A Statistical Interpolation of Satellite Data with Rain Gauge Data over Papua New Guinea

ZHI-WENG CHUA, a, b YURIY KULESHOV, a, b ANDREW B. WATKINS, a SUELYNN CHOY, b AND CHAYN SUN, b

ABSTRACT: Satellites provide a useful way of estimating rainfall where the availability of in situ data is low but their indirect nature of estimation means there can be substantial biases. Consequently, the assimilation of in situ data is an important step in improving the accuracy of the satellite rainfall analysis. The effectiveness of this step varies with gauge density, and this study investigated the effectiveness of statistical interpolation (SI), also known as optimal interpolation (OI), on a monthly time scale when gauge density is extremely low using Papua New Guinea (PNG) as a study region. The topography of the region presented an additional challenge to the algorithm. An open-source implementation of SI was developed on Python 3 and confirmed to be consistent with an existing implementation, addressing a lack of open-source implementation for this classical algorithm. The effectiveness of the analysis produced by this algorithm was then compared to the pure satellite analysis over PNG from 2001 to 2014. When performance over the entire study domain was considered, the improvement from using SI was close to imperceptible because of the small number of stations available for assimilation and the small radius of influence of each station (imposed by the topography present in the domain). However, there was still value in using OI as performance around each of the stations was noticeably improved, with the error consistently being reduced along with a general increase in the correlation metric. Furthermore, in an operational context, the use of OI provides an important function of ensuring consistency between in situ data and the gridded analysis.

SIGNIFICANCE STATEMENT: The blending of satellite and gauge rainfall data through a process known as statistical interpolation (SI) is known to be capable of producing a more accurate dataset that facilitates better estimation of rainfall. However, the performance of this algorithm over a domain such as Papua New Guinea, where gauge density is extremely low, is not often explored. This study reveals that, although an improvement over the entire Papua New Guinea domain was slight, the algorithm is still valuable as there was a consistent improvement around the stations. Additionally, an adaptable and open-source version of the algorithm is provided, allowing users to blend their own satellite and gauge data and create better geospatial datasets for their own purposes.

KEYWORDS: Gauges; Satellite observations; Interpolation schemes; Optimization

1. Introduction

Accurate estimation of rainfall is critical to the knowledge of water availability, which is essential for the proper functioning of societies (Ponting 2007). Rain gauges offer a direct estimation through the collection and measurement of precipitation. Although subject to instrumental biases (Michaelides et al. 2009), rain gauges typically provide the most accurate estimate (Beck et al. 2019a), though radar and satellite precipitation estimates can be more valuable over larger areas (Wood et al. 2000) and for hydrological modeling (Gilewski and Nawalany 2018). One of the biggest limitations of rain gauges is the barrier to their installation, which can be both physical and financial, resulting in limited coverage over many parts of the world, including over oceans (Kidd et al. 2017). For a gauge-based estimate to be formed away from gauges, interpolation must be performed, but since rainfall is a variable that can possess high spatiotemporal variation, estimates away from rain gauges may be a poor representation of the actual rainfall over the area (Habib et al. 2001).

One of the countries where rainfall estimation is severely impacted by the number of operational gauges is Papua New Guinea (PNG); there are only seven stations for a land area of around 463 000 km² (Bhardwaj et al. 2021). The distribution of these stations exacerbates the limitations of the network as the majority of these stations are along the coast, leading to much of the mainland’s interior being unobserved. This region includes a major topographical feature in the form of the New Guinea Highlands, which complicates rainfall estimation due to the increased spatiotemporal variation induced by topography (Amjad et al. 2020). Spatial interpolation methods can be employed to produce a gridded rainfall analysis but performance can be expected to be greatly impacted by the very low gauge densities (Hofstra et al. 2010). For example, in an examination of gauge-based, reanalysis, and satellite-based rainfall datasets over PNG, Smith et al. (2013) noted that gauge analyses were particularly limited by their coarse horizontal resolution and struggled to resolve finer-scale spatial features like topographical effects.
The use of alternative data sources to bolster the estimation from rain gauges would be highly valuable over PNG. Global validations of rainfall estimates from model reanalyses and satellites suggest satellite datasets are the preferred alternative data source over the region (Beck et al. 2019a; Tang et al. 2020). In a validation against PNG station data, Global Satellite Mapping of Precipitation (GSMaP) satellite estimates had similar performance to ERA5 reanalysis estimates, though notably, over the one station with significant elevation, GSMaP performed better (Chua et al. 2020). Considering this, we chose to investigate the production of a satellite–gauge blended dataset over PNG.

To our knowledge, there is no satellite–gauge rainfall analysis specifically developed for PNG. Blended satellite–rainfall analyses such as Integrated Multi-satellite Retrievals for Global Precipitation Mission (IMERG), its predecessor Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA), and Climate Prediction Center (CPC) morphing method blended version (CMORPH BLD) have been developed over a quasi-global domain (Huffman et al. 2020, 2007; Xie and Xiong 2011), while datasets such as the Multi-Source Weighted Ensemble Precipitation (MSWEP) and CPC Merged Analysis of Precipitation (CMAP) also include model reanalysis data, extending coverage to the polar latitudes (Beck et al. 2019b; Xie et al. 2007). However, these existing datasets have mediocre performance over PNG (Chua et al. 2020; Wild et al. 2021), in part due to a lack of stations to incorporate. Wild et al. (2021) found the performance of contemporary satellite datasets including GSMaP and IMERG was the worst in PNG out of six countries in the southwest Pacific.

One way of improving the performance of existing analyses is to perform a postprocessed correction, with a variety of algorithms available including distance-based methods (Adhikary et al. 2016), the use of corrective ratios (Lin and Wang 2011), and data-driven approaches (Zhang et al. 2021). In tandem with this, many contemporary satellite datasets already possess a form of calibration or correction to station data in their generation process (Huffman et al. 2020; Mega et al. 2019). However, these corrections typically only utilize a subset of available stations. For example, GSMaP Gauge Near Real Time (GSMaP-GNRT, hereafter referred to as GSMaP) is calibrated to CPC Gauge Unified, a gauge analysis which only utilizes around three stations over PNG (Becker et al. 2013), a subset of the six stations utilized in this study. In this study, we aim to employ the statistical interpolation (SI) algorithm to develop a satellite–gauge rainfall analysis for PNG that improves upon the performance of GSMaP.

Also known as optimal interpolation (OI), SI is based on assimilating station observations onto a background field using weights that minimize the error of the resultant analysis.

In our previous study, we demonstrated the viability of using SI with a background field formed from monthly satellite precipitation estimates (Chua et al. 2022). Over Australia, we produced a satellite–gauge rainfall analysis for monthly rainfall that matched the performance of the Bureau of Meteorology’s (BOM) operational gauge analysis, the Australian Gridded Climate Dataset (AGCD) rainfall dataset, in addition to outperforming it over gauge-sparse regions. It also performed similarly to other top-performing satellite–gauge blending algorithms we identified. Importantly, one of the advantages of SI we identified was that the improvement from using satellite data corrected to gauge analysis through a preliminary step was relatively slight compared to other blending algorithms.

Removing the requirement of having a gauge analysis for correction is extremely valuable for PNG due to the low performance of gauge analyses over the region. The adaptation of this technique for PNG addresses multiple novel and valuable points:

1) The creation of a monthly rainfall analysis over PNG. Given its extreme gauge paucity, a performant rainfall analysis provides information over many parts of the country which would not be possible by solely using in situ station data. A gridded dataset is valuable for both operational and research purposes. For example:

(i) An important operational use is its potential to be used as input into drought early warning systems (DEWS); PNG is listed as the ninth most at-risk country in the world to natural hazards (Aleksandrova et al. 2021). One example of a potential DEWS that utilizes a gridded rainfall dataset is described in Bhardwaj et al. (2021). Dataset accuracy and resolution were mentioned as a system limitation with a more accurate rainfall analysis enabling better performance and trust in the DEWS.

(ii) A more accurate rainfall analysis facilitates better climate and environmental analysis. This would allow PNG National Weather Service (NWS) climatologists to better understand the spatiotemporal variability of PNG rainfall, as well as the climate drivers influencing their region. Bhardwaj et al. (2021) explored the impacts of El Niño–Southern Oscillation and the Indian Ocean dipole on PNG rainfall through the creation and analysis of rainfall decile maps based on MSWEP data. Gridded rainfall datasets have also seen use in other recent environmental and climate monitoring studies, including for investigating the temporal relationship of precipitation with vegetation growth (Ghaderpour et al. 2023) and for identifying breakpoints in the historical precipitation record across the globe (Kazemzadeh et al. 2022). Increasing the accuracy of the rainfall analyses that underpin climatological research can lead to new insights and bolster confidence in known teleconnections.

2) Evaluating the effectiveness of SI over a gauge-sparse region. Chua et al. (2022) demonstrated SI was effective over Australia, but in terms of the overall domain, Australia has a much denser gauge network than PNG (1 per 1300 km² compared to 1 per 66 100 km²). A degradation of analysis performance was observed over the gauge-sparse interior of Australia (though the SI dataset still performed favorably compared to the pure satellite or gauge analyses), and it is expected that performance will also be reduced over PNG due to its comparatively lower
gauge density. It is an important research question to find out to what extent degradation of performance will occur and whether the algorithm still has merit when only a small number of stations are available for assimilation. This knowledge is vital for understanding the applicability of the technique to other regions around the world, especially since the utility of a satellite–gauge dataset is greater for gauge-sparse regions.

3) Although SI is a well-regarded classical data assimilation technique, open-source Python implementation is extremely limited. PyDA (Ahmed et al. 2020) is the closest example we could find, being a Python module developed for data assimilation. However, PyDA does not seem to be developed for 2D geospatial data assimilation, with the examples provided being related to 1D interpolation. The code used in this study will be made open-access and consequently will be a valuable contribution to filling this void. Although developed for rainfall assimilation, it can easily be adapted for 2D assimilation of other variables. The code we previously used (Chua et al. 2022) was FORTRAN based, a less accessible language than Python, and was not available for open-source access. Its legacy development also meant further development was difficult, including modification for use outside of an Australian domain. This motivated the development of the algorithm in Python.

If the blending process over PNG is successful, this will open its usage to many other regions of the Pacific which share the characteristics of having a limited gauge network.

2. Materials and methods

a. Study domain

Papua New Guinea is a country in the southwestern Pacific which is comprised of the eastern half of the New Guinea island (commonly referred to as its mainland) along with around 700 offshore islands (Smith et al. 2013). The mainland possesses a significant amount of topography, the main form being a mountain range known as the New Guinea Highlands that traverses through its center (Smith 1985). The New Guinea Highlands peak at 4510 m and are high enough to receive snowfall (Smith 1985). A map of the study domain along with topography derived from NOAA’s ETOPO1 dataset is shown in Fig. 1. The ETOPO1 dataset provides information on topographical relief as derived from global and regional surveyed data and satellite altimetry (NOAA 2016).

Being close to the equator, most of PNG is primarily classed as equatorial climate according to the Köppen–Geiger climate classification, indicating the occurrence of substantial rainfall throughout the year (Beck et al. 2018). However, southern parts of the mainland are classified as tropical savanna signifying the presence of a marked wet–dry season (Beck et al. 2018). The climate of PNG is heavily influenced by the northwest monsoon from December to April and the southeast monsoon from May to October (Pereira et al. 2019), resulting in wetter and drier periods over the year for most parts of the country. For example, the capital Port Moresby is categorized as having a “wet” season between October and April and a “dry” season for the remainder of the year (Smith et al. 2013).

b. SI algorithm

Statistical interpolation (SI) is a method of assimilating in situ data onto a gridded background field (Reynolds and Smith 1994). It was independently developed by Kolmogorov (in 1941) and Wiener (in 1949) and with the advent of increased computing power, it was adopted by meteorological agencies from the mid-1970s onward (Foster 1961). This includes for the purposes of creating a rainfall analysis from gauges (Evans et al. 2020), as well as for blending gauge data with satellite estimates (Wu and Xie 2016). The analysis produced by SI is a weighted average of the in situ data (typically within a search radius) and the background field with the
weights being calculated so that the error variance of the resultant analysis is a minimum, with respect to both the in situ data and the background field. SI relies on the assumption that corrections to the background field depend linearly on the background–observation residuals, that the background and observation errors are unbiased and uncorrelated, and that rainfall errors are nonstationary and anisotropic (Heo et al. 2018). The ramifications of these assumptions are discussed in section 4a.

Given the development of the algorithm in Python is one of the key outcomes of this study, an outline of the mathematics is provided in this section. With the assumptions of no correlation between the background and observation errors, the SI algorithm can be written as Eq. (1):

$$A_i = B G_i + W_i \times (F_O - F_B)$$  \hspace{1cm} (1)

with

$$W_i = B_i \times (B + O)^{-1}$$  \hspace{1cm} (2)

where $A_i$ is the analysis value, $B G_i$ is the background value, $W_i$ is the vector of weights, $F_O$ is the vector of observations, $F_B$ is the vector of background values at the observation coordinates, $B$ is the background error covariance matrix, $O$ is the observation error covariance matrix, and $B_i$ is the vector of background error covariance between the analysis grid point and the observations. The subscript $i$ refers to values calculated for a particular grid point.

The terms $B$ and $O$ can be calculated a priori and are based on the observation locations. For the interpolation to be truly optimal, the “truth” is required to calculate the error variances required for $B$ and $O$. However, since the truth is unavailable, $B$ and $O$ must be estimated, which is why the process can never be “optimal.” There are several methods of estimating $B$ and $O$; Daley’s method (Daley 1991) was adopted in this study as it is a well-known implementation of the algorithm, as well as being the method used by BOM to create AGCD. This method involved breaking down an error covariance matrix into a product of its error variance component and its correlation component.

The error variance component was estimated from the error variance between the observation and the background. By assuming error homogeneity and no correlation between the background and observation errors, the sum of the observation error variance ($E_O^2$) and the background error variance ($E_B^2$), which can also be referred to as the total error variance, was represented as the error variance between the observation ($O$) and the background ($B$) computed for every station $k$ (with $K$ representing the total number of stations). This is denoted by Eq. (3):

$$\sum_{k=1}^{K} \frac{(O_k - B_k)^2}{R} = \frac{1}{K} \sum_{k=1}^{K} (O_k - T_k)^2 + \frac{1}{R} \sum_{k=1}^{K} (B_k - T_k)^2$$

$$= E_O^2 + E_B^2,$$  \hspace{1cm} (3)

where $T$ is the (unknown) truth.

The background error variance can be written as a proportion ($R_c$) of the total error variance, as denoted by Eq. (4):

$$E_B^2 = R_c \times (E_O^2 + E_B^2).$$  \hspace{1cm} (4)

The $R_c$ can be estimated using $R(r)$, a plot of the correlation of observation–background errors ($R$) with distance ($r$). By assuming that the observation error was not correlated with distance, its only contribution to $R(r)$ was at the $R$ intercept of $R(r)$ and it is this $R$ intercept that provides an estimate of $R_c$. Note for example, the smaller this intercept is, the lower the correlation of perfectly collocated stations, which implies a larger observation error variance (since the background value is the same at collocated points). The $R_c$ is thereby estimated by finding the limit of $R(r)$ as $r$ approaches zero. In this study, observation–background error pairs from Australian stations to the satellite field values were used as there were not sufficient pairs over the PNG domain. The pairs were categorized by month to provide monthly values of $R_c$. Several ways of determining $R(r)$ from these distributions were trialed:

1) exponential fitting,
2) polynomial fitting,
3) nonparametric fitting using support vector regression (SVR).

The SVR fit was identified to be the most suitable fit through visual inspection in conjunction with chi-square tests. Based on these distributions, $R_c$ was calculated and ranged from 0.58 to 0.84. Equations (3) and (4) were then solved to obtain estimates for $E_O^2$ and $E_B^2$.

The correlation component for the background error was modeled through a correlation model proposed by Thiebaux (Daley 1991) where the correlation function, $R(r)$, is based on $r$, the distance between the stations (or between the analysis point and stations for $B$) and $L$, an empirical length parameter according to Eq. (5):

$$R = \left(1 + \frac{r}{L}\right) \times e^{-r/L}.$$  \hspace{1cm} (5)

The term $L$ was selected based on the knowledge that the correlation between stations falls off sharply. Previous studies suggest correlation decreases to 0.4 for distances around 100–300 km (Svoboda et al. 2015; Şen and Habib 2001). This is in line with the correlation functions for the observation–background error calculated earlier. Following this, a value of $L$ was selected that resulted in a correlation of 0.4 at 100 km. However, this value was problematic because it occasionally resulted in scenarios where large negative differences between station and satellite values (i.e., when the station values were much smaller than the satellite values) were transferred too strongly away from the station resulting in unrealistic-looking broad regions of zero rainfall. For example, one of the areas most prone to these artifacts was Morobe Province where flat terrain morphs into the Highlands over distances on the order of tens of kilometers. Consequently, $L$ was manually reduced until these artifacts disappeared, with a value of
This relationship is based on the Thiebaux model using an \( L \) value of 25.

25 being chosen in the end. The correlations produced by this model are shown in Fig. 2.

Based on this model, the correlations at distances of 10, 50, and 100 km were roughly 0.94, 0.41, and 0.09, respectively. These lower correlations are in line with correlations found for rainfall values on shorter time scales (Habib et al. 2009). This is reasonable as the problematic regions for this study were close to significant topography, where the increased spatiotemporal variability induced by the topography dictates that the correlation length scale for these stations must be shorter than the typical length scale.

As conventional use of OI assumes that the background field and the observations are unbiased (Daley 1991), it was also reasonable to assume that the observation errors have no correlation with each other and so an identity matrix was used for the correlation matrix \( O \).

All the parameters required for calculating \( A \) have now been explained, and following Eq. (1), the analysis value for each grid point was calculated. Once the analysis was produced, all values less than 0.1 mm were set to 0 to improve rain/no-rain consistency.

A simple superobservation routine was included as highly correlated stations could potentially lead to the error covariance matrices being singular, preventing Eq. (2) from being solved. A maximum number of stations that were considered for each grid point along with a correlation threshold were selected (20 and 0.95, respectively). If stations had correlations to each other that were greater than this threshold, they were considered to be highly correlated and were combined into a superobservation. The values of the stations were combined into a weighted average with the weights being based on each station’s correlation as a fraction of the sum of the correlations. The superobservation coordinates were based on an arithmetic average of the component stations. Using a reduced set of stations containing the superobservation, the algorithm could then proceed as normal.

Note that the implementation of this routine did not affect the analysis in this study as the number of stations available for use in PNG was less than the maximum number of stations selected and were also geographically distant to each other (meaning their correlations to each other were less than the correlation threshold).

c. Datasets

The GSMaP dataset and rain gauge stations from PNG NWS were used as the SI inputs in this study. GMSaP was selected along with IMERG, it is one of the best-performing satellite-based rainfall datasets globally as well as in smaller regional studies (e.g., Tang et al. 2020; Shi et al. 2020; Wang and Yong 2020). GMSaP is also provided as part of the World Meteorological Organization (WMO) Space-Based Weather and Climate Extremes Monitoring (SWCEM) (Kuleshov et al. 2019), guaranteeing its provision in the future as a reliable data feed for operational implementation. Soil Moisture to Rain (SM2R) and ERA5 were also used as reference datasets in the triple collocation analysis. For the analysis over Australia, monthly data from all datasets were available across the full study period, i.e., from 2001 to 2020. For the analysis relying on in situ data over PNG, a period from 2001 to 2014 is used to account for the PNG rain gauge data having gaps after 2014, while a period from 2007 to 2015 is used for the TCA due to the SM2R rain data beginning in 2007. Details on these datasets are described in Table 1.

**Dataset Biases**

The performance of all datasets can be expected to suffer over gauge-sparse areas. Gauge density is the largest control on the accuracy of gauge analyses (Hofstra et al. 2010). Satellite datasets rely on calibration to gauges, and while ERA5 does not explicitly ingest in situ rainfall values, in situ data are used for other meteorological variables (such as humidity, pressure, and temperature) which affects the rainfall values modeled (Hersbach et al. 2020). Reduced observations have a significant impact on the quality of model reanalyses (Bosilovich et al. 2008).

The latest generation of satellite datasets generated from the Global Precipitation Mission (GPM) satellite constellation has demonstrated their superior performance over reanalysis datasets in nonpolar areas where gauge and radar coverage are lacking (Tang et al. 2020; Xu et al. 2022).

Mountainous terrain leads to increased biases for all datasets, in part due to increasing the spatiotemporal heterogeneity of rainfall, but also from the estimation biases unique to each dataset (Amjad et al. 2020; Saddique et al. 2022). For gauges, the main issue is increased wind speeds causing increased undercatch, with underestimations exceeding 20% for unshielded rain gauges (Pollock et al. 2018). Reanalyses are affected by modeling complexity increases due to topographical effects such as lapse rate changes and mesoscale circulations (Amjad et al. 2020). Satellite retrieval algorithms have difficulty detecting orographic rainfall that commonly is associated with low warm clouds (Dinku et al. 2007). Biases due to cold surfaces (Stampoulis and Anagnostou 2012) are unlikely to be too problematic over PNG given snowfall is confined to the highest peaks. However, biases from topography
Global Satellite Mapping of
Precipitation (GSMaP) and
Gauge Near Real Time
Version 6, Japan
Aerospace Exploration
Agency (JAXA) (Mega
et al. 2019)

Rain gauge data from PNG
NWS

Rain gauge data from BOM

Fifth-generation European
Centre for Medium-Range
Weather Forecasts
(ECMWF) reanalysis
(ERAS), ECMWF
(Hersbach et al. 2020)

Soil Moisture to Rain (SM2R)
(Brocca et al. 2019)

<table>
<thead>
<tr>
<th>Dataset name and provider</th>
<th>Description</th>
<th>Resolution/coverage/length of record</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Satellite Mapping of Precipitation (GSMaP)</td>
<td>Primarily based on microwave estimates from the Global Precipitation Mission (GPM) constellation. Microwave brightness temperatures from the surface are empirically linked to rain rates. Cloud motion vectors formed using infrared (IR) data along with the application of a Kalman filter are used to improve spatiotemporal coverage. A correction based on the values of CPC Gauge Unified over the past 30 days is performed as a final step.</td>
<td>0.1° × 0.1° global from 60°S to 60°N, hourly. Monthly record from 2001 to 2020 used over Australia, from 2007 to 2015 over PNG in this study.</td>
</tr>
<tr>
<td>Gauge Near Real Time Version 6, Japan Aerospace Exploration Agency (JAXA) (Mega et al. 2019)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain gauge data from PNG NWS</td>
<td>Stations maintained by the PNG NWS.</td>
<td></td>
</tr>
<tr>
<td>Rain gauge data from BOM</td>
<td>Stations maintained by the BOM.</td>
<td></td>
</tr>
<tr>
<td>Fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERAS), ECMWF (Hersbach et al. 2020)</td>
<td>Model reanalysis formed from 4D-Var assimilation of observations into the Integrated Forecast System (IFS) weather model. Rain gauge data are not assimilated.</td>
<td>0.1° × 0.1° global, hourly. Monthly record from 2007 to 2015 used over PNG in this study.</td>
</tr>
<tr>
<td>Soil Moisture to Rain (SM2R) (Brocca et al. 2019)</td>
<td>Accumulated rainfall rate estimates derived from inputting satellite-estimated soil moisture values into an inversion of the soil water balance equation.</td>
<td>0.1° × 0.1° global, daily. Monthly record from 2007 to 2015 used over PNG in this study for TCA.</td>
</tr>
</tbody>
</table>

are highly relevant to PNG given it is highly mountainous along the central spine of its mainland in addition to significant topography also being present on some of the smaller islands.

The presence of inland water bodies can also cause issues with satellite retrievals. Background surface emissivity over water is low, meaning emission information from hydrometeors can be utilized by satellite retrieval algorithms (Prigent 2010). Over land, the higher background surface emissivity complicates the detection of hydrometeor emission and thus, the scattering-induced reduction in brightness temperatures (which is assumed to be from hydrometeors) is used instead (Prigent 2010). Inland water bodies can lead to confusion over which algorithm should be used with overestimations in the past 30 days is performed as a final step.

Three validation routines are presented in section 3, each with a different function. The function of each validation is explained in its respective section with general points relating to the methodology of the validations presented here.

When a gridded dataset was compared to a in situ station value, the gridded dataset was bilinearly interpolated to the coordinates of a station. A spatial representation error exists since a gridded average was being compared to a point value. Typically, gridded datasets underrepresent high-end variability and overrepresent the number of rain days. One method that has been used to improve spatial consistency is forming an in situ gridded estimate by averaging the number of stations within a certain radius. However, the small number of stations available in this study made this unfeasible as at most, the grid cells would only contain one station. Instead, we acknowledge that this spatial representation error would have been present to a similar degree across all the gridded datasets, making comparisons between the datasets reasonable.

When gridded datasets were compared to each other, a land–sea mask (derived from the Python Basemap module) was applied so that only land grid cells were compared.

TRIPLE COLLOCATION ANALYSIS

The methodology of triple collocation analysis (TCA) is presented in greater detail given it is a lesser-known validation technique. TCA allows the ranking of three datasets in the absence of a known truth. This is particularly valuable for this study as the commonly used forms of truth—gauge data and radar data (Sun et al. 2018)—are extremely limited in coverage over the study domain. Furthermore, although gridded datasets that cover the whole domain exist, the uncertainty of datasets over the domain is very large (Smith et al. 2013; Wild et al. 2021) due to factors such as the aforementioned sparse observational network, the high spatiotemporal heterogeneity in rainfall brought about by the topography and the relatively high amounts of rainfall received climatologically. This uncertainty means reliance on a single dataset as truth is problematic as it is likely to contain significant bias that confounds the validation.
TCA provides a way to alleviate both of these factors and has proven itself to be a robust form of validation for monthly rainfall (e.g., Massari et al. 2017), including over PNG (Wild et al. 2021). The methodology will be briefly explained; for further details, readers are referred to Gruber et al. (2016).

TCA produces validation metrics of error variances ($\sigma$), as well as correlations (\(\rho\)), as defined by Eqs. (6) and (7):

\[
\sigma_1 = \sqrt{Q_{11} - \frac{Q_{13}Q_{13}}{Q_{23}}}
\]

\[
\sigma_2 = \sqrt{Q_{22} - \frac{Q_{13}Q_{13}}{Q_{23}}}
\]

\[
\sigma_3 = \sqrt{Q_{33} - \frac{Q_{13}Q_{13}}{Q_{23}}},
\]

\[
\rho_{x,X_1} = \frac{Q_{13}Q_{23}}{\sqrt{Q_{11}Q_{23}}}
\]

\[
\rho_{x,X_1} = \text{sign}(Q_{13}Q_{23}) \frac{Q_{13}Q_{23}}{\sqrt{Q_{11}Q_{23}}}
\]

\[
\rho_{t,X_1} = \text{sign}(Q_{13}Q_{23}) \frac{Q_{13}Q_{23}}{\sqrt{Q_{12}Q_{33}}},
\]

where $Q$ represents the covariance between a pair of datasets; the subscripts 1, 2, and 3 refer to each of the three datasets; and $t$ refers to the unknown truth dataset. There is an ambiguity of sign in Eq. (7) but the assumption of a positive correlation with truth is considered safe (Massari et al. 2017). The error variances refer to errors of a random nature, rather than a systematic bias.

Successful application of TCA relies on three key assumptions TCA: 1) linearity between the datasets and the truth, 2) stationarity of the truth and its errors, and 3) independence in the errors of the datasets. To reduce nonlinearity between the datasets due to different climatologies, the climatology-removed time series are used (Gruber et al. 2016). The appendix reveals the degree to which these assumptions are satisfied and hence, the appropriateness of TCA in this study though it should be noted that violations of these assumptions also affect the robustness of traditional validation metrics such as the root-mean-square error (RMSE) and Pearson’s correlation (Gruber et al. 2016).

3. Results

a. Validation of the implementation of the algorithm

To confirm the algorithm was implemented correctly, three checks were performed. First, the developed Python version (hereafter referred to as SI-P) was compared to the existing FORTRAN version (hereafter referred to as SI-F) that was previously utilized (Chua et al. 2022). As mentioned earlier, SI-F was limited to generation over Australia, and so this comparison had to be performed over Australia.

The difference between the two datasets was computed for each year for all land grid cells in the Australian domain. This was completed over the full period where both datasets were available (2001–20). The median value of this difference was calculated to be 5.03 mm month$^{-1}$.

A relatively slight discrepancy between the two analyses was expected with five reasons identified:

1) SI-F includes a cross-validation routine where station observations which are too different from an intermediate analysis created from their exclusion are excluded from the final analysis.
2) SI-F creates the analysis by performing SI on sectors before merging the sectors into a final analysis.
3) Values used for empirical parameters (e.g., the values of $R$, and $L$) are different between the two versions.
4) The superobservation routine is different between the two versions.
5) SI-F is generated from a background field of monthly anomalies while SI-P is generated from the monthly totals. To create a monthly anomaly field, a monthly climatological average field based on the 2001–20 GSMaP values is subtracted from the monthly total field.

Second, visual comparisons were completed to further investigate the similarity of the algorithms. These provided insight into physical rainfall features in the analyses which were not necessarily captured by statistical analysis. All months over the full period were investigated; an arbitrary example of an individual month is provided in Fig. 3 to demonstrate the similarity between SI-F and SI-P. The comments made regarding this month are made without loss of generalization as they also applied across the study period.

The similarity between SI-P and SI-F is high, more so than their similarity to GSMaP. Both datasets importantly show strong adjustment over areas where stations exist such as over western Tasmania and along the central coast of Queensland, indicating the successful incorporation of station data. The main difference is that SI-P appears to be a bit noisier. SI-F is smoother likely because of its generation through the blending of sectors. The shorter station correlation length used in SI-P also means the radius of influence for stations in SI-P is smaller, which has an evident effect when a remote station is included, such as over central Australia.

Additionally, seasonal averages over the full period are shown in Fig. 4. The northern wet and dry seasons were selected to respectively represent very high and very low rainfall periods of the year for northern Australia, an area which has relatively low gauge paucity.

Figure 4 further demonstrates how SI-F and SI-P have improved upon GSMaP where stations exist. Again, SI-F is smoother than SI-P and tends to spread out the effects from the station correction more. During the dry season, SI-P and GSMaP have a few spots of localized elevated rainfall over the interior of Australia (e.g., in the Northern Territory and Western Australia) which are not in SI-F. These spots are likely an effect of GSMaP being too strongly calibrated to the stations in these areas and are transferred across to SI-P as well. However, they do not appear in SI-F because SI-F is computed from the monthly anomalies. The excessive calibration would be present
in both the monthly total, as well as the climatological average, and so the anomaly produced from the two has the effect of the excessive calibration removed. Originally, it was thought these spots were due to SI-F possessing an additional error routine that omitted stations that represented an excessive departure from the background field. This assumption that stations in the area are being erroneously included is not likely to be the main reason given GSMaP also displays the spots (and to a greater degree than SI-P). SI-F contains unnatural-looking straight-edged features to the west of the junction of Western Australia, the Northern Territory, and South Australia. This could be due to SI-F’s use of sectors in generating its analysis, which may result in problems over gauge-sparse areas when there is little rainfall. Overall, SI-F and SI-P show a high amount of consistency.

The final check was to evaluate if SI-P reduced error at stations, a key feature of SI. This was checked over the Australian and PNG domains. A modified mean absolute error (MAE) was used over Australia as the number of gauges changed each year. The median MAE for all the stations each year was calculated, with the median of all years then being calculated. This is in contrast to the PNG case, where the MAE for each station was computed, and then the median across the stations was found. SI-P demonstrated a clear reduction in error as seen in Table 2.

Overall, SI-P validated well against SI-F and we can be confident the algorithm was implemented correctly. The remaining validations are performed solely over the PNG domain.

b. Split-sample validation of SI-P against input station data

A split-sample validation was performed by removing one station from the algorithm, generating SI-P using this reduced set, and comparing the resultant analysis to the removed station. This was repeated for all six stations. The MAE was used as the

---

**FIG. 3.** Visual comparison of (a) GSMaP, (b) SI-F, and (c) SI-P for June 2001.
validation metric with the mean calculated across all years and the median calculated across all stations.

This split-sampled MAE was compared to the MAE of GSMaP to examine if there was a notable improvement from using SI. The results were virtually identical, with the median mean absolute error for GSMaP being only $3.12 \times 10^{-5}$ mm month$^{-1}$ greater than that of SI-P. Split-sample validation is unable to provide much insight because the adjustment for each station is limited to a small radius; this is discussed further in section 4.

c. Triple collocation analysis of SI-P

To evaluate how the two datasets compare across the entire study domain, GSMaP and SI-P are compared in a TCA. Table 2 displays summary statistics, aggregating the metric over the study domain and across the study period.

The difference between SI-P and GSMaP over the entire study domain is very small, with SI-P having slightly better metrics. Again, the difference is very small as the adjustment from SI is limited to being only around the six stations; this is discussed further in section 4.

To enable examination of the difference around the stations, the correlation and RMSE were plotted spatially with results shown in Figs. 6 and 7, respectively.

It is clear that the difference between SI-P and GSMaP is limited to around the stations. SI-P generally displays improved correlation around the six stations. Port Moresby and Momote show the most consistent improvement while the trend in performance is mixed over Wewak and Madang. Kavieng does not show much difference while no difference is evident at Misima. This is because of the land–sea mask employed; the resolution of SM2R was not sufficient to provide data over these islands (in the case of Misima, no data were available) and consequently, evaluation using TCA was hindered. The correlation appears to be lower where there is some form of topography (e.g., the mainland Highlands, east New Britain, south New Ireland, and north Bougainville).
Spatial representation of the RMSE supports the finding that the difference between GSMaP and SI-P is limited to around the stations. However, the error is consistently reduced around the stations, which is different to the case of the correlation. This is encouraging as SI is designed to reduce the analysis error. Although error appears to be always reduced, this reduction in error did not always correspond to an improvement in the correlation.

d. Time series of SI-P against included stations

To inspect the temporal variation of the algorithm over time, time series of SI-P and GSMaP at the station locations in addition to the in situ station values from 2001 to 2014 is shown in Fig. 8. Note that a comparison of split-sampled values of SI-P against removed stations was not used as split-sampled SI-P would be virtually identical to GSMaP at the removed stations given how far apart the stations in this study are.

A moving average of 4 months is used to ease visual interpretation, with four months also roughly matching the period where a clear wet and dry season can be identified across the domain (the actual period varies with location and year with remaining months being classified as transition months) (Smith et al. 2013). To further improve clarity, the stations are plotted in pairs based on proximity to each other.

GSMaP generally demonstrates greater variability than the station values, with higher averages during the wet season and lower averages during the dry season. This is unexpected as gridded analyses typically exhibit less variability than in situ observations. Over Kavieng and Momote, this tendency breaks down with more frequent occurrences of the values of GSMaP being lower than the in situ observations.

The use of SI greatly increases the consistency of GSMaP with the in situ values though there are times when the algorithm appears to have low performance, at least in terms of the 4-month rolling average. For example, in Kavieng’s dry season during 2004, there is a noticeably large discrepancy between SI-P and the in situ averages, with even GSMaP displaying more closely matching 4-month averages to the in situ data over that period. Another large discrepancy exists for Misima around June 2012 where a mismatch in the timing of the peak average is also evident. It is encouraging that a consistent direction of bias is not visually evident, with the use of a rolling average assisting in highlighting the possibility of one.

e. Representation of seasonality

As outlined in section 2a, the presence of seasonal climate drivers such as monsoonal winds over the domain is known to exist. From Fig. 8, it is evident that a degree of seasonality exists in the time series of some of the stations (such as Madang). Using GSMaP, SI-P and in situ data, this section will attempt to quantify the seasonality that may exist at the station locations, examining whether SI is able to improve the representation of seasonality. This section is also an example of how blended datasets can be used to improve upon existing climate knowledge.

The seasonality of the time series was investigated through the creation of their periodograms. These were computed using the SciPy library in Python (Virtanen et al. 2020) and shown in Fig. 9. Periodograms decompose a time series into its constituent frequencies, allowing identification of periodicity which is significant, as well as the frequency at which it occurs. A sampling frequency of 1 was selected as each month corresponded to its own data point. This yielded a frequency unit of cycles per month for the x axis while the power unit for the y axis was the square of the original unit (i.e., mm²).
For all three of the datasets, Port Moresby, Misima, and Madang had periodograms that showed a clear peak in power at a frequency of about 0.083 cycles per month, indicating the presence of a significant periodic component at this frequency. This frequency corresponds to a period of 12 months, or a year, indicating the presence of a significant annual cycle to the rainfall at these locations. The other locations that did not have a clear singular peak in their periodogram still exhibited a peak (albeit similar or smaller than their other peaks) at this frequency, indicating an annual seasonality was still present but to a less significant extent.

SI had a slight to moderate effect on increasing the similarity of GSMaP to the in situ data, as seen by some of the smaller peaks in the periodograms of SI-P having a closer match to the in situ data than GSMaP had.

Overall, where a clear seasonality exists, it was already captured in GSMaP though SI generally did have an improving effect on the representation of seasonality. The stations further away from the equator were the ones which demonstrated significant seasonality on an annual time scale.

4. Discussion

a. Residual errors in the corrected dataset

Importantly, the version of SI implemented is shown to be effective at assimilating station observations, and thus effective at improving analysis performance near stations. Section 3a demonstrated the resultant analysis had much lower biases at stations compared to the background field used as an input. Section 3c also showed that SI generally led to improvements around the stations, especially in terms of reducing error.

The heavy weighting of the analysis to stations where they exist means it is critical that the station data included are accurate. A proper quality-check routine that can remove erroneous station data would be valuable and should be considered for future implementation. SI-F uses a cross-validation routine that compares input station values to the background field and to an intermediate analysis that excludes the station being checked, with the station being excluded if the difference to either is too large. This is an example of a routine that could also be implemented in SI-P though care has to be taken to ensure accuracy.
as there have been cases in SI-F where extreme values have been improperly excluded.

However, although performance at and around the stations appears to be improved from implementing SI, performance is not significantly improved when the entire PNG domain is considered. This is evident in all three validation sections and is because of two reasons. First, the number of stations used in comparison to the total study area is very small. Furthermore, the influence of stations was designed to drop off sharply with distance to account for the high spatiotemporal variation in rainfall exhibited in some areas of PNG. As a result of these two factors, the total area for which station information could be used to significantly adjust the analysis was very small compared to the total area of the analysis.

This lack of improvement contrasts with the notable improvement observed in previous studies for monthly rainfall (Chua et al. 2022; Bhargava and Danard 1994; Ly et al. 2013). These studies were completed over a study region where there was a greater number of stations used and their relative area of influence compared to the overall domain was larger. The use of SI over PNG is further complicated because of the large amount of topography present in the domain.

Complex topography is associated with high spatiotemporal variation in rainfall which led to a greater-than-usual violation of the assumptions of error homogeneity and isotropy required for SI. Other factors such as spatial variation in rainfall errors due to different climatic zones and rainfall modes also contribute to these violations. To constrain the effects of these violations, a much lower-than-usual station correlation length scale had to be used, explaining the small station radii of influences in this study. Previous studies (e.g., Diodato 2005) note the breakdown of geostatistical interpolation methods based on homogenous assumptions over regions with complex topography, with the use of elevation information being a valuable input, albeit one which is unable to be directly utilized in SI.

It is important to consider that the large uncertainty in truth and shared biases between the reference datasets make it difficult to perform a gridded comparison with great certainty.
across the domain. Even though TCA is considered more robust to biases in the reference datasets confounding validation (compared to validation against a single dataset), it is not immune, and the two reference datasets used in TCA in this study (ERA5 and SM2R) are expected to have significant biases of their own over PNG. If these biases spatially align with the biases of GSMaP, the performance of GSMaP would be inaccurately inflated or deflated.

It should also be noted that even for a grid point which is collocated exactly with an included station, the analysis would not be adjusted to be equal to the station value (explaining the nonzero RMSEs obtained from the in situ validation in section 3a). This is by design as the algorithm accounts for the existence of observational errors including from instrumental biases and from the spatial representation difference incurred from translating from an in situ station value to a gridded average.

In this study, the proportion of the total error variance that can be apportioned to the background field ($R_z$) was between 0.58 and 0.84, which is lower than the values obtained when a
FIG. 9. Periodograms of the time series based on GSmAP, SI-P, and in situ data at the station locations.
background field of station climatology was used in Australia (Evans et al. 2020). This is logical as a background field based on satellite precipitation estimates for the month should be closer to the “true” field than a climatological field would be. This factor also contributes to a smaller station influence than that observed in earlier SI studies (Evans et al. 2020; Chua et al. 2022).

b. Future work

The optimization of priors \((R, L)\) used in this algorithm is considered an important research topic for the future. A key point is improving the representation of the inhomogeneity of rainfall error structure across the domain. This could be accomplished by the creation of different regions or climatic zones, for which the priors can be individually computed. The use of a different correlation model could further improve this. Addressing nonstationarity and anisotropy in rainfall modeling is a complex topic; one contemporary way of achieving this is by modeling rainfall stochastically and using the ensemble of modeled fields to create an anisotropic correlation model (Nerini et al. 2017).

Another approach to improving the performance of this algorithm would be to investigate the addition of extra explanatory variables to complement the information provided by rain gauges. This could be valuable, especially over domains like PNG that have a low number of rain gauges. A natural variable to consider would be elevation given the strong influence topography has on rainfall, in addition to digital elevation models being a readily accessible dataset. Geospatial interpolation techniques that use elevation as an additional variable have been explored in the past, including in cokriging (Adhikary et al. 2017) and empirical Bayesian kriging with regression prediction (Ali et al. 2021), with its inclusion generally improving performance. The degree of improvement has varied with study area, though given the paucity of gauge information over PNG coupled with its significant topography, the addition of elevation data would be a good candidate to trial. Meteorological variables such dry-bulb temperature and wind speed (Babel et al. 2015), and airflow indices based on mean sea level pressure (Kilsby et al. 1998), could also be considered as additional explanatory variables but are likely limited in value given they are affected by the same data paucity as the rainfall observations over PNG. As mentioned in section 4a, additional variables cannot be directly used in SI but a simple way of including them could be using linear regression to create additional correction factors that can then be applied to the SI output.

To further quantify seasonality, the jumps upon spectrum and trend (JUST) method which is based on least-squares spectral analysis (LSSA) could be used (Ghaderpour 2021). LSSA attempts to break a time series down into trend and seasonal components by iteratively fitting sinusoidal components to it (Lomb 1976). However, the use of JUST requires careful selection of the involved parameters and thus would be better suited for a more detailed study focused on seasonality.

5. Conclusions

Satellite precipitation datasets offer an effective way of estimating rainfall where in situ data are limited. However, they can also possess significant biases, meaning the assimilation of in situ data is extremely valuable in improving accuracy. Statistical interpolation (SI) is a classical data assimilation technique that forms a weighted average between a background field and in situ observations based on correlation and error information. Over a gauge-sparse region, the use of satellite precipitation estimates as the background field has been demonstrated to produce superior performance over relying purely on in situ rain gauge data. However, the effectiveness of the algorithm in terms of improving upon the satellite estimates is underexplored.

Papua New Guinea (PNG) was selected as a study area, which would pose complications for the algorithm, both in terms of its gauge paucity and its significant topography. PNG also lacks an operational gauge-based analysis which the OI-derived dataset produced in this study could satisfy. There does not appear to be an open-source Python-based version of the algorithm which is another gap this study can fill.

The OI algorithm was successfully implemented in Python 3 for monthly rainfall and shown to be consistent with a known existing implementation. Next, split-sample in situ and triple collocation analysis (TCA) validations over PNG were performed. When performance is considered across the whole domain, the improvement gained from OI is slight, with only the error statistic from the TCA showing a perceptible decrease from 2.45 to 2.43 mm day\(^{-1}\). The lack of a significant improvement is because the area of influence of the station data is small, especially compared to the area of the overall domain. The area of influence is small as there are only six stations in this comparatively large domain, in addition to the radius of influence of the individual stations having to be forced to be small to account for the high spatiotemporal variation of rainfall induced by the topography near some of the stations. This demonstrates that the value of OI is generally heavily limited when gauge density is extremely low, and the spatiotemporal variation of rainfall is high (such as where topography is complex).

However, when only performance around the stations is considered, there was a noticeable improvement gained from using OI with the error consistently being reduced and a general increase in correlation metric. This means that although OI did not yield significant domain-wide improvement over the input background field when gauge paucity is extremely high, it is still valuable given the performance increase around the included stations. In an operational context, this value is increased as OI results in better consistency between the gridded analysis and the in situ data.

In an example of the value a gridded dataset can provide for climatological knowledge, the seasonality of the station locations was investigated through the use of periodograms. Only half of the stations analyzed (Port Moresby, Misima, and Madang) demonstrated a significant identifiable seasonality, which occurred on an annual cycle. This seasonality was represented in both the corrected and uncorrected GSMaP,
though the use of SI led to a slight to moderate improvement in its representation.

Although the algorithm was used to create monthly rainfall analyses in this study, it was designed so that both the background field and in situ datasets can be easily adapted. In addition to being amendable to different time scales and rainfall datasets, it can also be adapted for the assimilation of other geospatial variables.

Acknowledgments. We are grateful to Nathan Eizenberg and Dr. Yan Wang for their contributions to our understanding of the statistical interpolation (SI) algorithm. We are also appreciative of colleagues from the Climate Monitoring and Long-Range Forecasts sections of the Australian Bureau of Meteorology for their helpful advice and guidance. Author contributions: Conceptualization, Z.-W. C., Y. K., A. W., S. C. and C. S.; methodology, Z.-W. C. and Y. K.; software, Z.-W. C.; validation, Z.-W. C.; formal analysis, Z.-W. C.; investigation, Z.-W. C.; resources, Z.-W. C.; data curation, Z.-W. C.; writing—original preparation, Z.-W. C.; writing—review and editing, Z.-W. C., Y. K., A. W., S. C. and C. S.; visualization, Z.-W. C.; supervision, Z.-W. C., Y. K., A. W., S. C. and C. S.; project administration, Z.-W. C. and Y. K. All authors have read and agreed to the published version of the manuscript. The authors declare no conflict of interest.

Data availability statement. GSMaP data were provided by EORC, JAXA. Station gauge data were provided by the Papua New Guinea National Meteorological Service. Contains modified Copernicus Climate Change Service Information (2019). Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus Information or Data it contains. A repository for the SI code in this study can be found at https://github.com/ZC-BOM/optimal-interpolation.

APPENDIX

Satisfaction of TCA Criteria

The degree to which the datasets used in the TCA (SM2R, ERA5, GSMaP, and SI-P) satisfy the assumptions required for TCA are discussed below and displayed in Table A1:

1) Orthogonality of errors (the expected sum of the errors is zero). A time series of errors for each dataset was computed by using MSWEP as truth. These error time series were then normalized by the time series of mean values to obtain relative magnitudes. The biases for the datasets are less than 20% of the mean. GSMaP, SI-P, and ERA5 have larger biases, but this is in part due to the high uncertainty in truth where MSWEP is also likely to contain biases of its own that would inflate the biases obtained.

2) No cross correlation among the errors of the datasets, as well as with the truth. Using the error time series, the linear correlation of GSMaP and SI-P with ERA5 and SM2R was calculated. Correlations between errors are reasonably low, with the highest value being around 0.5. Complete independence is unrealistic as there are factors that commonly affect accuracy between datasets (e.g., topography). Some correlation from shared data sources is also expected (SM2R contains bias correction to gauges and ERA5 ingests some satellite-based moisture-related information (not precipitation estimates).

Table A1. Metrics testing whether the assumptions required for TCA are satisfied.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ADF statistic</th>
<th>Ratio of bias to mean (%)</th>
<th>R to SM2R</th>
<th>R to ERA5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM2R</td>
<td>−2.76</td>
<td>−3.68</td>
<td>7.14</td>
<td>0.84</td>
</tr>
<tr>
<td>ERA5</td>
<td>−1.02</td>
<td>−4.15</td>
<td>19.99</td>
<td>0.84</td>
</tr>
<tr>
<td>GSMaP</td>
<td>0.59</td>
<td>−7.39</td>
<td>14.29</td>
<td>0.70</td>
</tr>
<tr>
<td>SI-P</td>
<td>0.59</td>
<td>−7.40</td>
<td>14.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

3) Stationarity of data. For each of the datasets, an augmented Dickey–Fuller test (ADFT) was performed on both the time series of monthly values and the time series of errors. The ADFT tests the null hypothesis that a unit root exists in the dataset, thereby indicating nonstationarity (Said and Dickey 1984). The more negative a test value is, the greater the confidence that the dataset is stationary. SM2R and ERA5 demonstrate a high degree of stationarity but there appears to be some nonstationarity to GSMaP and SI-P. However, when the ADFT is calculated for GSMaP over the longer period of 2007–2020, the statistic decreases to −2.05 suggesting that the nonstationarity is likely an effect of the shorter study period and not of the dataset itself.

4) The datasets can be linearly related to each other. The linear correlation between the time series of each dataset was computed. All the datasets demonstrate a high linear correlation to each other.

Overall, the datasets generally satisfy the assumptions required.

REFERENCES


