DESIGN AND CREATION OF A DATABASE TO ASSESS THE INFORMATION NEEDS OF HYDROLOGICAL MODELS

LUIS IZQUIERDO-HORNA, JOSE ZEVALLOS, THURIAN CEVALLOS & DANIELA RIOS
Department of Engineering, Universidad Tecnológica del Perú, Peru

ABSTRACT

Hydrological models are vital in water resource management and emergency response, providing essential insights for managing river basins and mitigating flood risks. For these models to effectively function, they require accurate and comprehensive data. This study focuses on designing and creating a database specifically for the SAC, GR4J and SOCONT hydrological models used within the RS MINERVE hydrological modelling hub in the context of the Piura River basin. The research involves identifying and collecting precipitation, temperature and river discharge data from national organisations such as the National Service of Meteorology and Hydrology of Peru (SENAMHI) and the National Water Authority (ANA). Characterised by their temporal and spatial resolutions, the data are organised using standardised data formats and relational database structures to ensure interoperability and ease of access. To ensure the models function optimally, the collected data are subjected to rigorous validation for quality and consistency through statistical testing. The developed database aims to support the accurate assessment of information needs in hydrological modelling. Preliminary results indicate considerable improvements in model accuracy and reliability, facilitating effective water resource management and decision-making processes in the region.

Keywords: hydrological models, database creation, data quality validation, water resource management, Piura River basin.

1 INTRODUCTION

Hydrological models are designed to study and simulate water behaviour in watersheds [1]; however, their accuracy and reliability largely depend on the quality and availability of input data [2]. In addition, the lack of understanding the real impact of hydrometeorological forecasts on vulnerable populations hinders the ability to make informed decisions [3]. To address this, the creation of a comprehensive database is proposed to provide the necessary data to enhance the application of tools such as SOCONT, GR4J, and SAC-SMA, thereby increasing the accuracy and usefulness of forecasts in water management and risk mitigation.

The SOCONT, GR4J and SAC-SMA hydrological models, part of the RS-MINERVE software, are conceptual tools of varying complexity that allow for the detailed simulation of the hydrological behaviour of a watershed [4]. The SOCONT (observation and control system for water levels) hydrological model integrates snow, infiltration, and runoff models, using input data such as precipitation, temperature, and evapotranspiration [5]. Recently, SOCONT has been used to simulate extreme flood events under summer conditions [6] and to assess changes in the hydrological regime of watersheds [7].

Conversely, the GR4J (Spanish acronym for Daily Rural Engineering with Four Parameters) model is a conceptual rainfall-runoff model characterised by its simplicity, using precipitation and temperature as input data [8]. It has been applied to simulate hydrological processes in various tropical and subtropical watersheds [9]. Finally, the SAC-SMA conceptual hydrological model is widely used to simulate runoff flow using precipitation and potential evapotranspiration data [10].

www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/SDP240511

2 AREA OF STUDY

The Piura River basin is located in the department of Piura, northern Peru, covering an area of 10,872 km^2 across the provinces of Piura, Morropón, Huancabamba, Ayabaca and Sechura [11]. The region experiences a semiarid climate with an average annual precipitation of 400 mm [12]. There are 47 hydrometeorological stations distributed throughout the basin. Fig. 1 shows the location map of the study area.

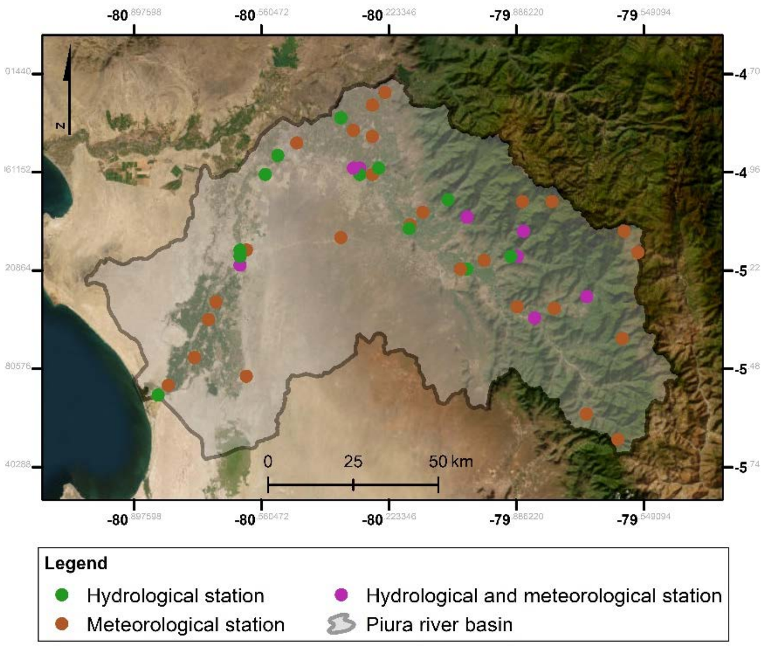


Figure 1: Location map of the Piura River basin, Department of Piura, Peru.

3 INPUT VARIABLES FOR THE DATABASE

3.1 Precipitation

Precipitation is a meteorological phenomenon where water vapor in the atmosphere condenses and falls to the surface of the Earth in various forms such as rain, snow, or hail [13]. It is measured using instruments called rain gauges, which record the amount of accumulated water in millimetres over a specified period (e.g., day, month, and year) [14]. In this study, we used daily precipitation data.

3.2 Temperature

Temperature is a crucial physical property of matter that measures the intensity of heat in a system, be it solid, liquid, or gaseous [15]. At meteorological stations, temperature is measured using sensors that record daily temperature data over extended periods [16]. Furthermore, air temperature influences the formation and accumulation of ozone in the atmosphere, impacting both air quality and public health [17].

3.3 Flow

The flow of a river refers to the minimum amount of water required to flow through its course to maintain ecological values, measured in $m^3 s^{-1}$ [18]. This value is determined by carefully analysing the minimum needs of the ecosystems within the basin, considering any alterations to the river's natural flow [19]. This variable serves as the primary parameter for adjusting and validating the model's performance by minimising the discrepancy between simulated and observed flow [20].

3.4 Digital elevation model

A digital elevation model (DEM) is a three-dimensional depiction of the terrain surface, offering detailed spatial information about the topography of an area of interest [21]. The GR4J, SOCONT, and SAC-SMA hydrological models can greatly benefit from incorporating spatial and land use data, improving the accuracy of their simulations [22]. In particular, the DEM is used to define the watershed, assess the topography, and determine flow directions [23].

3.5 Soil type

The soil variable provides information about soil properties such as infiltration capacity and water retention [24]. While soil maps are not strictly necessary for the basic operation of the selected models, their use greatly benefits the calibration and adjustment of model parameters, thereby enhancing simulation accuracy [25].

3.6 Land use

Different types of land use and vegetation cover in the watershed considerably impact infiltration and evapotranspiration [26]. This information is valuable due to the data it provides on evapotranspiration (ETP). Each type of vegetation cover exhibits different evapotranspiration rates, while urbanised and paved areas tend to generate more surface runoff due to lower infiltration capacity, leading to considerable variability in soil water retention based on predominant land use [27], [28].

The following sections outline the process conducted to assess the precipitation variable, identified as the most relevant due to its high variance over short time intervals [29], especially in arid climates.

4 DESIGN AND IMPLEMENTATION OF THE DATABASE

Database creation with data from rainfall and meteorological stations owned by the National Water Authority (ANA) and the National Service of Meteorology and Hydrology of Peru (SENAMHI) aimed to gather hydrometeorological information by systematically organising data. This effort included collecting geographical coordinates, station data, recording periods, and data sources to enhance information visualisation and management. For the selection of the recording period, a prolonged interval was established because the accuracy of frequency estimates is strongly linked to sample size [30]. This period spanned from 1980 to 2016, excluding the years affected by the El Niño events of 1982–1983, 1997–1998, and 2016–2017. The reason for these exclusions was that the selected hydrological models lacked the capability to manage sudden trend changes, potentially resulting in either overestimated average flows or underestimated critical flows [9].

The data from these stations was consolidated into a unified input dataset, maintaining consistency in recording periods across all station types. After a first filter, the number of stations was reduced to five meteorological stations with information on the precipitation variable: Chusis, Chalaco, Huarmaca, Huancabamba and Miraflores. On the other hand, the RSMINERVE software requires input data to be in a specific format; therefore, the data file must be in CSV format, where columns are separated by commas and rows by line breaks. The first column of the file must contain variable names. Mandatory variables include Station, which refers to the name of the meteorological station; X and Y, representing the geographic longitude and latitude of the station in UTM coordinates, respectively; Z, indicating station altitude above sea level; Category, describing the type of information provided by the station, such as precipitation or temperature; Sensor, an abbreviation for the type of sensor used to measure the variable, like P for precipitation; and Measurement Units, which for precipitation would be mm/month. Finally, the Interpolation Method must be defined, with 'ConstantBefore' used for precipitation. It should be noted that there are no spaces between words. If necessary, an underscore (_) must be used to replace blank spaces. Stations must be listed vertically, and station data must be organised by selected periods.

5 VALIDATION OF THE DATABASE

The validation of the database involves three main stages (Fig. 2). The first stage is an initial evaluation of the collected data, which includes a visual inspection to ensure there are no negative values, detect any distortions in the series mean, and confirm the presence of credible limits, with a daily rainfall limit of 305 mm being the maximum value a station can detect [31]. The next step involves developing double mass curve graphs to visually inspect for potential breaks in the data. These graphs plot the cumulative average precipitation of all basin stations on the x-axis and the cumulative daily volumes of each station under study on the y-axis.

Finally, statistical homogeneity analyses (e.g., Pettitt, SNHT and Buishand) are conducted to determine if there are notable breaks in the data time series. Using multiple statistical tests increases our confidence in the results if all tests indicate a similar trend direction [32]. These analyses are conducted annually due to daily data variability [33]. The Pettitt method is used to identify notable breakpoints in chronologically ordered meteorological data without assuming any specific distribution. The standard normal homogeneity test (SNHT) tends to detect changes at the beginning and end of the time series. Meanwhile, the Buishand model uses cumulative deviations from the mean to identify change points. These models are combined to verify homogeneity in the time series, confirming breaks if at least two of them reject the null hypothesis of homogeneity ($p > 0.5$) [34]–[37]. Finally, if a station with a breakpoint is identified, a homogeneity adjustment will be applied using the mean and standard deviation correction method for the oldest discordant period [38], [39].

With the validated information, missing data were completed using linear regression and random forest methods, optimised with an algorithm based on the root mean square error (RMSE). This statistical index measures the accuracy of a predictive model. Ideally, RMSE results should approach zero, indicating high model precision and predictions that are very close to the actual values [40]. This approach ensured that gaps were filled with the highest possible accuracy by evaluating the performance of both methods to select the most suitable one for each station. Linear regression predicts unknown data values based on related and known data, establishing a linear relation between a reference station and the one with missing data [41]. In contrast, random forest is a decision tree-based algorithm that uses multiple trees for prediction, displaying predictive variables at each node [42], [43].

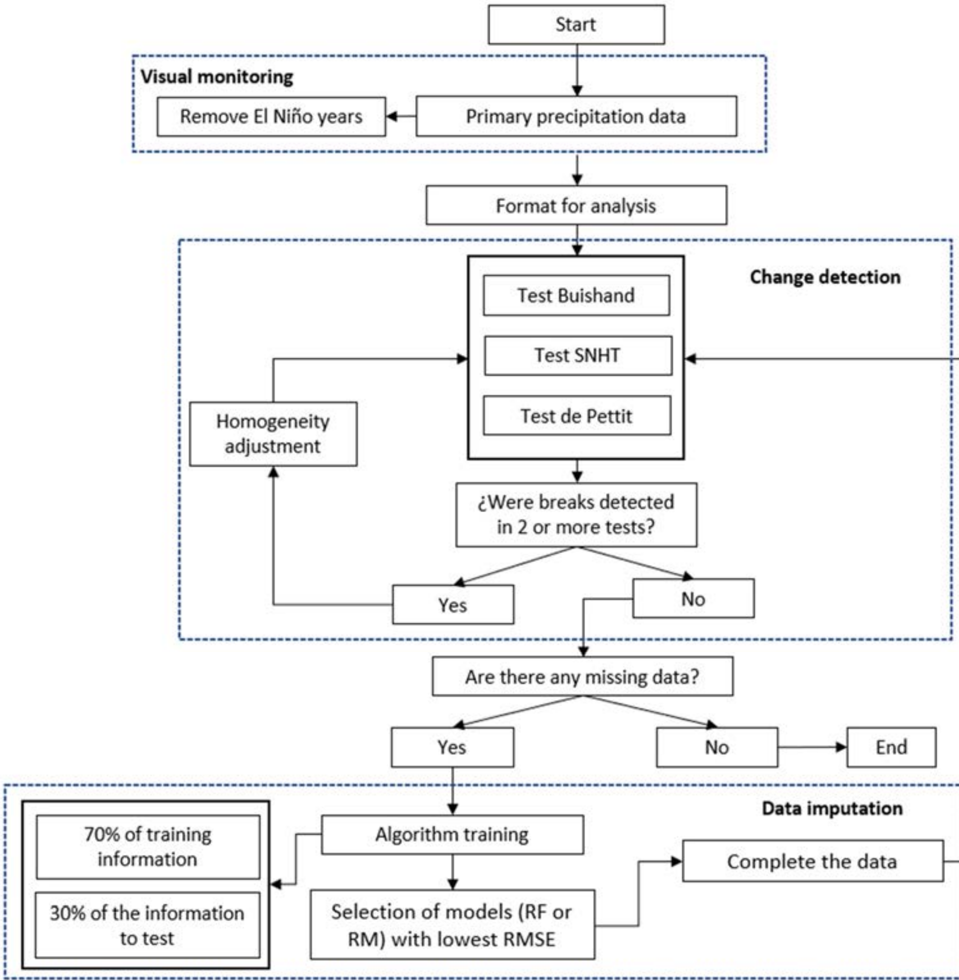


Figure 2: Flowchart representing the links between the main steps (i.e., visual monitoring, change detection, and data imputation) in the proposed research methodology.

6 RESULTS AND DISCUSSION

6.1 Visual analysis of the precipitation variable

According to Fig. 3, the Chalaco and Huancabamba stations indicate good data consistency, while the graphs for Chusis, Miraflores and Huarmaca show some interruptions in the diagram due to incomplete data.

6.2 Homogenisation

The statistical analysis of the incomplete data revealed that at the Chusis station, the SNHT and Buishand tests did not reject the null hypothesis of homogeneity, suggesting that the

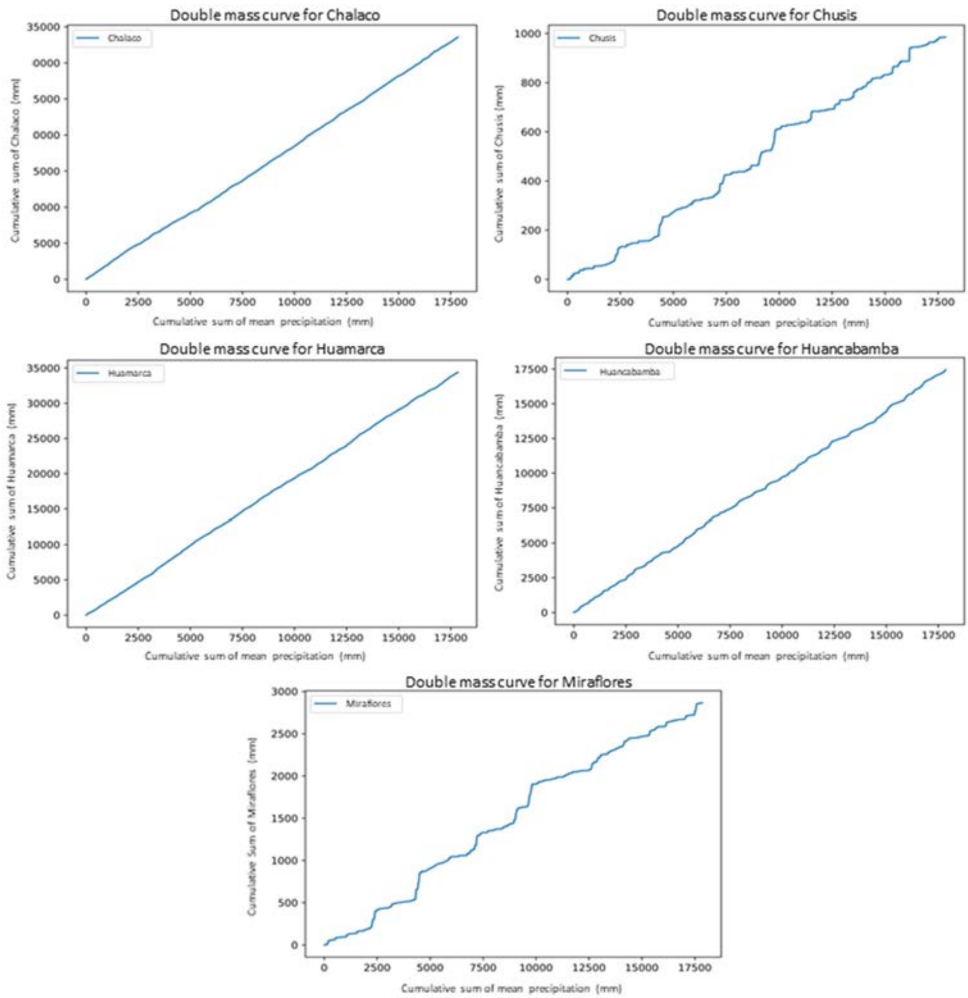


Figure 3: Double mass curve of the primary precipitation data for each meteorological station representing an overview of potential breaks.

change detected by the Pettitt test might not be notable or could have been influenced by random factors [44] (Table 1). A similar situation was observed at the Chalaco station, where only the Buishand test detected a break, while the Pettitt and SNHT tests did not. Finally, at the Huarmaca, Huancabamba, and Miraflores stations, none of the three tests rejected the null hypothesis, indicating no considerable change in the data distribution during the period assessed.

Notably, the results of the trained model indicate that linear regression is a more effective data imputation method than random forest for the stations of interest, as evidenced by a lower RMSE (Table 2). This effectiveness can be attributed to linear regression's greater interpretability, which makes it easier to understand the influence of explanatory variables on the response variable [45]. In addition, linear regression is a simpler model and less prone to overfitting, especially in large datasets or those with many explanatory variables [46].

Table 1: Statistical test results for precipitation from meteorological stations assessing primary data.

Gauge station	Pettit	Pettit *p*-value	SNHT	SNHT *p*-value	Buishand	Buishand *p*-value
Chusis	**True**	0.0093	False	0.1333	False	0.0888
Chalaco	False	0.0132	False	0.0622	**True**	0.0297
Huarmaca	False	0.1463	False	0.2310	False	0.0793
Huancabamba	False	0.0239	False	0.0574	False	0.0803
Miraflores	False	0.0854	False	0.9530	False	0.8167

Table 2: Results from the RMSE values and model selection for the imputation of precipitation data for each meteorological station.

Gauge station	RMSE LR	RMSE RF
Chusis	3.4724	3.8338
Chalaco	4.5171	4.7253
Huarmaca	6.1898	6.6643
Huancabamba	1.9306	2.0917
Miraflores	0.8075	0.8746

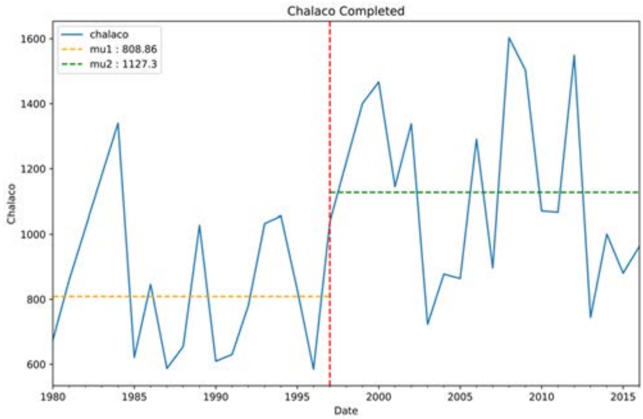
In the statistical analysis of the completed data (Table 3), it was observed that at the Chalaco station, the Pettitt, SNHT and Buishand tests rejected the null hypothesis of homogeneity, indicating the detection of an abrupt change in precipitation amount at some point during the analysed period. From the initial analysis, the Chalaco station exhibited signs of a trend change according to the Buishand model. In contrast, at the Chusis, Huarmaca, Huancabamba and Miraflores stations, none of the three statistical models found evidence of a considerable change in means. Therefore, these four stations do not evidence heterogeneous data.

Table 3: Statistical test results on precipitation data from meteorological stations using imputed data.

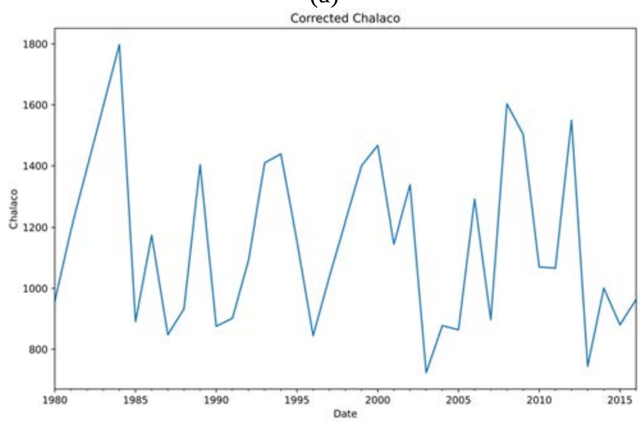
Gauge station	Pettit	Pettit *p*-value	SNHT	SNHT *p*-value	Buishand	Buishand *p*-value
Chusis	False	0.3805	False	0.5276	False	0.4179
Chalaco	**True**	0.0033	**True**	0.0189	**True**	0.0130
Huarmaca	False	0.2138	False	0.3723	False	0.1016
Huancabamba	False	0.0281	False	0.0701	False	0.0643
Miraflores	False	0.4322	False	0.8608	False	0.8103

In the same line, Fig. 4(a) denotes a break in the annual precipitation mean starting from the year 1995 at the Chalaco station, increasing from approximately 800 to 1100 mm. Fig. 4(b) displays the corrected data, reflecting an adjusted time series following the identification of this change point.

www.witpress.com, ISSN 1743-3541 (on-line)



(a)



(b)

Figure 4: (a) Annual series from the Chalaco meteorological station with uncorrected imputed precipitation data; and (b) Annual series from the Chalaco meteorological station with corrected imputed precipitation data.

7 CONCLUSION

Based on the study findings, initially all stations demonstrated homogeneity, though a 17% missing data rate was addressed using regression algorithms. However, the Chalaco station exhibited heterogeneity that required correction. The cause of this heterogeneity in Chalaco could not be conclusively attributed to non-climatic factors due to the lack of confirmatory metadata.

To improve the robustness of the findings, increasing the number of stations and conducting grouped analyses with nearby stations is recommended. Likewise, no outliers were identified outside of the El Niño years, suggesting minimal impact on the homogeneity analysis of the precipitation time series.

Ultimately, the methodology proposed in this study is expected to not only facilitate future hydrometeorological research in similar regions but also provide crucial tools for validating and continually improving data quality.

ACKNOWLEDGEMENTS

This project was funded by Universidad Tecnológica del Perú, within the framework of the 'Research Projects I+D+i 2024' agreement. The authors would like to thank anonymous reviewers for their valuable comments on the previous version of this manuscript.

REFERENCES

[1] Llauca, H., Leon, K. & Lavado-Casimiro, W., Construction of a daily streamflow dataset for Peru using a similarity-based regionalization approach and a hybrid hydrological modeling framework. *J. Hydrol. Reg. Stud.*, **47**, 101381, 2023.

[2] Baltazar, L.A. et al., Hydrological modeling in a region with sparsely observed data in the eastern Central Andes of Peru, Amazon. *J. South Am. Earth Sci.*, **121**, 104151, 2023. https://doi.org/10.1016/j.jsames.2022.104151.

[3] Kruczkiewicz, A. et al.. Earth observations for anticipatory action: Case studies in hydrometeorological hazards. *Earth Observation for Flood Applications*, ed. G.J.-P. Schumann, Elsevier, pp. 237–253, 2021. https://doi.org/10.1016/B978-0-12-819412-6.00011-0.

[4] Muñoz, R. et al.. Comparing model complexity for glacio-hydrological simulation in the data-scarce Peruvian Andes. *J. Hydrol. Reg. Stud.*, **37**, 100932, 2021. https://doi.org/10.1016/j.ejrh.2021.100932.

[5] Todorović, A., Grabs, T. & Teutschbein, C., Improving performance of bucket-type hydrological models in high latitudes with multi-model combination methods: Can we wring water from a stone? *J. Hydrol.*, **632**, 130829, 2024. https://doi.org/10.1016/j.jhydrol.2024.130829.

[6] Zeimetz, F. et al., Relevance of the correlation between precipitation and the 0°C isothermal altitude for extreme flood estimation. *J. Hydrol.*, **551**, pp. 177–187, 2017.

[7] Osuch, M., Wawrzyniak, T. & Majerska, M., Changes in hydrological regime in High Arctic non-glaciated catchment in 1979–2020 using a multimodel approach. *Adv. Clim. Chang. Res*., **13**, pp. 517–530, 2022.

[8] Cantoni, E. et al.. Hydrological performance of the ERA5 reanalysis for flood modeling in Tunisia with the LISFLOOD and GR4J models. *J. Hydrol. Reg. Stud*., **42**, 101169, 2022.

[9] Andrade, J.M. et al., Efficiency of global precipitation datasets in tropical and subtropical catchments revealed by large sampling hydrological modelling. *J. Hydrol.*, **633**, 131016, 2024. https://doi.org/10.1016/j.jhydrol.2024.131016.

[10] Zhang, Z. et al., SAC-SMA a priori parameter differences and their impact on distributed hydrologic model simulations. *J. Hydrol*., **420–421**, pp. 216–227, 2012.

[11] ANA, Surface water resource assessment in the Piura River Basin, 53.

[12] Yakirevich, A. et al., Modeling the impact of solute recycling on groundwater salinization under irrigated lands: A study of the Alto Piura aquifer, Peru. *J. Hydrol*., **482**, pp. 25–39, 2013.

[13] Wu, X. et al., Multi-source precipitation products assessment on drought monitoring across global major river basins. *Atmos. Res*., **295**, 106982, 2023.

[14] Si, P. et al., Construction of daily precipitation series and the observational characteristics of extreme precipitation in Tianjin, China during 1888–2022. *Adv. Clim. Chang. Res*., **15**, pp. 52–61, 2024.

[15] SENAMHI, Air temperature and precipitation atlas of Peru, 2021. http://repositorio.senamhi.gob.pe/handle/20.500.12542/1339.

www.witpress.com, ISSN 1743-3541 (on-line)

[16] Song, B. & Park, K., Temperature trend analysis associated with land-cover changes using time-series data (1980–2019) from 38 weather stations in South Korea. *Sustain. Cities Soc*., **65**, 102615, 2021.

[17] Wang, J. et al.. Relationship between ozone and air temperature in future conditions: A case study in sichuan basin, China. *Environ. Pollut*., **343**, 123276, 2024.

[18] Sai Louie, A.J. et al., Testing the reproducibility of active-distributed temperature sensing for measuring groundwater specific discharge beneath a braided river. *J. Hydrol*., **633**, 130877, 2024.

[19] ANA, Guide for the elaboration and implementation of integrated water resources management plans in Nicaraguan watersheds.

[20] Sirisena, T.A.J.G., Maskey, S. & Ranasinghe, R., Hydrological model calibration with streamflow and remote sensing based evapotranspiration data in a data poor basin. *Remote Sens.*, **12**, pp. 1–24, 2020.

[21] Moges, D.M. et al., How does the choice of DEMs affect catchment hydrological modeling? *Sci. Total Environ.*, **892**, 164627, 2023. https://doi.org/10.1016/j.scitotenv.2023.164627.

[22] Tran, T.N.D. et al., Quantification of global digital elevation model (DEM): A case study of the newly released NASADEM for a river basin in Central Vietnam. *J. Hydrol. Reg. Stud*., **45**, 101282, 2023.

[23] Pearson, R.A. et al., GeoFabrics 1.0.0: An open-source Python package for automatic hydrological conditioning of digital elevation models for flood modelling. *Environ. Model Softw*., **170**, 105842, 2023. https://doi.org/10.1016/j.envsoft.2023.105842.

[24] Horta, A. et al., Assessing the use of digital soil maps in hydrological modeling for soil-water budget simulations: Implications for water management plans in southern Portugal. *Geoderma Reg.*, **36**, e00741, 2024.

[25] Nan, Y. & Tian, F., Isotope data-constrained hydrological model improves soil moisture simulation and runoff source apportionment. *J. Hydrol.*, **633**, 131006, 2024.

[26] Dalla Torre, D. et al., Suitability of ERA5-Land reanalysis dataset for hydrological modelling in the Alpine region. *J. Hydrol. Reg. Stud.*, **52**, 101718, 2024.

[27] Dash, S.S., Naik, B. & Kashyap, P.S., Assessment of land use/land cover change derived catchment hydrologic response: An integrated parsimonious hydrological modeling and alteration analysis based approach. *J. Environ. Manage.*, **356**, 120637, 2024.

[28] Liu, C. et al., Modeling hydrological consequences of 21st-century climate and land use/land cover changes in a mid-high latitude watershed. *Geosci. Front.*, **15**, 101819, 2024.

[29] Pendergrass, A.G. et al., Precipitation variability increases in a warmer climate. *Sci. Rep.*, **7**, pp. 1–9, 2017.

[30] Porth, L.S. et al.. Development of a technique to determine adequate sample size using subsampling and return interval estimation. *J. Hydrol.*, **251**, pp. 110–116, 2001.

[31] Huerta, A., Combination of observed and satellite data: Chillón, Rímac and Lurín basins (CHIRILU), 2020. https://doi.org/10.6084/m9.figshare.13260020.v3.

[32] Assefa, A. et al., Impact of sustainable land management on vegetation cover using remote sensing in Magera micro watershed, Omo Gibe Basin, Ethiopia. *Int. J. Appl. Earth Obs. Geoinf.*, **103**, 102495, 2021.

[33] Alexandersson, H. & Moberg, A., Homogemzation of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol.*, **17**, pp. 25–34, 1997.

[34] Deb, S., Analyzing trends and change points in hydro-meteorological parameters and groundwater level in the Barak river basin in India. *Phys. Chem. Earth*, **134**, 103542, 2024.
[35] Khoualdia, W. et al., Climate variability and its impact on water resources: case study of the Souk Ahras region, northeastern Algeria. *Desalin. Water Treat.*, **296**, pp. 10–18, 2023.
[36] Animashaun, I.M. et al., Rainfall analysis over the Niger Central Hydrological Area, Nigeria: Variability, trend, and change point detection. *Sci. African*, **8**, e00419, 2020.
[37] Tahiri, A. et al., Influence of climate variability on water resource availability in the upper basin of Oum-Er-Rabiaa, Morocco. *Groundw. Sustain. Dev.*, **19**, 100814, 2022.
[38] Peterson, T.C. et al., Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.*, **18**, pp. 1493–1517, 1998.
[39] White, M.D. & Greer, K.A., The effects of watershed urbanization on the stream hydrology and riparian vegetation of Los Peñasquitos Creek, California. *Landsc. Urban Plan.*, **74**, pp. 125–138, 2006.
[40] Mentaschi, L. et al., Problems in RMSE-based wave model validations. *Ocean Model*, **72**, pp. 53–58, 2013.
[41] Daniel Francisco, C.-A., Modelling the monthly rainfall-runoff process using regressions. *Ing. Investig. y Tecnol.*, **15**, pp. 625–636, 2014.
[42] Shen, Y. et al., Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. *Comput. Geosci.*, **159**, 105019, 2022.
[43] Mosavi, A., Ozturk, P. & Chau, K.W., Flood prediction using machine learning models: Literature review. *Water (Switzerland)*, **10**, pp. 1–40, 2018.
[44] Aryal, A. et al., Spatio-temporal variability of streamflow in major and medium rivers of Nepal. *J. Hydrol. Reg. Stud.*, **50**, 101590, 2023.
[45] Suárez, M.G., Retail shelf space occupied by private label brands: estimation by neural networks vs. multiple regression. *Cuad. Econ. y Dir. la Empres.*, **12**, pp. 37–66, 2009.
[46] Li, M. et al.. Assessing landslide susceptibility based on the random forest model and multi-source heterogeneous data. *Ecol. Indic.*, **158**, 111600, 2024.

www.witpress.com, ISSN 1743-3541 (on-line)