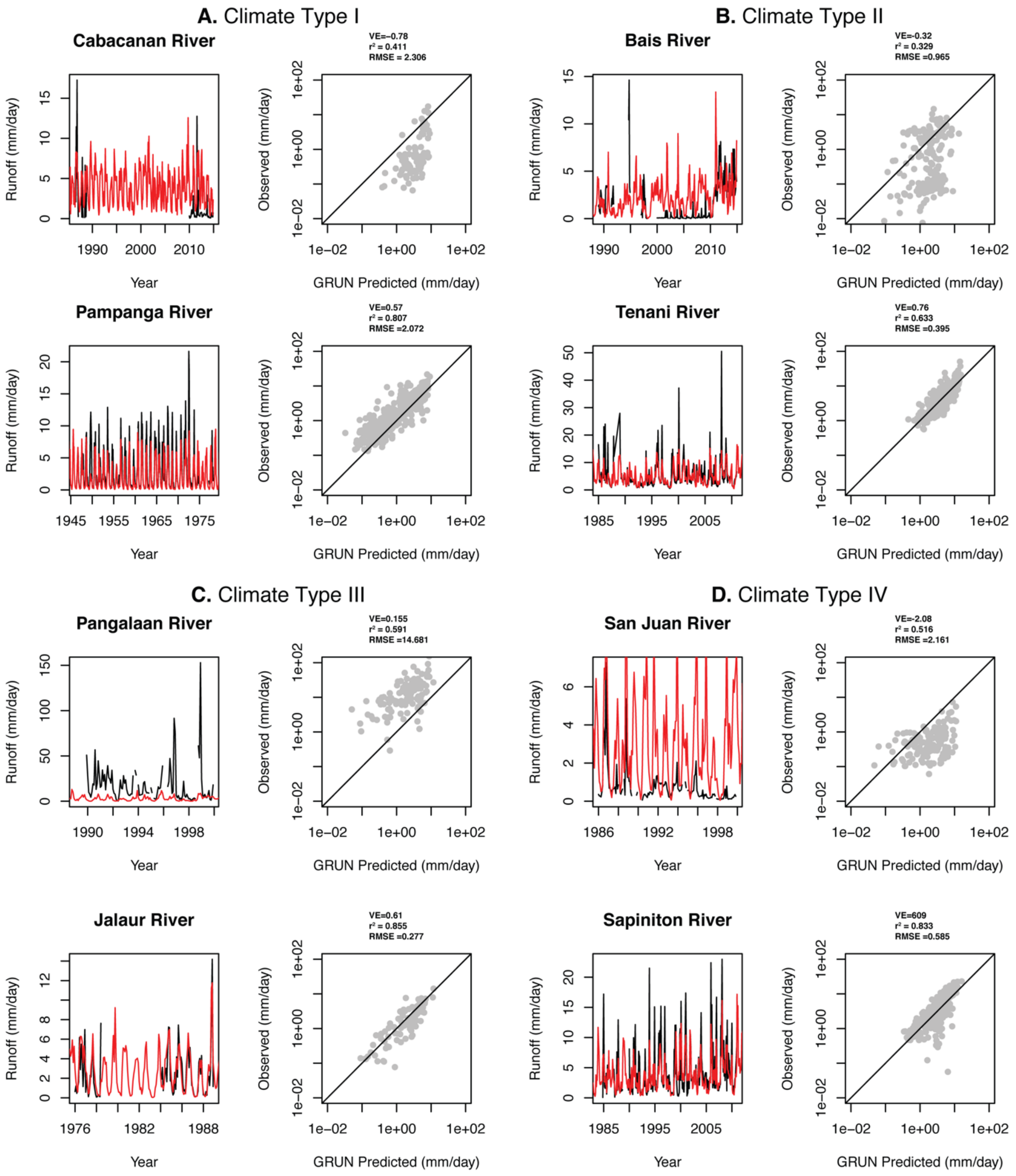

Figure 2. Example timeseries of GRUN predicted (red lines) and observed (black lines) runoff values, and cross-plots (log-scale) with VE, r^2 and RSME values for the worst (top) and best (bottom) performing river basins within Climate Types I, II, III and IV (panels **A-D**, respectively).

3 Results and Discussion

Figure 2 and the supplemental figures (Figure A1) show time series comparisons between the GRUN runoff values and runoff (area normalized discharge). Statistics performance metrics across all data as well as by climate types I to IV are listed in Table 1. Given the emphasis on a country scale evaluation of GRUN we primarily focus below on results in aggregate grouped by climate type or for all catchments. In the following sections we break down the comparison between the streamflow observations and GRUN by reporting summary statistics, comparing runoff distributions and extreme values at the individual basin level and by Climate Type, analyse flow duration curves, and finally look at several correlations of VE to watershed characteristics. Following this we calculate bias correction regressions and provide an outlook for future work.

3.1 Comparison of runoff distributions

Average runoff values among all catchments are somewhat well predicted by GRUN. Across all observations the r-squared of the correlation between GRUN prediction and observation was of 0.372 and VE of 0.363 (Table 2). Using log(runoff) values (following Criss and Winston, 2008) this improves to an r-squared of 0.546 and a volumetric efficiency (VE) of 0.733, suggesting reasonable utility in the GRUN product at a country scale for the Philippines, despite no training data from the Philippines being used in the creation of GRUN. The RMSE across the dataset was 2.648 mm/day (Table 2). NSE and NSE-log10 values, which scored overall values across the dataset of 0.091 to 0.453, respectively, ranged, from -10.70 to 0.68 and -11.53 to 0.76, respectively, for individual catchment comparisons, with median values of 0.02 and 0.24, respectively. More than half of the catchments, 29 of 55 scored NSE values greater than 0, with only 5 catchments scoring values greater than 0.5. Similarly, 32 of 55 catchments scored NSE-log10 values greater than 0, with 12 catchments scoring values greater than 0.5.

In Figure 3A median values of runoff (black dots) given a volumetric efficiency (VE) metric of 0.509 across all catchments, the average (mean) difference between the observed median and simulated values is +16%. The median and interquartile ranges (IQR, 25% to 75%) shown in Figure 2E overlap between GRUN and the observations. For five catchments of large (n=2) and relatively small sizes (n=3) the IQR of the observations does not overlap with the GRUN runoff IQR. The three small catchments are Climate Type III (yellow) and the two large catchments are Climate Type IV. In two catchments of moderate size the GRUN IQR is greater than the observed IQR runoff range. Looking at extreme monthly values (maximum and minimum) over the period of observation demonstrates significant underprediction in wettest conditions (orange dots in Figure 3A and 3D) with almost all catchments' maximum observations falling above the 1:1 line and a lower VE score (compared to the median runoff VE score) for maximum values of 0.194. The minimum values plot around the 1:1 line and are more evenly distributed, however the VE score of 0.154 is similarly low due to greater spread than the median values.

Regardless of Climate Type, a general underestimation of the model is seen for the highest runoff months, looking at the distributions by basin. This is especially evident in Climate Types I and II with pronounced wet seasons as also shown by their lower r-squared values (Figure 4) and lower VE values. Climate Type II also has the highest RMSE value of 4.55 mm/day (compared to an average observed flow of 9.03 mm/day). Climate Types III and IV have comparable r-squared and VE values though skewness towards underprediction during the highest runoff months is still evident, particularly for Climate Type IV. These patterns are particularly evident looking at Figure 5, which shows flow duration curves (FDC) by climate type (see Figure A2 for individual catchments). Such an analysis allows for inspection of runoff distributions and biases across the range of observed and predicted values. At low flow (high exceedance probability, >80%) there is reasonable agreement in the shape and magnitude of the distributions for Climate Types I and IV between GRUN and the observations (bottom right of the FDC plots). Climate Type III demonstrates consistent bias across all runoff values less than an exceedance probability of ~90%. At high flow (low exceedance probability, <20%), as also noted above, runoff for all Climate Types is underestimated by GRUN, with Climate Types I and II showing the greatest discrepancy (top left of each FDC plot).

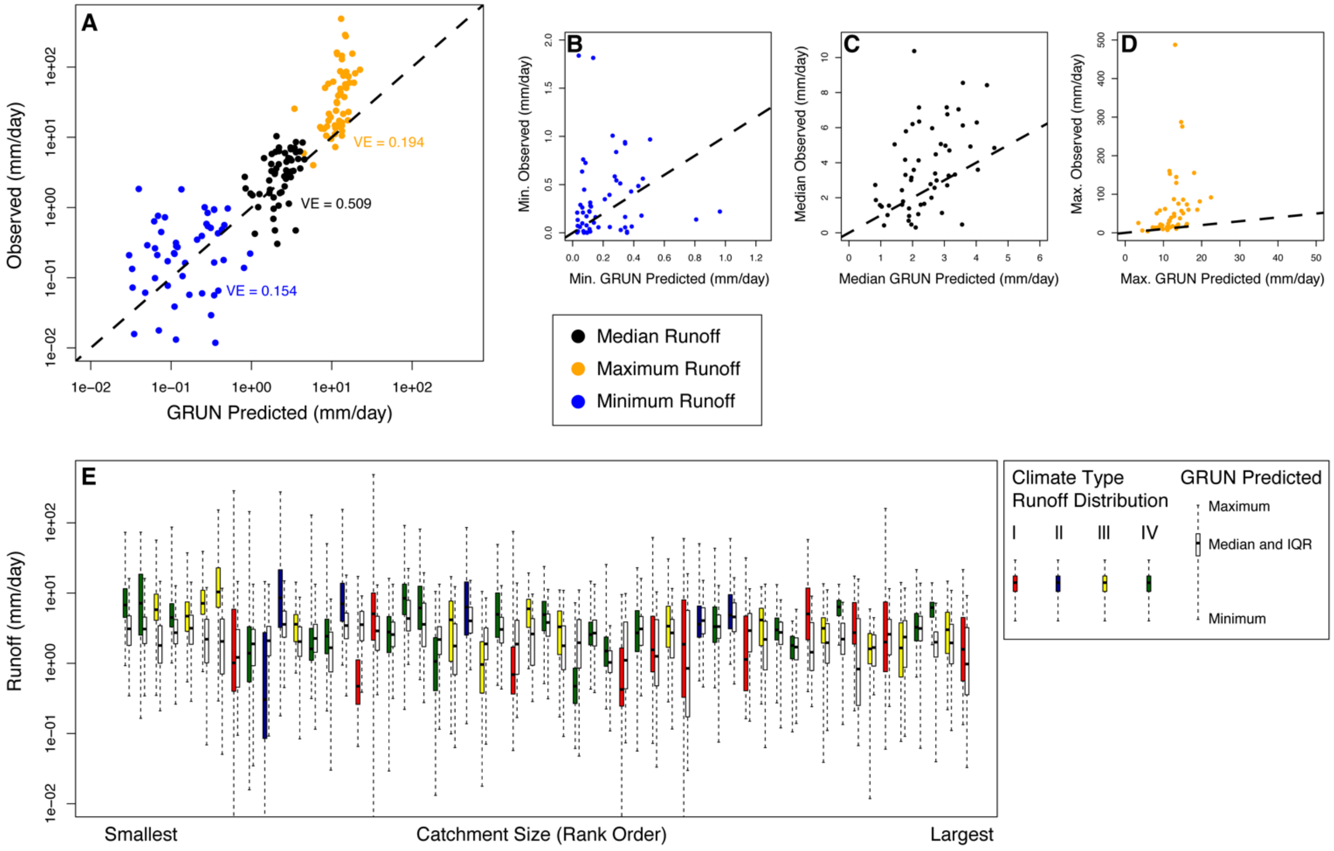

Figure 3. Comparison of runoff ranges and distributions. **(A)** Comparison of median and extreme (maximum and minimum) monthly values between the observations and GRUN in log space. **(B-D)** As in **(A)** for the minimum, median and maximum monthly values, respectively, in linear space. **(E)** Distribution of runoff between observations (coloured) and GRUN (white). Plots show the median (line), interquartile range (box) and maximum/minimum values (whiskers). The boxplots for GRUN only include months for which observations are available.

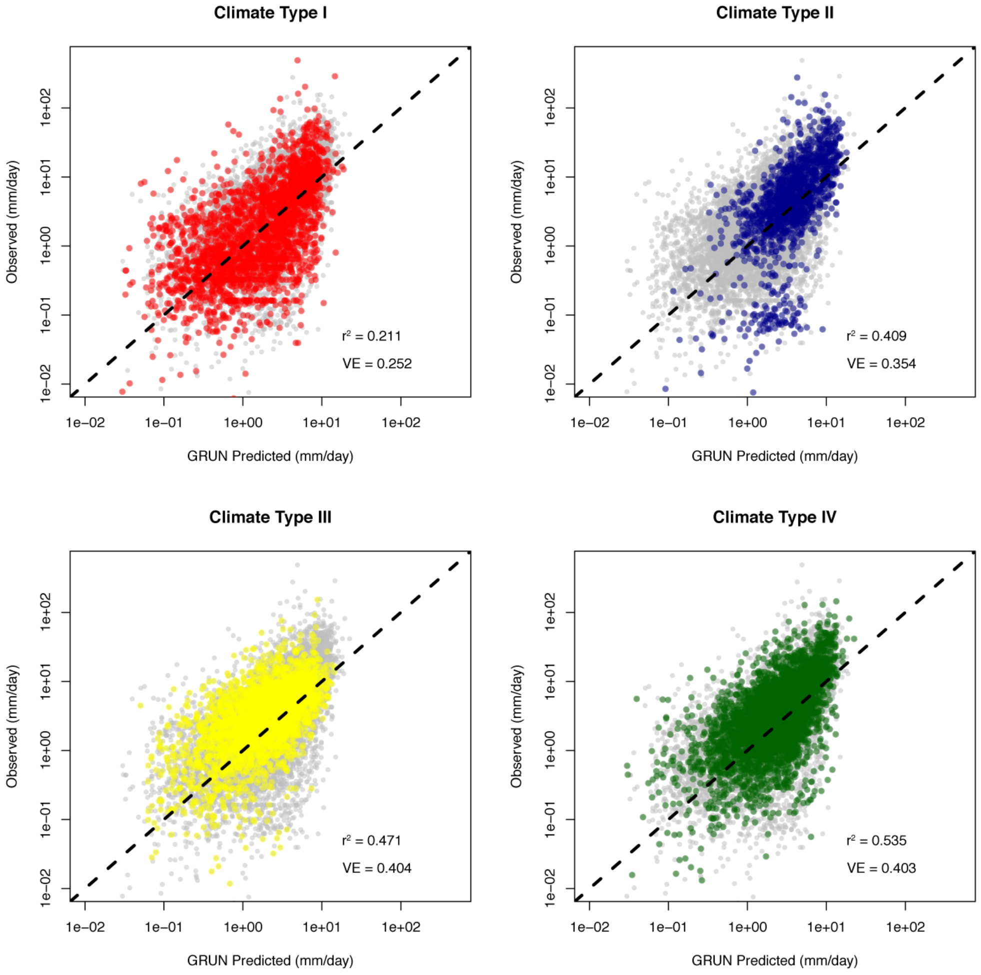

Figure 4: Cross plots of GRUN predicted vs. observed monthly runoff by climate type (see Figure 1 for climate type distributions). Grey dots represent all data, colored dots represent data points from that region. The squared pearson correlation coefficient (r^2) and volumetric efficiency (VE) metrics are listed for each panel.

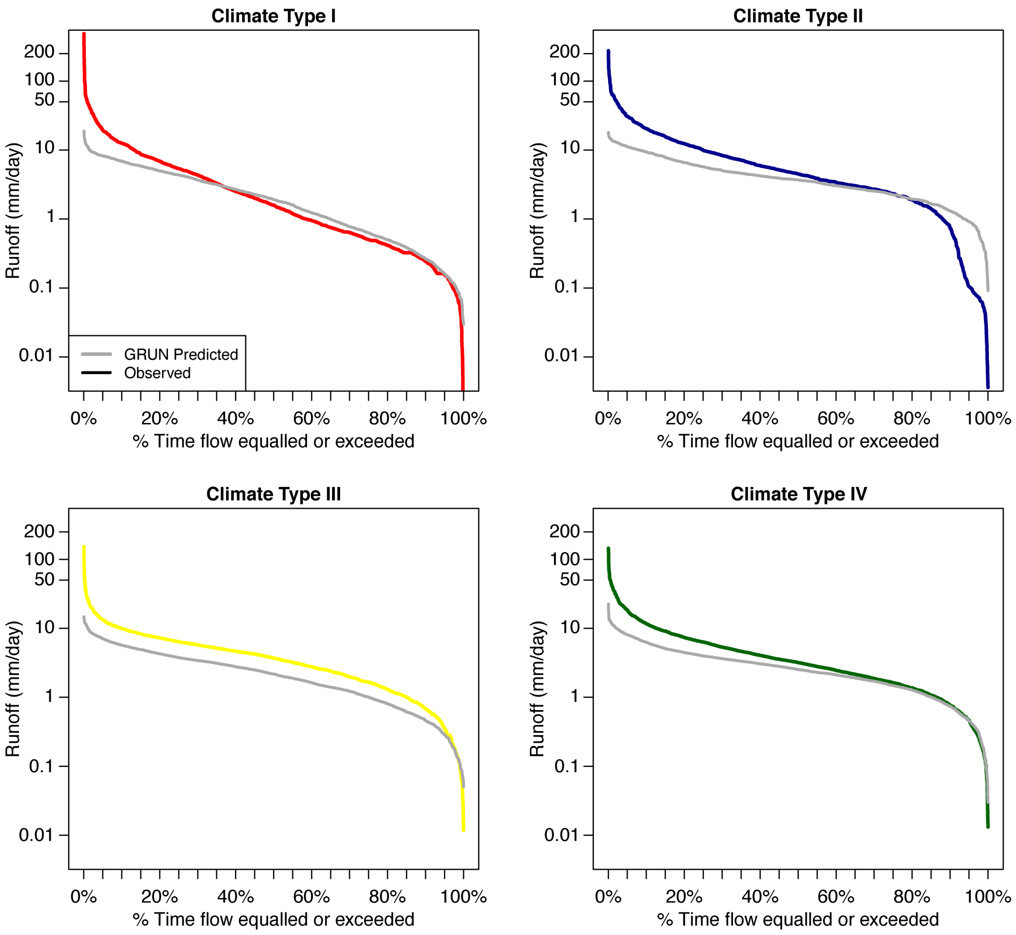

Figure 5. Flow Duration Curves (FDC) by Climate Type. For each Climate Type the observed (coloured) and GRUN predicted (grey) runoff distributions are shown. These plots represent the rank ordered data shown in Figure 4. FDC's for individual catchments are compiled in Figure A2.

3.2 Correlation and trends with watershed characteristics

The biases noted above are likely due to the high uncertainty and underprediction in monsoonal precipitation inputted into GRUN. There is a significant negative correlation (at $p<0.01$, r^2 =0.391) between log values of maximum runoff difference (observed minus predicted) versus catchment area (not shown). This suggests two possibilities: first, that particularly for small catchments which may have steeper average slope, GRUN underpredicts monthly runoff values during the wet season due to the precipitation datasets used to create GRUN; and second model-data agreement improves with catchment size. In this section we explore these possibilities.

There was a weak positive correlation ($r^2 = 0.041$, $p = 0.137$) between VE and log(catchment size) (Figure 6) and a stronger negative correlation with mean runoff ($r^2 = 0.182$, $p < 0.01$). However, at low runoff there is significant spread in VE score, driven primarily by Climate Type I catchments (red box and whisker plot in Figure 6C). These catchments experience distinct wet and dry seasons and are dominantly located in the northwest Philippines. The positive correlation with catchment size is likely primarily due to the extreme wet months. This is particularly evident from the Nash-Sutcliff Efficiency (NSE) and log(Nash-Sutcliff Efficiency) (NSE-log10) scores in Table 2 and Table A1. Since NSE puts more weight on large flow (Criss and Winston, 2008), it is not surprising that our NSE-log10 scores are better because extremely wet months are weighted less than using the raw runoff values compared to raw NSE scores. It is also notable that the VE scores using log10 values across the entire dataset is are better (0.363 vs. 0.733; Table 2). The physical significance of these observations could be that for large basins the time of concentration of any given flood event will be much longer, thus flood peaks will be wider and subdued due to infiltration into the shallow aquifers. However, a more likely explanation is that the average rainfall intensity is too low in the GSWP3 precipitation data used in producing GRUN. A bias at high and moderate flow conditions is particularly evident in the flow duration curves (Figure 5) and is likely due to downscaling from averages over larger areas with topographic complexity. While GSWP3 uses downscaled 20th century reanalysis products (Kim et al., 2017) at T248 resolution (~0.5º, the same resolution as GRUN), the topographic complexity of tropical islands such as the Philippines on the sub-0.5º scale would likely results in smoothing of the variability and lowering of the absolute magnitude of the precipitation fields, particularly during the wet season and during large synoptic precipitation events during the monsoon season. Further, because of the monthly (rather than daily or sub-daily) output of GRUN and the comparison carried out here, the biases observed in our analysis could be because of averaging of the input products and/or discharge datasets to monthly values. Finally, while the GSWP3 precipitation inputs are bias-corrected using the Global Precipitation Climatology Centre precipitation network, previous work has highlighted data quality issues with some historical data from the Philippines (Schneider et al., 2014).

Previous studies have investigated the correlation between runoff and catchment size (Mayor et al., 2011), and the different hydrologic and geologic factors that cause non-linear relationships between these two variables (Rodriguez et al., 2014). Recently, Zhang et al. (2019) point out that runoff coefficients increase logarithmically as catchment size decreases. Moreover, the same paper reports for smaller catchments the effects of vegetation cover, slope, and land use are larger compared to larger catchments. This implies that predictability of basin runoff for smaller catchments are more difficult due to the variances in the compounding factors mentioned above. We hypothesize that these effects proposed by Zhang et al. (2019) also influence the Philippines streamflow dataset use in this study. As such, we suggest that GRUN, is a useful new tool for studying trends, seasonality and average runoff from tropical catchments in the Philippines (such as in previous work: e.g., Merz et al., 2011; Wanders and Wada, 2015). However, we qualify this finding by noting, based on the limitations discussed above, that GRUN is not suitable for extreme value analyses associated with major tropical storms during the wet seasons unless suitable bias corrections (see next section) can be effectively carried out.

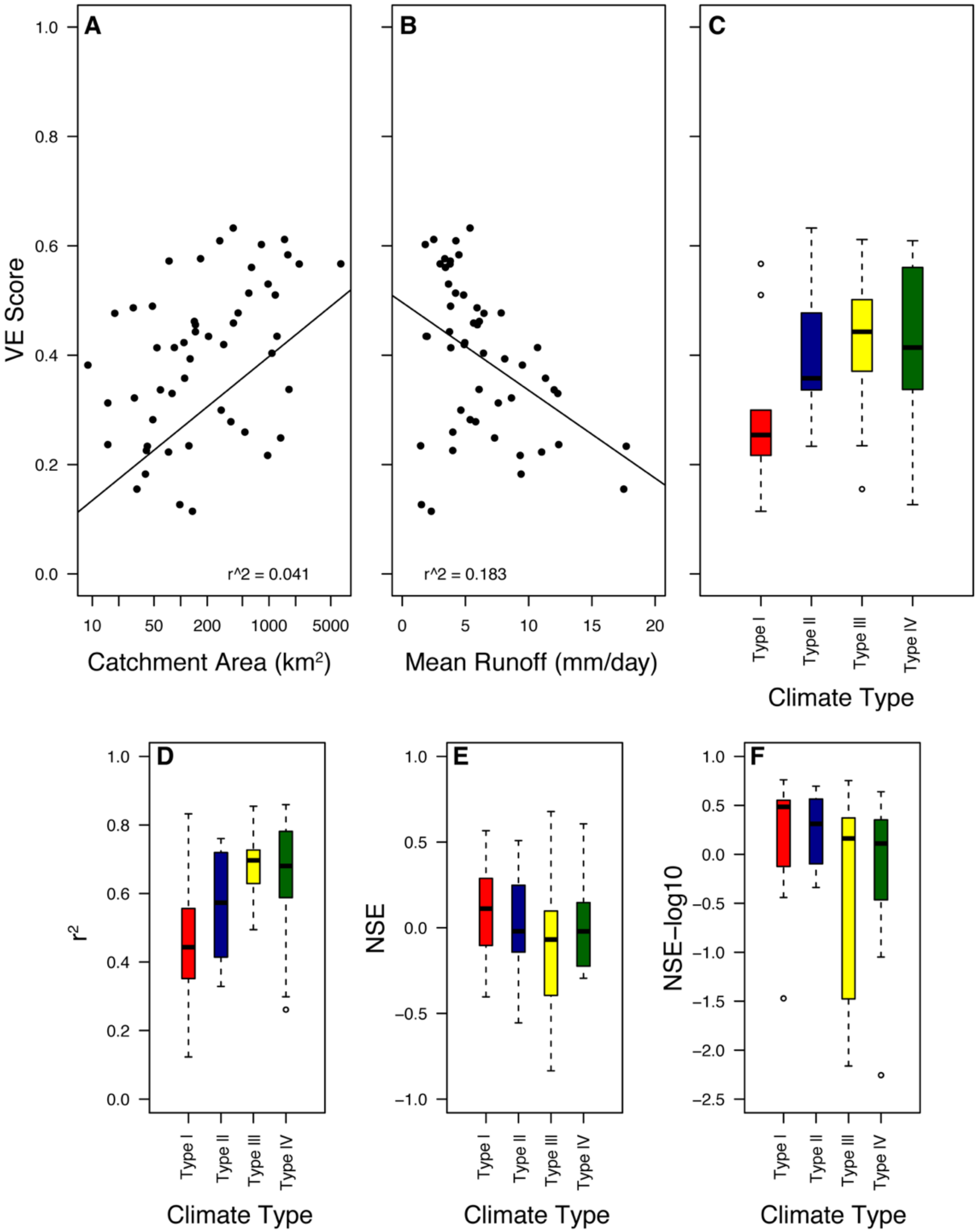

Figure 6: Diagnostic plots of volumetric efficiency (VE) results and comparison across metrics. Cross plots show the correlation of VE with **(A)** catchment area and **(B)** mean runoff. **(C)** Box and whisker plots show data from distribution of VE by climate type. Box and whisker plots show the median, interquartile range and 95% confidence intervals and outliers (dots). The regression in (A) is between the VE scores and ln(Catchment Area). **(D-F)** As in **(C)** for r^2, NSE and NSE-log10.

3.3 Bias Correction and Outlook

Overall, GRUN underestimates the actual observed runoff for Philippine basins. The GRUN dataset predicts a range of 0 to 10mm/day for most basins and up to 20mm/day for larger basins. The observed maximum runoff values are on average higher and exceed 50mm/day during months with extreme rain events (Figure 4). Furthermore, the GRUN dataset also appears to underestimate minimum flow in streams from highly seasonal catchments (e.g. Types I and II).

The underestimation of runoff values during extreme rain events may be a result of the fast saturation of the overlying soil and exceedance of rates of infiltration partly as a result of shallow aquifers filling up and consequently the conversion of excess rainfall into direct runoff (e.g., Tarasova et al., 2018). On the other hand, the underestimation of flow during low flow events may be a result of not accurately accounting for stream baseflow which is fed by shallow aquifers, as well as other effects such as land use and surface properties. These effects may be buffered in larger catchments leading the increase in that model-data agreement with catchment size (Figure 6B). Alternatively, this could be due to an underestimation in the wet season rainfall products from GSWP3 used to create GRUN as described above.

Given the biases and in particular the clear underprediction of streamflow in GRUN during the wettest months we perform a bias correction of the GRUN dataset at a nationwide level using all the available data used in our analysis. We do so in a two-step process to both correct the mean offset and stretch the wettest months to higher values with all transformations occurring in log-transform space (i.e., as displayed in cross plots in Figures 2 and 4). Thus, we first add the mean log10(runoff) difference between the observations and the predicted values (0.117 ± 0.045). Following this, using the *lm* function in R, we fit a linear regression between the observations and the GRUN predicted values (log10(runoff, observed) = m × log10(runoff, predicted) + b) and correct the predicted values using the slope (m=0.774 ± 0.058) and intercept (b=0.099 ± 0.030) derived from this regression. Uncertainties reported here are 68% confidence intervals and were assessed by bootstrap resampling observation and prediction pairs from 20 catchments (vertical line in Figure 7) without replacement 10,000 times. In Figure 7 we show the influence of including an increasing number of catchments from the dataset in our bootstrap resampling to assess how the mean value of the coefficients asymptote as more catchments are included. While the mean log10(runoff) offset is relatively unaffected, the slope and intercept of the bias correction do not asymptote until more than 10 catchments are included in the analysis. Finally, a leave-one-out approach (i.e., calculating the coefficients with 54 of 55 catchments) indicates 68% confidence uncertainties of ± 0.005, ± 0.006, and ± 0.003 for the log10(runoff) offset, the slope and intercept, lower than those reported above, as expected.

By carrying out these calculations in log-transform space the highest GRUN runoff values are the most affected, which are the data points that were most underpredicted (Figures 3A and 4). Because these corrections were carried out in log10 space statistical bias in the form of underestimation is possible (Ferguson, 1986). Following Ferguson (1986) we calculate the unbiased estimate of the variance (notated as ‘s’) as 0.0686 mm/day which gives a correction factor (calculated as exp(2.65s^2) of 1.0126. This correction factor, a multiplier, can be applied to the bias corrected values to adjust for possible the bias due to the log10 space regression we have implemented.

To assess this bias correction, we calculated RMSE values at a catchment, climate type and countrywide level (Figure 8 and Tables 2 and A1). The log-transform bias correction improves the nationwide RMSE value by an order of magnitude (2.648 vs. 0.292) and most significantly improves catchments in Climate Types III and IV (Figure 8; 2.285 vs. 0.432 and 2.398 vs. 0.131, respectively; Table 2). Interestingly, the median RMSE value for Climate Type I and II catchments are not notably improved, however, the RMSE range for both have been reduced (red and blue boxes in Figure 8, respectively).

This analysis and the improvement of RMSE values, as well as some of performance metrics such as NSE (see scores tabulated in Table 2), using a simple log-transform based bias correction demonstrates the importance of either: 1) including smaller catchments in future products such as GRUN, or 2) performing similar bias corrections on a country, region or even catchment scale as appropriate. This is particularly important because if taken at face value the proportional contribution of relatively small tropical land areas to global discharge (e.g., Dai and Trenberth, 2002) would be underestimated without such corrections.

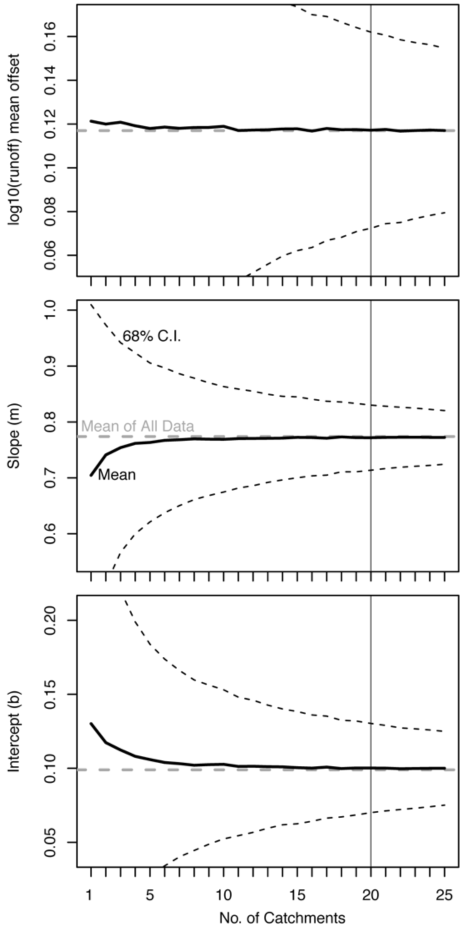

Figure 7. Bootstrap uncertainty of bias correction. Plotted versus the number of catchments is the mean and confidence interval (68%) log10(runoff) offset, the slope and intercept of the bias correction determined via random sampling at the 335 catchment level without replacement. The vertical line at 20 catchments is the uncertainty reported in the text where all three coefficients asymptote to the mean values determined using the entire dataset.

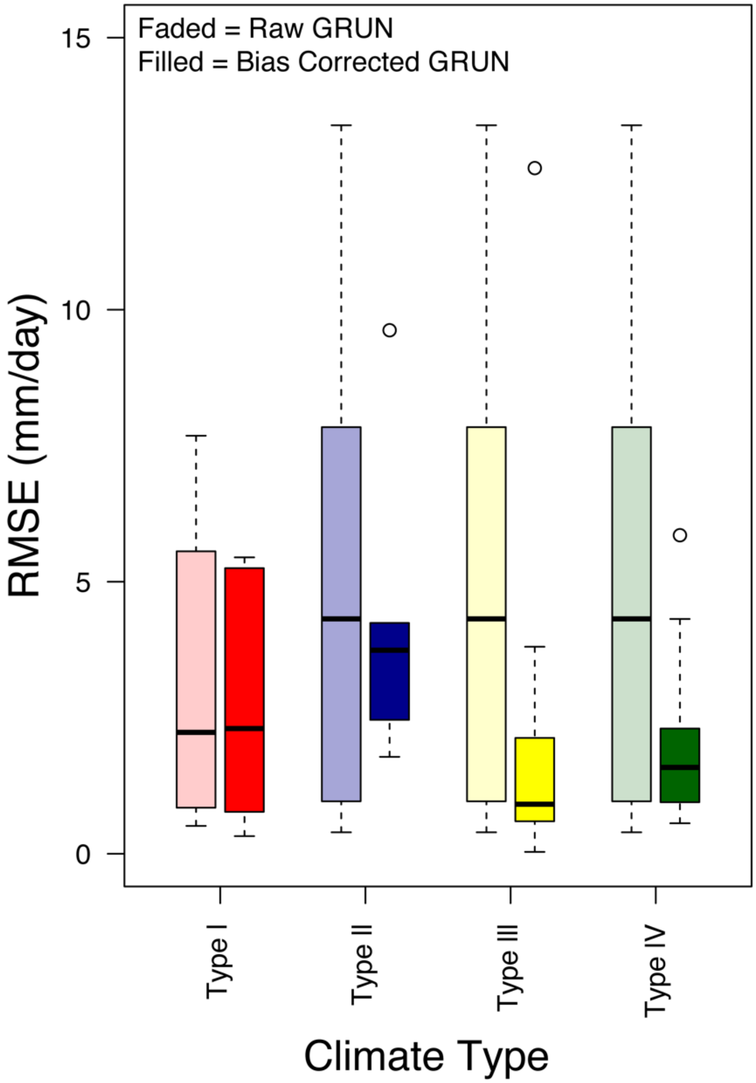

Figure 8. Box and whisker plots of the Root Mean Square Error (RMSE) for catchments grouped by climate type of 340 observed values versus raw GRUN values (light-coloured boxes) and bias-corrected GRUN values (bold-coloured boxes). For bias correction equation and country-wide results see Table 2.

4 Conclusion

Using monthly runoff observations from catchments in the Philippines with more than 10 years of data between 1946 and 2014, we a significant but weak correlation ($r^2 = 0.372$) between the GRUN-predicted runoff values and actual observations. The results indicate a somewhat skilful prediction for monthly runoff (Volumetric Efficiency = 0.363 and log(Nash-Sutcliff Efficiency) = 0.453). Looking at different hydrometeorological regimes, we demonstrated that GRUN performs best among low rainfall catchments located in climate types III and IV. There was a weak negative positive correlation between volumetric efficiency and catchment area. Further, we found that particularly for smaller catchments, maximum wet season values are grossly underpredicted by GRUN. The application of a nationwide bias correction to stretch high runoff values using log-transform runoff values greatly improved the RMSE of the predicted values. Global databases such as GRUN are applicable for aggregated stream discharge estimates and to investigate general trends in the hydrologic characteristics of a region. The recommended bias correction presented here will likely improve such estimates and analysis for the Philippines. GRUN was not intended to be used for estimating discharge for single small catchments, however it is applicable for use in regional and country scale analyses provided that proper statistical comparison of modelled versus actual gauged data are performed. We thus propose that the use of the GRUN dataset can be extended to other ungauged tropical regions with smaller catchments after at least applying a similar correction as described in this study.

Acknowledgements

We thank the Republic of the Philippines Department of Science and Technology Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) Balik Scientist program for facilitating collaboration between the authors in 2019. This project was partially supported through the DOST-PCIEERD project entitled, Catchment Susceptibility to Hydrometeorological Events supporting Tolentino and David, and a DOST-PCIEERD Balik Scientist award to Ibarra. Ibarra is funded by the UC Berkeley Miller Institute and UC President's Postdoctoral Fellowship. We acknowledge the help of Mart Geronia, Allen Marvin Yutuc, Iris Joy Gonzales, Jia Domingo and EJ Terciano Jr. for help with dataset compilation, and Tyler Kukla for discussing the statistical methods. We thank DPWH-BRS, GISM and GRDC for datasets availability. Finally, we thank Jaivime Evaristo and an anonymous referee for thorough and helpful reviews and edits, and Gionata Ghiggi for ensuring accurate representation of the GRUN product and comments on the manuscript.

Data availability

Data was compiled from the DPWH-BRS, GISM and GRDC datasets (see links in text) and is made available as a supplemental file. The supplemental file "PhilippinesRiverDischarge_Ibarra_HESS.xlsx" contains individual tabs for each catchment in the same order as Figures A1 and A2. The runoff time series start with January of the first year that data is available. Blank cells indicate no measurement. The first three rows include the major river basin, the river name and the location of the station. Further metadata including location (latitude and longitude), area and data source can be found in Table 1.

Author contributions

DEI and CPCD designed the study, DEI and PLMT carried out the dataset compilation and screening, PLMT verified catchment areas, DEI and PLMT carried out the analysis, and DEI and CPCD prepared the manuscript with contributions from PLMT.

Competing interests

The authors declare that they have no conflict of interest.

References

Abon, C. C., David, C. P. C., and Bellejera, N. E. B.: Reconstructing the Tropical Storm Ketsana flood event in Marikina River, Philippines, Hydrol. Earth Syst. Sci., 15, 1283–1289, doi:10.5194/hess-15-1283-2011, 2011.

Abon, C. C., Kneis, D., Crisologo, I., Bronstert, A., David, C. P. C., Heistermann, M.; Evaluating the potential of radar-based rainfall estimates for streamflow and flood simulations in the Philippines, Geomat. Nat. Haz. Risk., 7, 1390-1405, doi:10.1080/19475705.2015.1058862, 2016.

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., and Salamon, P.: A global streamflow reanalysis for 1980–2018. Journal of Hydrology X, 6, 100049, doi: 10.1016/j.hydroa.2019.100049, 2020.

Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals. Hydrol. Process., 22(14), 2723-2725, doi:10.1002/hyp.7072, 2008.

Dai, A., and Trenberth, K. E.: Estimates of Freshwater Discharge from Conntinents: Latitudinal and Seasonal Variations, Journal of Hydrometeorology, 3 (6), 660-687, 2002.

David, C. P., Cruz, R. V. O., Pulhin, J. M., and Uy, N. M., Freshwater Resources and Their Management, In, Philippine Climate Change Assessment Report WG2: Impacts, Vulnerabilities and Adaptation, OML Foundation, 34-54, 2017.

Davie, J.C., Falloon, P.D., Kahana, R., Dankers, R., Betts, R., Portmann, F.T., Wisser, D., Clark, D.B., Ito, A., Masaki, Y. and Nishina, K.: Comparing projections of future changes in runoff from hydrological and biome models in ISI-MIP, Earth System Dynamics, 4(2), 359-374, doi: 10.5194/esd-4-359-2013, 2013

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata, Earth Syst. Sci. Data, 10, 765–785, https://doi.org/10.5194/essd10-765-2018, 2018.

Evaristo, J., and McDonnell, J. J.: Global analysis of streamflow response to forest management: Nature, 570, 455-461, doi:10.1038/s41586-019-1306-0, 2019.

Ferguson, R. I.: River Loads Underestimated by Rating Curves, Water Resources Research, 22, 74-76, doi; 10.1029/WR022i001p00074, 1986.

Ghiggi, G., Humphrey, V., Seneviratne, S.I., Gudmundsson, L. 2019. GRUN: An observations-based global gridded runoff dataset from 1902 to 2014. Earth Sys. Sci., 11, 1655-1674, doi:10.5194/essd-11-1655-2019, 2019. Dataset analysed: https://figshare.com/articles/GRUN_Global_Runoff_Reconstruction/9228176

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst. Sci. Data, 10, 787–804, https://doi.org/10.5194/essd-10-787-2018, 2018.

Hagemann, S., Chen C., Haerter, J. O., Heinke, J., Gerten, D., and Piani, C.: Impact of a Statistical Bias Correction on the Projected Hydrological Changes Obtained from Three GCMs and Two Hydrology Models, Journal of Hydrometerology, 12 (4), 556-578, doi: 10.1175/2011JHM1336.1, 2011.

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., and Pappenberger, F.: GloFAS-ERA5 operational global river discharge reanalysis 1979-present. Earth System Science Data, doi: 10.5194/essd-2019-232, 2020

Jose, A. M., and Cruz, N. A.: Climate change impacts and responses in the Philippines: water resources, Clim Res., 12 (2-3), 77–84, 1999.

Kim, H., Watanabe, S., Chang, E. C., Yoshimura, K., Hirabayashi, J., Famiglietti, J., and Oki, T.: Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1) [Data set], Data Integration and Analysis System (DIAS), doi:10.20783/DIAS.501, 2017.

Kintanar R. L.: Climate of the Philippines, PASAGA report, 38, 1984.

Kumar, P., Masago, Y., Mishra, B. K., and fukushi, K., Evaluating future stress due to combined effects of climate change and rapid urbanization for Pasig Marikina river, Manila, Groundwater Sustain. Develop., 6, 227-34, doi:10.1016/j.gsd.2018.01.004, 2018.

Kummu, M., Guillaume, J. H. A., de Moel, H., Eisner, S., Flörke, M., Porkka, M., Siebert, S., Veldkamp, T. I. E., and Ward, P. J.: The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability, Nat. Sci. Rep., 6, 38495, https://doi.org/10.1038/srep38495, 2016.

Meybeck, M., Kummu, M, and Dürr, H. H.: Global hydrobelts and hydroregions: improved reporting scale for water-related issues, Hydrol. Earth Syst. Sci., 17, 1093–1111, doi: 10.5194/hess-17-1093-2013, 2013.

Mayor, A. G., Bautista, S., and Bellot, J.: Scale-dependent variation in runoff and sediment yield in a semiarid Mediterranean catchment, J. Hydrol., 397 (1-2), 128-135, doi:10.1016/j.jydrol.2010.11.039, 2010.

Merz, R., Parajka, J., and Bloschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resources Research, 47 (2), doi: 10.1029/2010WR009505, 2011.

Mulligan, M.: WaterWorld: a self-parameterising, physically based model for application in data-poor but problem-rich environments globally, Hydrology Research, 44 (5), 748-769, doi: 10.2166/nh.2012.217, 2013.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Nurse, L. A., McLean, R. F., Agard, J. Briguglio, L.P., Duvat-Magnan, V., Pelesikoti, N., Tompkins, E. and Webb, A.: Small islands. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Barros, V.R., et al. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, Ch. 29, 1613-1654, 2014.

Paronda, G. R. A., David, C. P. C., and Apodaca, D. C.: River flow patterns and heavy metals concentrations in Pasig River, Philippines as affected by varying seasons and astronomical tides, IOP Conf. Series: Earth and Environmental Science 344, 012049, doi: 10.1088/1755-1315/344/1/012049, 2019.

Rodríguez-Caballero, E., Cantón, Y., Lazaro, R., and Sole-Benet, A.: Cross-scale interactions between surface components and rainfall properties. Non-linearities in the hydrological and erosive behavior of semiarid catchments. Journal of Hydrology 517, 815–825, doi: 10.1016/j.jhydrol.2014.06.018, 2014.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. Theoretical and Applied Climatology 115, 15-40, doi: 10.1007/s00704-013-0860-x, 2014.

Tarasova, L., Basso, S., Zink, M. and Merz, R.: Exploring Controls on Rainfall-Runoff Events: 1. Time Series-Based Event Separation and Temporal Dynamics of Event Runoff Response in Germany, Water Resources Research, 54 (10), 7711-7732, doi:10.1029/2018WR022587, 2018.

Tolentino, P. L. M., poortinga, A., Kanamaru, H., Keesstra, S., Maroulis, J., David, C. P. C., Ritsema, C. J.: Projected Impact of Climate Change on Hydrological Regimes in the Philippines, PLoS ONE, 11(10), e0163941, doi: 10.1371/journal.pone.0163941, 2016.

Wanders, N. and Wada, Y.: Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century, Geophys. Res. Lett., 42, 10689–10695, https://doi.org/10.1002/2015GL066929, 2015.

Winsemius, H.C., Aerts, J.C., Van Beek, L.P., Bierkens, M.F., Bouwman, A., Jongman, B., Kwadijk, J.C., Ligtvoet, W., Lucas, P.L., Van Vuuren, D.P. and Ward, P.J.: Global drivers of future river flood risk. Nature Climate Change, 6(4), 381-385, doi: 10.1038/nclimate2893, 2016.

WEF: The Global Risks Report 2018, available at: http://reports. weforum.org/global-risks-2018/.

Zhang, Q., Liu, J., Yu, X. and chen, L.: Scale effects on runoff and a decomposition analysis of the main driving factors in the Haihe Basin mountainous area. Sci. Tot. Env., 690, 1089-1099, doi: 10.1016/j.scitotenv.2019.06.540, 2019

Table 1: List of stations used in this analysis including full station names, updated catchment areas, years of coverage, division and climate type.

River Name	Station Name	Latitude	Longitude	Coverage	Years of Coverage	Catchment Area (km^2)	Dataset	Climate Type
Sinalang River	Penarrubia, Abra	17.622	120.715	1984-2015	32	136.128	BRS	1
Antequera River	Sto. Rosario, Antequera, Bohol	9.493	123.890	1984-2016	33	54	BRS	4
Amparo River	Brgy. Mabini, Macrohon, So. Leyte	10.042	124.018	1985-2007	23	74	BRS	4
Hira-an River	Upper Hiraan, Rarigara, Leyte	11.258	124.672	1986-2010	27	8.93	BRS	4
Leyte River	San Joaquin, Capocan, Leyte	11.880	124.829	1985-2007	23	29.15	BRS	3
Surigao River	Surigao City	9.796	125.808	1986-2010	25	85	BRS	4
Bais River	Cabanlutan, Bais City, Negros Oriental	9.876	124.140	1989-2015	27	41	BRS	2
Lingayaon River	Lingayon, Alang-Alang, Leyte	11.192	124.863	1957-1991	35	18	BRS	4
Sapiniton River	Libton, San MIguel, Leyte	11.188	124.795	1984-2010	27	277.3	BRS	4
Laoag River	Poblacion, Laoag City, Ilocos Norte	18.203	120.590	1984-2016	33	1355	BRS; updated from Tolentino et al. (2016)	1
Pared River	Baybayog, Alcala, Cagayan	17.682	121.270	1983-1996	14	966	BRS; used in Tolentino et al. (2016)	1
Ganano River	Ipil, Echague, Isabela	16.812	121.211	1986-2001	16	977	BRS; used in Tolentino et al. (2016)	3
Magat River	Baretbet, Bagabag, Nueva Vizcaya	16.992	121.073	1986-2002	17	2199	BRS; used in Tolentino et al. (2016)	3
Camiling River	Poblacion, Mayantoc, Tarlac	15.018	120.503	1985-2017	33	288	BRS; updated from Tolentino et al. (2016)	1
Gumain River	Sta. Cruz, Lubao, Pampanga	14.960	120.441	1985-2001	17	370	BRS; used in Tolentino et al. (2016)	1
Rio Chico River	Sto. Rosario, Zaragosa, Nueva Ecija	15.658	120.088	1985-2006	22	1177	BRS; used in Tolentino et al. (2016)	1
San Juan River	Porac, Calamba, Laguna	14.498	121.267	1986-1999	14	165	BRS; used in Tolentino et al. (2016)	4
Pangalaan River	Pangalaan, Pinamalayan, Oriental Mindoro	13.275	121.260	1989-1999	11	32	BRS; used in Tolentino et al. (2016)	3
Das-ay River	Sto. Nino II, Hinungangan, Leyte	10.385	125.202	1987-2007	21	59	BRS; used in Tolentino et al. (2016)	2
Tukuran River	Tinotongan, Tukuran, Zamboanga del Sur	7.627	123.030	1986-2009	24	147	BRS; used in Tolentino et al. (2016)	3
Hijo River	Apokan, Tagum, Davao del Norte	7.812	125.211	1986-2016	31	634	BRS; updated from Tolentino et al. (2016)	4
Cagayan de Oro River	Cabula, Cagayan de Oro City, Misamis Oriental	8.316	124.811	1991-2004	14	1079	BRS; used in Tolentino et al. (2016)	4
Davao River	Tigatto, Davao City	7.329	125.634	1984-1999	16	1683	BRS; used in Tolentino et al. (2016)	4
Allah River	Impao, Isulan, Sultan Kudarat	6.568	124.085	1980-1994	15	1231	BRS; used in Tolentino et al. (2016)	3
Agusan Canyon River	Camp Philips, Manolo Fortich, Bukidnon	8.296	124.450	1986-2004	19	48	BRS; updated from Tolentino et al. (2016)	3
Wawa River	Wawa, Bayugan, Agusan del Sur	8.261	125.501	1981-2010	30	396	BRS; updated from Tolentino et al. (2016)	4
Buayan River	Malandag, Malungon, South Cotabato	6.317	125.749	1986-2004	19	207	BRS; used in Tolentino et al. (2016)	4
Gasgas River	Manalpac, Solsona	18.080	120.830	1978-1988	11	73	GISM	1
Jalaur River	Calyan, Pototan, Iloilo	10.930	122.670	1976-1988	13	1499	GISM and GRDC	3
Padsan River	Bangay	18.080	120.700	1946-1979	34	534	GRDC	1
Pampanga River	San Agustin	15.170	120.780	1946-1977	32	6487	GRDC	1
Sipocot River	Sabang	13.810	122.990	1946-1970	25	447	GRDC	2
Mambusao River	Tumalalud	11.260	122.570	1950-1978	29	307	GRDC	3
Padada River	Lapulabao	6.660	125.280	1949-1978	30	821	GRDC	4
Aloran River	Juan Bacay, Aloran, Misamis Occ.	8.420	123.820	1978-2003	26	30	GRDC + BRS	3
Cabacanan River	Baduang, Pagudpud	18.580	120.800	1979-2017	39	60	GRDC + BRS	1
Maragayap River	Sta. Rita, Bacnotan, La Union	16.750	120.374	2004-2017	14	40	BRS	1
Abacan River	San Juan, Mexico, Pampanga	15.118	120.703	2004-2017	14	217	BRS	1
Hibayog River	La Victoria, Carmen, Bohol	9.876	124.141	2004-2017	14	41	BRS	4
Manaba River	Calma, Garcia-Hernandez, Bohol	9.631	124.131	2001-2016	16	98	BRS	4
Gabayan River	Canawa, Candijay, Bohol	9.848	124.450	2001-2017	17	48.5	BRS	4
Bangkerohan River	Brgy. Tagaytay, Bato, Leyte	10.342	124.834	1984-1990; 2000-2009	17	168	BRS	4
Borongan River	Brgy. San Mateo, Borongan City	11.628	125.403	1990-2008	19	111	BRS	2
Loom River	Brgy. Calico-an, Borongan City	10.594	125.404	1986-2004	19	42	BRS	2
Pagbanganan River	Brgy. Makinhas, Baybay City	10.637	124.865	1984-2008	25	128	BRS	4
Rizal River	Brgy. Rizal, Babatngon, Leyte	11.389	124.908	1990-2008	18	15	BRS	4
Tenani River	Brgy. Tenani, Paranas (Wright), Samar	11.806	125.127	1985-2001	17	394	BRS	2
Disakan River	Disakan, Manukan, Zamboanga del Norte	8.480	123.048	1985-1991; 1997-2000	11	109	BRS	3
Kabasalan River	Banker, Kabasalan, Sibugay, Province	7.831	122.778	2002-2011	10	143	BRS	3
Sindangan River	Dicoyong, Sindangan, Zamboanga del Norte	8.217	123.057	1990-2003	14	590.5	BRS	3
Alubijid River	Alubijid, Misamis Oriental	8.570	124.476	1991-2009	19	124	BRS	3
Kipaliko River	Tiburcia, Kapalong, Davao del Norte	7.602	125.681	2004-2016	13	147	BRS	4
Banaue River	Poblacion, Banaue, Ifugao	16.915	121.061	1987-1995; 2005-2010	15	15	BRS	3
Aciga River	Santiago, Agusan del Norte	9.269	125.570	2002-2015	14	80	BRS	4
Agusan River	Sta. Josefa, Agusan del Sur	7.993	126.036	1982; 1984-1987; 1989-2010	27	1633	BRS	4

Table 2: Results of statistical agreement between GRUN aggregated by Climate Type and for the entire dataset (see Table A1 for individual catchments)

	Pearsons Coeff (r^2)*	**Volumetric Efficiency (VE)**	**Nash-Sutcliff Efficiency (NSE)**	**Nash-Sutcliff Efficiency (NSE-log10)**	**Root Mean Square Error**	**Root Mean Square Error Bias Corrected ****	**Volumetric Efficiency (VE) Bias Corrected ****	**Nash-Sutcliff Efficiency (NSE) Bias Corrected ****	**Nash-Sutcliff Efficiency (NSE-log10) Bias Corrected ****
Entire Dataset	0.372	0.363	0.091	0.453	2.648	0.292	0.323	0.182	0.385
Entire Dataset log10(runoff)	0.546	0.733	n/a	n/a	n/a	n/a	1.067	n/a	n/a
Climate Type 1 (n=12)	0.211	0.252	0.062	0.538	2.476	0.298	0.168	0.111	0.432
Climate Type 2 (n=6)	0.409	0.354	0.05	0.49	4.554	0.544	0.349	0.188	0.457
Climate Type 3 (n=15)	0.471	0.404	0.026	0.23	2.285	0.432	0.345	0.011	0.188
Climate Type 4 (n=22)	0.535	0.403	0.159	0.414	2.398	0.131	0.377	0.323	0.36

Notes

* For regressions forced through intercept of 0

** Two-step bias correction procedure where first mean offset is added to the predicted GRUN values and then a log-transform stretch correction is applied (see text for details)