Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería Agrícola
Programa de Doctorado en Recursos Hídricos y Energía para la Agricultura.

"USO DE DATOS DE PRECIPITACIÓN GRILLADA PARA MEJORAR LA MODELACIÓN HIDROLÓGICA"

Tesis para optar el grado de Doctor en Recursos Hídricos y Energía para la Agricultura

DOMINGO MARCELO PORTUGUEZ MAURTUA
CHILLÁN - CHILE
2023

Profesor guía: José Luis Arumí Ribera
Dpto. de Recursos Hídricos
Facultad de Ingeniería Agrícola
Universidad de Concepción

Esta tesis ha sido realizada en el Departamento de Recursos Hídricos de la Facultad Ingeniería Agrícola, Universidad de Concepción.

Profesor Guía

_____

Dr. José Luis Arumí Ribera
Facultad de Ingeniería Agrícola
Universidad de Concepción

Comisión Evaluadora:

_____

Dr. Enrique Muñoz Ortiz
Facultad de ingeniería
Universidad Católica de la Santísima Concepción

_____

Dra. Alejandra Stehr Gesche
Facultad de Ingeniería
Universidad de Concepción

_____

Dr. Octavio Lagos Roa
Facultad de Ingeniería Agrícola
Universidad de Concepción

Director de Programa

_____

Dr. Sebastián Alberto Krogh Navarro
Facultad de Ingeniería Agrícola
Universidad de Concepción

AGRADECIMIENTO

A mi esposa y compañera de vida, Eliana Gabriela, por todo su amor y apoyo incondicional en cada momento, sobre todo en los momentos difíciles; a mis hijas Josselyn y Abigail, mis motores y motivos de inspiración en la vida para ser cada vez mejor; a mis padres, Toribio y Bartola, por brindarme sus consejos, motivación y todo su amor.

A mi profesor guía Dr. José Luis Arumí por su constante apoyo, confianza y amistad, además de alentarme en cada una de las fases de la investigación que fueron fundamentales para finalizar la tesis. A los profesores Dra. Alejandra Stehr Gesche, Dr. Octavio Lagos Roa y Dr. Enrique Muñoz Ortiz, miembros de la Comisión Evaluadora, por su tiempo de revisión y sugerencia de mejora de la tesis.

Al Centro de Recursos Hídricos para la Agricultura y Minería (CRHIAM) a través del proyecto ANID/FONDAP/15130015, cuyo financiamiento fue de gran importancia para poder llevar a cabo la investigación.

A mis colegas del Departamento de Recursos Hídricos de la Universidad Nacional Agraria La Molina, Lima – Perú, Dr. Néstor Montalvo, Dr. Eduardo Chavarri, Dr. David Ascencios, Dr. Miguel Sánchez, Dra. Lía Ramos, Dra. Rocío Pastor, Mg.Sc. Ángel Becerra, Mg.Sc. Antonio Enciso, Mg.Sc. José Arapa y demás colegas.

A Clara Castro, Secretaria de Postgrado de la Universidad, por brindarme su apoyo durante mi permanencia como estudiante en el programa de Doctorado. A la Sra. Mirna Schafer una bellísima persona, por su cálida hospitalidad y amabilidad, me hizo sentir como en casa durante mi estadía en Chile.

RESUMEN

Uno de los fenómenos naturales más devastadores, capaces de causar una gran destrucción en muy poco tiempo son las inundaciones, producidas por las fuertes lluvias estacionales, que se caracterizan por su alta velocidad y poder destructivo. La precipitación es el principal aporte de recurso hídrico en una cuenca; pero al mismo tiempo es responsable de los eventos climático-meteorológicos extremos, como sequías o inundaciones. Además, la cantidad limitada de estaciones terrestres, la dificultad en su obtención, ausencia de registros continuos de longitud considerable, son factores limitantes que asocian la incertidumbre a la variación temporal y espacial de la precipitación. Por lo tanto, para abordar la estimación de precipitaciones en lugares donde no se dispone de series de datos confiables, se considera como alternativa la interpolación espacial. La interpolación espacial es una técnica eficiente que permite generar precipitaciones grilladas, insumo de datos de entrada en los modelos hidrológicos distribuidos basados físicamente.

El objetivo del estudio fue evaluar la mejora de la modelación hidrológica con datos grillados mediante la aplicación del modelo hidrológico HEC-HMS en cuenca interandina. Se plantearon tres objetivos específicos. Primero (Capitulo II), analizar la calidad y relleno de serie de precipitación diaria, esto se realizó mediante gráfica de serie de tiempo, diagrama de caja y prueba de homogeneidad; y para la completación de precipitaciones faltante se utilizaron técnica de regresión y machine learning. Segundo (Capitulo III), analizar el uso de parámetros morfométricos para el mapeo de áreas vulnerables a inundaciones, mediante la caracterización de los parámetros morfométricos se identificaron zonas vulnerables a inundaciones. Tercero (Capitulo IV), evaluar el desempeño de un modelo hidrológico utilizando datos grillados, se utilizaron como insumos de entrada información grillada de precipitación, topografía y numero de curva al modelo hidrológico. Capítulo V, se presentan las conclusiones generales, hipótesis y futuras investigaciones.

Además, siguiendo la metodología desarrollada en la investigación, se menciona de gran importancia la evaluación la calidad de los datos, antes de ser usada como insumo en la generación de precipitación grillada extremas. Finalmente, esta investigación desarrolló una metodología que permitió mejorar la modelación hidrológica a partir de datos grillados.

ABSTRACT

One of the most devastating natural phenomena, capable of causing great destruction in a very short time, are floods, produced by heavy seasonal rains, which are characterized by their high speed and destructive power. Precipitation is the main source of water resources in a basin; but at the same time it is responsible for extreme climatic-meteorological events, such as droughts or floods. In addition, the limited number of ground stations, the difficulty in obtaining them, the absence of continuous records of considerable length, are limiting factors that associate uncertainty to the temporal and spatial variation of precipitation. Therefore, to address the estimation of precipitation in places where reliable data series are not available, spatial interpolation is considered as an alternative. Spatial interpolation is an efficient technique to generate gridded precipitation, input data for physically based distributed hydrological models.

The objective of the study was to evaluate the improvement of hydrological modeling with gridded data through the application of the HEC-HMS hydrological model in the inter-Andean basin. Three specific objectives were set. First (Chapter II), to analyze the quality and filling of daily precipitation series, this was done by means of time series plot, box plot and homogeneity test; and for the completion of missing precipitation, regression and machine learning techniques were used. Second (Chapter III), analyze the use of morphometric parameters for the mapping of areas vulnerable to flooding, through the characterization of morphometric parameters were identified areas vulnerable to flooding. Third (Chapter IV), to evaluate the performance of a hydrological model using gridded data, gridded information on precipitation, topography and curve number were used as input to the hydrological model. Chapter V presents the general conclusions, hypotheses and future research.

In addition, following the methodology developed in the research, the evaluation of the quality of the data, before being used as input in the generation of extreme gridded precipitation, is mentioned as being of great importance. Finally, this research developed a methodology that allowed improving the hydrological modeling from gridded data.

INDICES

CAPITULO I: Introducción

1.1. Introducción

El cambio climático es una de las mayores amenazas para el mundo entero que afecta el equilibrio natural de la tierra y los ecosistemas (Teng et al., 2018). Este fenómeno se presenta más a menudo, mostrando cambios en las variables climáticas primarias, como la temperatura y la precipitación (Wang et al., 2021, Pérez et al., 2021, Kropp, 2015). La precipitación es el principal aporte de recurso hídrico en una cuenca; pero al mismo tiempo es responsable de los eventos climático-meteorológicos extremos, como sequías o inundaciones. La predicción y el control de estos eventos han sido muy difíciles, debido a la naturaleza altamente dinámica del clima y a su repentina aparición (Cahyono and Adidarma, 2019). Por lo tanto, es necesario realizar un estudio detallado para mejorar la precisión de dichos análisis (Chen et al., 2017).

Sin embargo, la cantidad limitada de estaciones terrestres, la escasez de información pluviométrica, la dificultad en su obtención, ausencia de registros continuos de longitud considerable, son factores limitantes que asocian la incertidumbre a la variación temporal y espacial de la precipitación, produciendo unas de las principales fuentes de error al ser aplicado en modelos precipitación - escorrentía (Pérez et al., 2021, Adhikary et al., 2017). Además, en las últimas décadas con el desarrollo de tecnología satelital, se dispone de numerosos productos grillados (Belayneh et al., 2020, Cho, 2020, Chen et al., 2019, Timmermans et al., 2019), que tienen la ventaja de presentar cobertura continua, pero introducen dudas porque la precipitación no se mide directamente sino que se infiere de las mediciones de microondas y radar infrarrojo, agregando así otra capa de procesamiento de datos que estaría sujeta a incertidumbre y error (Benkirane et al., 2023, Su et al., 2021, Timmermans et al., 2019). Por lo tanto, para abordar la estimación de precipitaciones en lugares donde no se dispone de series de datos confiables, se considera como alternativa la técnica de interpolación espacial, la cual permite conocer la distribución espacial de las precipitaciones (Ali et al., 2021, Das, 2021).

Para la generación de la distribución espacial de la precipitación, existen diferentes métodos de interpolación espacial a partir de estaciones pluviométricas, situadas dentro o cercas a una zona de interes (Salhi, 2022, Zou et al., 2021, Amini et al., 2019, Ma et al., 2018, Adhikary et al., 2017). La elección del método de interpolación es crucial en la generación de isoyetas, debido que las precipitaciones extremas suelen tener una gran variabilidad espacial, especialmente en

duraciones cortas (Zou et al., 2021, Ma et al., 2018). Muchos estudios han aplicado método geoestadístico como una alternativa de interpolación espacial, esto debido a las altas autocorrelaciones espaciales entre los puntos muestreados y predicciones insesgadas con varianza mínima (Ali et al., 2021, Amini et al., 2019, Adhikary et al., 2017, Chen et al., 2017). Por lo tanto, mediante el uso de la técnica geoestadística kriging ordinario, se generó la precipitación grillada para periodos de retornos (Tr) 5, 10, 20, 50 y 100 años.

Por otro lado, los modelos hidrológicos se han convertido en una herramienta indispensable para comprender los procesos hidrológicos y analizar sus cambios en entornos cambiantes (Zhang et al., 2023, Dwarakish and Ganasri, 2015). El uso de modelos hidrológicos para la predicción está motivado principalmente por la escasez de datos hidrológicos (Janicka et al., 2023). Además, el proceso de modelado hidrológico incluye, preparación de datos de entrada a los modelos, construcción de modelos hidrológicos y validación de estos modelos para simular procesos hidrológicos realistas (Zhang et al., 2023). Los modelos hidrológicos pueden ser clasificados según la naturaleza de representar los procesos (empírico, conceptuales y de base física), según la forma en que se consideren las variables hidrológicas (deterministas o estocásticos), según la forma en que se represente el medio físico (agregados, semidistribuidos y distribuidos) y según la naturaleza de los periodos de tiempo (basados en eventos o continua) (Al Khoury et al., 2023, Devi et al., 2015, Dwarakish and Ganasri, 2015, Picouet, 2014).

Además, para seleccionar el tipo y método de modelización suele depender de la finalidad, disponibilidad de datos y facilidad de uso antes de comenzar la simulación, pero el reto es elegir un modelo que pueda simular con precisión el proceso hidrológico en distintas condiciones climáticas y con los datos disponibles (Al Khoury et al., 2023, Hussain et al., 2021). Así mismo, para la implementación de un modelo hidrológico es primordial considerar las características espacio-temporales de una cuenca, útil para predecir con precisión la respuesta hidrológica y entender los procesos hidrológico en una cuenca (Al Khoury et al., 2023, Dwarakish and Ganasri, 2015). Además, en estudios anteriores sobre predicción de inundaciones mediante modelos hidrológicos a escala de cuenca, el uso de productos grillados, datos de entrada al modelo son escasas (Natarajan and Radhakrishnan, 2021, Jabbar et al., 2021, Abdessamed and Abderrazak, 2019, Natarajan and Radhakrishnan, 2019, Timmermans et al., 2019). Por lo tanto, en esta investigación para datos de entrada al modelo hidrológico, se utilizaron precipitación grillada, modelo digital de elevación (topografía) y curva número grillado. El objetivo del

estudio fue evaluar la mejora de la modelación hidrológica con datos grillados para Tr 5, 10, 20, 50 y 100 años, utilizando HEC-HMS, en la Cuenca del rio Cañete. El HEC-HMS es un programa de modelización hidrológica del tipo determinístico semidistribuido de base física y basado en eventos/continuo (Benkirane et al., 2023, Hussain et al., 2021, Castro and Maidment, 2020, Dwarakish and Ganasri, 2015). El rendimiento del modelo hidrológico fue evaluado mediante métricas estadísticas, para simulaciones de datos de entrada con precipitación grillada y precipitación desde estaciones terrestres. En general, esta investigación concluyo que la simulación basada en precipitación grillada superó a la basada en precipitación desde estaciones terrestre.

La tesis consta de cinco capítulos. Capítulo 1 describe la introducción general de la investigación, algunos conceptos teóricos utilizados, hipótesis y objetivos. Los capítulos 2, 3 y 4 se desarrollaron los tres objetivos específicos. Capítulo 2 se abarco el desarrollo del primero objetivo específico, donde se analizaron la calidad y relleno de serie de precipitación diaria, aplicando regresión y machine learning para el relleno de series de precipitaciones diarias faltante. Capítulo 3 abarca el desarrollo del segundo objetivo específico, donde se analizaron el uso de parámetros morfométricos para el mapeo de áreas vulnerables a inundaciones repentinas, aplicando criterio de ponderación. Además, los resultados de los capítulos 2 y 3 fueron insumo para la escritura de dos artículos de investigación, que están publicadas en revista indexada en la base Web of Science y Scopus. Capítulo 4 plamo el desarrollo del tercer objetivo específico, evaluándose el desempeño de un modelo hidrológico utilizando datos de precipitación grillada generada mediante técnica geoestadística kriging ordinario. Estos resultados permitieron escribir un tercer artículo, la cual está en la fase de traducción y próximamente enviada a una revista indexada de la base Web of Science y/o Scopus. Finalmente, Capítulo 5 presenta las principales conclusiones generales y futuras líneas de investigación.

1.2. Fundamentos Teóricos

1.2.1. Modelos Hidrológicos

Un modelo puede definirse, según Anderson et al.,(2015) como una representación simplificada de un sistema real complejo llamado prototipo, bajo forma física o matemática. En un modelo hidrológico, el sistema físico real que generalmente se representa es la cuenca hidrográfica y cada uno de los componentes del ciclo hidrológico (Anderson et al., 2015, Jajarmizadeh et al.,

2012). De esta manera un modelo matemático ayuda a tomar decisiones en materia de hidrología, por lo que es necesario tener conocimiento de entradas y salidas del sistema para verificar si el modelo es representativo del prototipo (Khan et al., 2019).

El mejor modelo es el que ofrece resultados cercanos a la realidad con el menor número de parámetros y complejidad (Devi et al., 2015). La utilidad de los modelos radica, entre otros aspectos, en la predicción de fenómenos a largo plazo en un tiempo relativamente corto, también permiten obtener relaciones de causa-efecto, sin haber realizado cambios en los sistemas reales (Jajarmizadeh et al., 2012). En la actualidad, los modelos de simulación hidrológica son herramientas utilizadas para la planificación del uso del suelo y ordenamiento territorial en cuencas hidrográficas, permitiendo analizar su respuesta a diferentes alternativas de manejo.

1.2.2. Principales tipos de modelos hidrológicos

Existe un gran número de modelos hidrológicos que varían en cuanto a su naturaleza y complejidad. Por lo tanto, es difícil clasificarlos sin ambigüedad, dada la variedad de criterios existente, se menciona los criterios de clasificación (Al Khoury et al., 2023, Devi et al., 2015, Dwarakish and Ganasri, 2015, Hingray et al., 2014):

Según la naturaleza de las relaciones utilizadas para representar los procesos, los modelos pueden clasificarse en empíricos, conceptuales o de base física.

- Los modelos empíricos se basan en las relaciones observadas entre las entradas y salidas del hidrosistema considerado. Expresan la relación entre las variables de entrada y salida del sistema (por ejemplo, la relación lluvia-escorrentía) utilizando un conjunto de ecuaciones desarrolladas y ajustadas a partir de los datos obtenidos para el sistema. Un modelo empírico no está diseñado para describir las causas del fenómeno hidrológico considerado ni para explicar el funcionamiento del hidrosistema. El hidrosistema se considera una caja negra.
- Los modelos hidrológicos conceptuales están concebidos para representar los principales procesos hidrológicos de forma razonable sin necesidad de parametrizar las leyes físicas que los rigen. La representación es conceptual en el sentido de que se basa en la percepción que tiene el hidrólogo del comportamiento hidrológico de la cuenca de drenaje. Esta percepción se deriva de la experiencia del hidrólogo y de su comprensión teórica, empírica y/o intuitiva del funcionamiento del hidrosistema estudiado.

- Los modelos de base física representan el funcionamiento hidrológico del hidrosistema mediante el acoplamiento de distintos submodelos, cada uno dedicado a determinados procesos hidrológicos. En general, se basan en una discretización espacial precisa del medio físico. Independientemente de la naturaleza y la resolución espacial de esta discretización, hay que tener en cuenta que la variabilidad espacial del medio físico y de los procesos sólo puede describirse explícitamente para escalas espaciales mayores a la discretización. Para escalas menores, debe describirse de manera conceptual.

Según la forma en que se consideren las variables hidrológicas o las relaciones entre estas variables, los modelos pueden clasificarse en deterministas o estocásticos.

- Los modelos determinísticos estipulan que, para unas condiciones iniciales y unas condiciones de contorno dadas, la relación entre las entradas y las salidas del sistema considerado son inequívocas. En otras palabras, para una entrada dada del modelo, existe una y sólo una salida.
- Los modelos estocásticos simulan procesos que son parcial o totalmente aleatorios. Para un conjunto dado de condiciones iniciales y de contorno, diferentes aplicaciones del modelo estocástico dan, para una entrada dada, diferentes salidas.

Según la forma en que se represente el medio físico y, en particular, según la naturaleza de la unidad espacial en la que se resuelven las ecuaciones utilizadas para describir los procesos, los modelos pueden clasificarse en agregados, semidistribuidos y distribuidos.

- Los modelos hidrológicos agregados (también llamado agrupado o lumped) describen la cuenca de drenaje como una única unidad hidrológica. Por lo tanto, no se tiene en cuenta la variabilidad espacial de los procesos hidrológicos considerados para describir su comportamiento.
- Los modelos semidistribuidos dividen la cuenca en unidades de subcuenca más pequeñas. Cada subcuenca tiene un conjunto independiente de valores de los parámetros del modelo. De ahí que estos tipos de modelos incluyan la variabilidad espacial de los parámetros y ofrezcan mejores resultados que los modelos agrupados.
- Los modelos distribuidos, la cuenca se discretiza de forma más detallada en celdas o en una malla regular o irregular. Estos tipos de modelos también incluyen la variabilidad espacial de los parámetros.

Dependiendo de la naturaleza de los periodos de tiempo considerados para la simulación, los modelos pueden clasificarse como modelos basados en eventos o modelos de simulación continua.

- Los modelos basados en eventos se utilizan para simular eventos hidrológicos seleccionados sin tener en cuenta los periodos entre eventos. Se desarrollan principalmente para simular la transformación de las precipitaciones en escorrentía para la previsión o predicción de la descarga fluvial. El uso de un modelo basado en eventos requiere la estimación de las condiciones iniciales de la simulación para cada evento considerado.
- Los modelos de simulación continua se utilizan para simular de forma continua el comportamiento hidrológico de la cuenca. La simulación puede realizarse durante largos periodos de tiempo, abarcando una variedad de situaciones hidrometeorológicas que van desde caudales bajos hasta descargas de crecidas. Estos modelos deben considerar todos los procesos que influyen significativamente en la respuesta de la cuenca de drenaje.

La figura 1 muestra la clasificación de modelos hidrológicos.

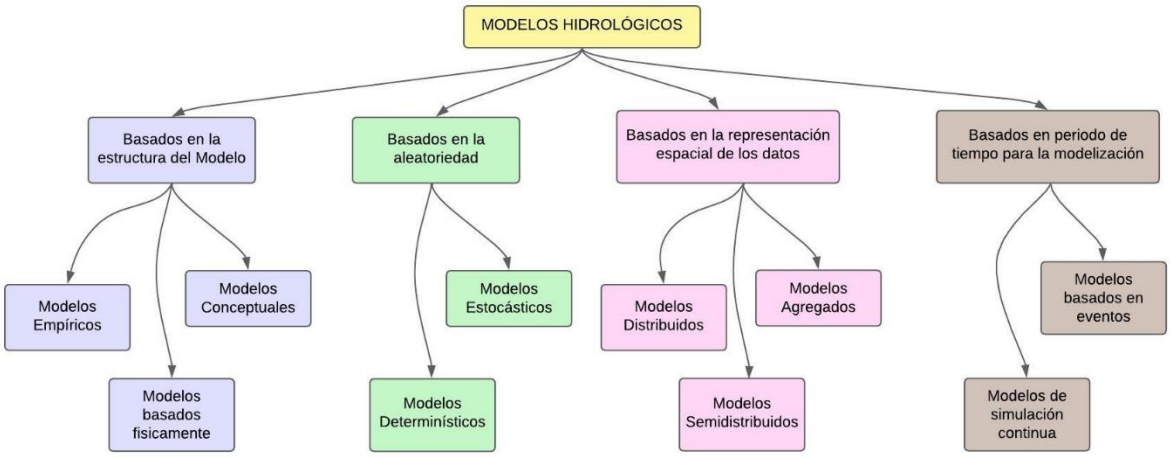


Figura 1: Clasificación de modelos hidrológicos

1.2.3. Modelo hidrológico HEC-HMS

El modelo HEC-HMS (Hydrologic Engineering Center's Hydrologic Modeling System) es un programa capaz de simular la precipitación-escorrentía (https://www.hec.usace.army.mil/software/hec-hms/). Está diseñado para simular los procesos de precipitación-escorrentía en una amplia gama de zonas geográficas, como inundaciones,

abastecimiento de agua, grandes cuencas fluviales o pequeñas escorrentías urbanas y naturales (Jabbar et al., 2021, Castro and Maidment, 2020). El HEC-HMS es uno de los modelos hidrológicos más utilizados y documentados, y en él se han realizado varios estudios diferentes (Castro and Maidment, 2020, Barbosa et al., 2019, Abdessamed and Abderrazak, 2019).

HEC-HMS es un modelo determinístico, semi-distribuido, basado en eventos/continuo y de base física (Benkirane et al., 2023, Hussain et al., 2021, Castro and Maidment, 2020). Realiza fácilmente una amplia gama de funciones de estudio hidrológico, como pérdidas, transformación de descargas, enrutamiento en canales abiertos y análisis de datos meteorológicos, simulación de precipitación y escorrentía, y estimación de parámetros (Benkirane et al., 2023). El modelo consta de cuatro módulos principales: modelo cuenca, modelo meteorológico, especificaciones de control y datos de entrada (series temporales, datos emparejados y datos grillados) (Natarajan and Radhakrishnan, 2021). Para simular las pérdidas por infiltración se utilizan diferentes métodos como Déficit y Constante, Exponencial, Inicial y Constante, SCS Curva Numero, y datos grillados (Déficit y Constante y SCS Curva Numero). Para transformar el exceso de precipitación en escorrentía superficial, se utilizan métodos como el hidrograma unitario Clark, onda cinemática, ModClark, hidrograma unitario SCS, hidrograma unitario Snyder, grafica S especificado por el usuario, hidrograma hidrograma unitario especificado por el usuario y Onda difusa 2D. Para transito hidrológico en cauce se utilizan los modelos como Onda Kinematic, Lag, Modified Plus, Muskingum, Muskingum-Cunge, Normal Depth y Straddle Stagger (Abdessamed and Abderrazak, 2019, Natarajan and Radhakrishnan, 2019).

1.3. Aporte al conocimiento

En muchas investigaciones de predicción de inundaciones mediante modelos hidrológicos a nivel de cuenca, los datos de entrada al modelo, como precipitación, temperatura, numero de curva, etc., generalmente son obtenidos como datos puntuales (estaciones meteorológicas). Así mismo, la cantidad limitada de estaciones terrestres, la escasez de información pluviométrica, la dificultad en su obtención, ausencia de registros continuos de longitud considerable, son factores limitantes que asocian la incertidumbre de la variación temporal y espacial de los datos de entrada al modelo. Además, mediante el análisis de parámetros morfométricos conocer el comportamiento hidrológico a nivel de cuenca y generar la distribución espacial de la precipitación en base a datos puntuales utilizando métodos geoestadísticos.

Los procedimientos metodológicos seguido en esta investigación, pueda brindar herramienta de aporte al conocimiento en el campo de la inteligencia artificial (*machine learning*) aplicado a la hidrología. Además, mapeo de áreas vulnerables a inundaciones repentinas en base a parámetros morfométricos, fácilmente de obtener utilizando herramienta SIG y finalmente, predecir con precisión las inundaciones con el uso de información grillada mediante modelación hidrológica. La metodología y las herramientas utilizadas en esta investigación ayudaran a entender la variabilidad espacial y temporal de las precipitaciones, los procesos hidrológicos distribuido espacialmente y la respuesta hidrológica de eventos extremos.

1.4. Hipótesis

El uso de información grillada pluviométrica, topográfica y curva número, permite predecir con precisión las inundaciones mediante el uso de modelo hidrológico.

1.5. Objetivos

Objetivo general

- Analizar el uso de información grillada pluviométrica, topográfica y curva número, para la predicción de inundaciones mediante modelo hidrológico en una cuenca interandina.

Objetivo especifico

- Analizar la calidad y relleno de serie de precipitación diaria
- Analizar el uso de parámetros morfométricos para el mapeo de áreas vulnerables a inundaciones
- Evaluar el desempeño de un modelo hidrológico utilizando datos grillados.

Referencias

Abdessamed, D. and Abderrazak, B. (2019). Coupling HEC-RAS and HEC-HMS in rainfall-runoff modeling and evaluating floodplain inundation maps in arid environments: case study of Ain Sefra city, Ksour Mountain. SW of Algeria. Environmental Earth Sciences, 78(19):586.

Adhikary, S. K., Muttil, N., and Yilmaz, A. G. (2017). Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments. Hydrological processes, 31(12):2143–2161.

Al Khoury, I., Boithias, L., and Labat, D. (2023). A review of the application of the soil and Water Assessment Tool (SWAT) in Karst Watersheds. Water, 15(5).

Ali, G., Sajjad, M., Kanwal, S., Xiao, T., Khalid, S., Shoaib, F., and Gul, H. N. (2021). Spatial–temporal characterization of rainfall in Pakistan during the past half-century (1961–2020). Scientific reports, 11(1):1–15.

Amini, M. A., Torkan, G., Eslamian, S., Zareian, M. J., and Adamowski, J. F. (2019). Analysis of deterministic and geostatistical interpolation techniques for mapping meteorological variables at large watershed scales. Acta Geophysica, 67(1):191–203.

Anderson, M. P., Woessner, W. W., and Hunt, R. J. (2015). Applied Groundwater Modeling, Second Edition: Simulation of Flow and Advective Transport. Academic Press, second edition edition.

Barbosa, J., Fernandes, A., Lima, A., and Assis, L. (2019). The influence of spatial discretization on HEC-HMS modelling: a case study. Int. J. Hydrol, 3:442–449.

Belayneh, A., Sintayehu, G., Gedam, K., and Muluken, T. (2020). Evaluation of satellite precipitation products using HEC-HMS model. Modeling Earth Systems and Environment, 6(4):2015–2032.

Benkirane, M., Amazirh, A., Laftouhi, N.-E., Khabba, S., and Chehbouni, A. (2023). Assessment of GPM satellite precipitation performance after bias correction, for hydrological modeling in a semi-arid watershed (high atlas mountain, morocco).

Atmosphere, 14(5).

Cahyono, C. and Adidarma, W. K. (2019). Influence analysis of peak rate factor in the flood eventsâ€™ calibration process using HEC-HMS. Modeling Earth Systems and Environment, 5(4):1705–1722.

Castro, C. V. and Maidment, D. R. (2020). Gis preprocessing for rapid initialization of HEC-HMS hydrological basin models using web-based data services. Environmental Modelling & Software, 130:104732.

Chen, F., Gao, Y., Wang, Y., Qin, F., and Li, X. (2019). Downscaling satellite-derived daily precipitation products with an integrated framework. International Journal of Climatology, 39(3):1287–1304.

Chen, T., Ren, L., Yuan, F., Yang, X., Jiang, S., Tang, T., Liu, Y., Zhao, C., and Zhang, L. (2017). Comparison of Spatial Interpolation Schemes for Rainfall Data and Application in Hydrological Modeling. Water, 9(5).

Cho, Y. (2020). Application of NEXRAD radar-based quantitative precipitation estimations for hydrologic simulation using ArcPy and HEC Software. Water, 12(1).

Das, S. (2021). Extreme rainfall estimation at ungauged locations: Information that needs to be included in low-lying monsoon climate regions like Bangladesh. Journal of Hydrology, 601:126616.

Devi, G. K., Ganasri, B., and Dwarakish, G. (2015). A Review on Hydrological Models. Aquatic Procedia, 4:1001–1007. International Conference on Water Resources, Coastal and Ocean Engineering (ICWRCOE'15).

Dwarakish, G. and Ganasri, B. (2015). Impact of land use change on hydrological systems: A review of current modeling approaches. Cogent Geoscience, 1(1):1115691.

Hingray, B., Picouet, C., and Musy, A. (2014). Hydrology : A Science for Engineers. Taylor and Francis.

Hussain, F., Wu, R.-S., and Yu, K.-C. (2021). Application of physically based semi-distributed HEC-HMS model for flow simulation in tributary catchments of kaohsiung area Taiwan. Journal of Marine Science and Technology, 29(1):4.

Jabbar, L. A., Khalil, I. A., and Sidek, L. M. (2021). HEC-HMS hydrological modelling for runoff estimation in Cameron Highlands, Malaysia. International Journal of Civil Engineering and Technology, 12(9):40–51.

Jajarmizadeh, M., Harun, S., and Salarpour, M. (2012). A review on theoretical consideration and types of models in hydrology. Journal of Environmental Science and Technology, 5(5):249–261.

Janicka, E., Kanclerz, J., Agaj, T., and Gizinska, K. (2023). Comparison of Two Hydrological Models, the HEC-HMS and Nash Models, for Runoff Estimation in Michalowka River. Sustainability, 15(10).

Khan, S. I., Flamig, Z., and Hong, Y. (2019). Flood Monitoring System Using Distributed Hydrologic Modeling for Indus River Basin. In Khan, S. I. and Adams, T. E., editors, Indus River Basin, pages 335–355. Elsevier.

Kropp, S. (2015). Climate Change and Risk of Flooding in Germany. Research Collection, page 155.

Ma, L., Zhang, G., and Lu, E. (2018). Using the Gradient Boosting Decision Tree to Improve the Delineation of Hourly Rain Areas during the Summer from Advanced Himawari Imager Data. Journal of Hydrometeorology, 19(5):761 – 776.

Natarajan, S. and Radhakrishnan, N. (2019). Simulation of extreme event-based rainfall-runoff process of an urban catchment area using HEC-HMS. Modeling Earth Systems and Environment, 5(4):1867–1881.

Natarajan, S. and Radhakrishnan, N. (2021). Simulation of rainfall-runoff process for an ungauged catchment using an event-based hydrologic model: A case study of koraiyar basin in Tiruchirappalli city, India. Journal of Earth System Science, 130(1):30.

Pérez, R. E., Cortés-Molina, M., and Navarro-González, F. J. (2021). Analysis of Rainfall Time Series with Application to Calculation of Return Periods. Sustainability, 13(14):8051.

Salhi, H. (2022). Evaluation of the Spatial Distribution of the Annual Extreme Precipitation Using Kriging and Co-Kriging Methods in Algeria Country. In Tiefenbacher, J. P., editor, Climate Change in Asia and Africa, chapter 4. IntechOpen, Rijeka.

Su, J., Li, X., Ren, W., Lu, H., and Zheng, D. (2021). How reliable are the satellite-based precipitation estimations in guiding hydrological modelling in South China? Journal of Hydrology, 602:126705.

Teng, F., Huang, W., and Ginis, I. (2018). Hydrological modeling of storm runoff and snowmelt in Taunton River Basin by applications of HEC-HMS and PRMS models. Natural Hazards, 91(1):179–199.

Timmermans, B., Wehner, M., Cooley, D., Oâ€™Brien, T., and Krishnan, H. (2019). An evaluation of the consistency of extremes in gridded precipitation data sets. Climate Dynamics, 52(11):6651–6670.

Wang, N., Lombardo, L., Gariano, S. L., Cheng, W., Liu, C., Xiong, J., and Wang, R. (2021). Using satellite rainfall products to assess the triggering conditions for hydro-morphological processes in different geomorphological settings in China. International Journal of Applied Earth Observation and Geoinformation, 102:102350.

Zhang, S., Lang, Y., Yang, F., Qiao, X., Li, X., Gu, Y., Yi, Q., Luo, L., and Duan, Q. (2023). Hydrological Modeling in the Upper Lancang-Mekong River Basin Using Global and Regional Gridded Meteorological Re-Analyses. Water, 15(12).

Zou, W.-Y., Yin, S.-Q., and Wang, W.-T. (2021). Spatial interpolation of the extreme hourly precipitation at different return levels in the haihe river basin. Journal of Hydrology, 598:126273.

CAPITULO II: Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds

Marcelo Portuguez-Maurtua, José Luis Arumi, Octavio Lagos, Alejandra Stehr, and Néstor Montalvo Arquiñigo. (2022).

2.1 Resultado clave

- Se realizó el análisis de datos de serie de precipitación diaria faltantes, evaluación de valores atípicos y detección de homogeneidades
- Proceso de agrupamiento de estaciones pluviométricas de comportamientos similares mediante el análisis de regionalización, identificándose cuatro grupos de regiones.
- Aplicación de métodos Regresión y Machine Learning en el rellenado de valores faltantes de series de precipitación diaria, los resultados mostraron que el modelo machine learning optimizado presentaron la menor variabilidad en los errores de estimación y la mejor aproximación a los datos reales.

2.2 Resumen en extenso

La precipitación es un componente fundamental del ciclo hidrológico global que gobierna la distribución de los recursos, el entendimiento de su comportamiento temporal y espacial es de gran interés, y las estimaciones exactas de estas son fundamentales en múltiples líneas de investigación. La precipitación, debido a su alta variabilidad espacio-temporal y al gran número de variables interconectadas que intervienen, es una de las variables atmosféricas más difíciles de caracterizar, estimar y predecir, especialmente a escala diaria, debido a su alta variabilidad espacial y temporal.

La distribución espacial de las estaciones pluviométrica en regiones en vías de desarrollo, generalmente se localizan de manera heterogénea. Además, es frecuente encontrar series de precipitación incompletas lo que dificulta la caracterización hidrológica o climatológica de un determinado lugar. Para el relleno de series de precipitaciones, se conocen gran cantidad de métodos, como regresión por mínimos cuadrados, promedios aritméticos, etc. Sin embargo, investigaciones sobre relleno de serie de precipitación faltante, han sido abordado principalmente a escala anual y mensual, con pocas investigaciones centradas en la escala diaria,

debido a la complejidad de las características orográficas y, en algunos casos, a la no linealidad de las series de precipitación entre estaciones vecinas.

El objetivo de esta investigación fue evaluar diferentes modelos de relleno de datos de precipitación diaria faltante, mediante las técnicas de Modelo de Regresión (MR) y Machine Learning (ML). Se consideró para MR los algoritmos de regresión lineal (LRM) y múltiples (MRM), y para ML se consideró los algoritmos de regresión múltiples (ML-MRM), K-Nearest neighbor (ML-KNN), Gradient boosting trees (ML-GBT) y Random Forest (ML-RF). Además, se utilizó un proceso de optimización, Machine Learning Optimizado (MLO), con los modelos de regresión múltiples (MLO-MRM), K-Nearest neighbor (MLO-KNN), Gradient boosting trees (MLO-GBT) y Random forest (MLO-RF).

Para evaluar el rendimiento de los modelos utilizados, se aplicaron diferentes métricas estadísticas. Los resultados mostraron que los modelos de MLO presentaron la menor variabilidad en los errores de estimación y la mejor aproximación a los datos reales de la zona de estudio. Además, esta investigación demostró que el modelo ML interpreta las relaciones no lineales entre los pluviómetros a escala diaria, presentándose como un método eficiente de relleno de series de precipitación faltante a escala diaria.

water

MDPI

Article

Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds

Marcelo Portuguez-Maurtua [1,2,3]*, José Luis Arumi [2,4], Octavio Lagos [2,4], Alejandra Stehr [5] and NestorMontalvo Arquiñigo [3]

[1] Doctoral Program in Water Resources and Energy for Agriculture, Universidad de Concepcion, Av. Vicente Mendez 595, Chillan Chile; mportuguez@lamolina.edu.pe
[2] CRHIAM Water Research Center, Universidad de Concepcion, Victoria 1295, Concepcion Chile; jarumi@udec.cl
[3] Water Resources Department, College of Agricultural Engineering, Universidad Nacional Agraria La Molina, Av. La Molina s/n, Lima, Peru; nmontalvo@lamolina.edu.pe
[4] Water Resources Department, College of Agriculture Engineering, Universidad de Concepción, Av. Vicente Mendez 595, Chillan Chile; octaviolagos@udec.cl
[5] Centro de Ciencias Ambientales EULA-Chile, Departamento de Sistemas Acuáticos, Facultad de Ciencias Ambientales, Universidad de Concepción, Concepción 4070386, Chile; astehr@udec.cl
* Correspondence: Correspondence: mportuguez@lamolina.edu.pe; Tel.: +51-1 949-377-610

Citation: Portuguez-Maurtua, M.; Arumi, J. L.; Lagos, O.; Stehr, A.; Montalvo Arquiñigo, N. Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds. *Water* **2022**, *14*, 1799. https://doi.org/10.3390/w14111799

Academic Editor: Zheng Duan and Scott Curtis

Received: 26 March 2022
Accepted: 27 May 2022
Published: 2 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: As precipitation is a fundamental component of the global hydrological cycle that governs water resource distribution, the understanding of its temporal and spatial behavior is of great interest, and exact estimates of it are crucial in multiple lines of research. Meteorological data are inputs for hydroclimatic models and predictions, which generally lack complete series. Many studies have addressed techniques to fill gaps in precipitation series at annual and monthly scales, but few have done so at a daily scale due to the complexity of orographic characteristics and in some cases the non-linearity of precipitation. The objective of this study was to assess different methods of filling gaps in daily precipitation data using regression model (RM) and machine learning (ML) techniques. RM included linear regression (LRM) and multiple regression (MRM) algorithms, while ML included multiple regression algorithms (ML-MRM), K-nearest neighbors (ML-KNN), gradient boosting trees (ML-GBT), and random forest (ML-RF). This study covered the Malas, Omas, and Cañete River (MOC) watersheds, located on the Pacific Slope of central Peru, and a nineteen-year period of records (2001-2019). To assess model performance, different statistical metrics were applied. The results showed that the optimized machine learning (OML) models presented the least variability in estimation errors and the best approximation of the actual data from the study zone. In addition, this investigation shows that ML interprets and analyzes non-linear relationships between rain gauges at a daily scale and can be used as an efficient method of filling gaps in daily precipitation series.

Keywords: Precipitation gap filling; Regression; Machine learning; Standard normal homogeneity test; K-nearest neighbors; Gradient boosting tree; Random forest

1. Introduction

Precipitation is a fundamental component of the hydrological cycle that governs water resource distribution [1]. The hydrological cycle of a given region is directly related to its topography, geology, physical mechanisms, and climate; precipitation is the most important phenomenon [2, 3]. Precipitation, due to its high spatio-temporal variability and the large number of interconnected variables involved, is one of the most difficult atmospheric variables to characterize, estimate and forecast [4], especially on a daily scale, due to its high spatial and temporal variability [5]. The understanding of the temporal and spatial behavior of precipitation is of great interest, especially in studies on climatic risks [6]. In addition, exact estimates of precipitation are fundamental in multiple lines of research, as they serve as the basis for statistical models and analysis [7, 8].

Peru is composed of three slopes, one that drains into the Pacific Ocean (Pacific Slope), another that drains into the Atlantic Ocean (Atlantic Slope), and a third that drains into Lake Titicaca (Titicaca Slope) [9]. The Peruvian Pacific Slope is located in tropical latitudes, and precipitation there is mainly influenced by orographic conditions, the ocean, and the atmosphere [10, 11]. The spatial distribution of rainfall stations in Peru is heterogenous. In addition, precipitation series are frequently incomplete, which complicates the hydrological or climatological characterization of a given place [12].

Recent studies on the Peruvian Pacific Slope and coast have allowed it to be classified into homogenous regions [11], which has aided the understanding of spatial and seasonable precipitation variability patterns [9, 13]. There are numerous methods of precipitation series gap filling, including least squares regression, predictive mean matching, nearest neighbor techniques, decision tree techniques, gradient boosting, and artificial neural networks [12, 14-17] In addition, geostatistical methods such as ordinary kriging tend to overestimate the number of rainy days and underestimate their magnitudes, and a negative correlation is even found in several reports between nearby stations [18-20]. Also, the authors Huang et al. [21] and Gorshenin et al. [22] have evaluated the k-neighest-neighbor algorithm, together with machine learning models, such as multilayer perceptron (MLP), support vector machine (SVM) and random forest (RF), with promising results. A study in Germany used machine learning (ML) techniques, analyzing non-linear relationships between spatially distributed rain gauges [12]. In addition, a recent study conducted by Bellido-Jiménez et al. [23] to fill possible gaps in precipitation datasets, in semi-arid regions, Andalusia, several machine learning models (MLP, SVM and RF) were tested showing good results using neighboring data with MLP.

However, studies on precipitation series gap filling have mainly addressed annual and monthly scales [24-26]. Similarly, there are other studies that have addressed regional-scale development techniques, merging estimates based on quantile mapping, spatial interpolation, machine learning, and multi-strategy fusion [27, 28], with few investigations focused on a daily scale, due to the complexity of orographic characteristics and in some cases non-linearity of precipitation series between neighboring stations [8, 29, 30].The objective of this study

was to fill gaps in daily precipitation series through comparative analysis of regression model (RM) and ML techniques. RM included linear (LRM) and multiple regression models (MRM). For ML, multiple regression models (ML-MRM), K-nearest neighbors (ML-KNN), gradient boosting trees (ML-GBT), and random forest (ML-RF) were used. In addition, an optimization process, optimized machine learning (OML), was used with the multiple regression (OML-MRM), K-nearest neighbors (OML-KNN), gradient boosting tree (OML-GBT), and random forest models (OML-RF), for a network of 17 rainfall stations located in the Malas, Omas, and Cañete River (MOC) watersheds. We assessed the efficiency of the results obtained from each model using statistical metrics. However, in order to guarantee reliable results using raw rainfall data, it is an essential requirement to perform the quality control process, such as the homogenization of the daily rainfall series, which allowed the detection of observation and measurement errors, problems that occur in a rainfall observation network [31]. In addition, identify homogeneous zones through the regionalization process, using up to three methods as a means of verifying the results.

The aim of this study was to demonstrate that ML techniques can interpret and analyze non-linear relationships between rain gauges at a daily scale and can be used as an efficient method of filling gaps in daily precipitation series. The results of the gap-filled precipitation series can be used in future investigations to evaluate the performance of the of the daily precipitation data obtained from satellite sensors based on a hydrological model and evaluate its performance based on time series of discharges measured at hydrometric stations. Finally, the results of this study showed that the ML models presented better approximations to the actual data than the RM models.

The structure of the paper is organized as follows. Section 2 shows the information about the locations, the dataset, the theoretical background of the different machine learning (ML) models evaluated, the preprocessing algorithms and evaluation metrics, in addition to the quality control of the dataset by homogenization and regionalization. Then, in Section 3 and 4, the results are reported and discussed, respectively. Finally, Section 5 describes the conclusions reached in this work.

2. Materials and Methods

2.1. Study Area

The study area comprised the Mala, Omas, and Cañete River (MOC) watersheds (Figure 1), located in the central part of the Peruvian Pacific Slope and coast; its total area is 9,496 km2 (2,250, 1,167 and 6,079 km2, MOC basins, respectively). The area is characterized by a significant latitudinal gradient that goes from 0 to 6,500 m.a.s.l.; above 2,500 m.a.s.l. is the wet watershed area [32]. The rivers flow from east to west from the Andes to the Pacific Ocean, with bare, steep slopes that favor significant swelling, floods, and erosion during heavy rainfall episodes [9]. In addition, in normal conditions, the region is influenced by the South Pacific High, in combination with the Humboldt current that produces dry, stable conditions with moist air trapped below the inversion layer at about 1,000 m.a.s.l., and presents major seasonal and interannual precipitation variability [9, 11, 13].

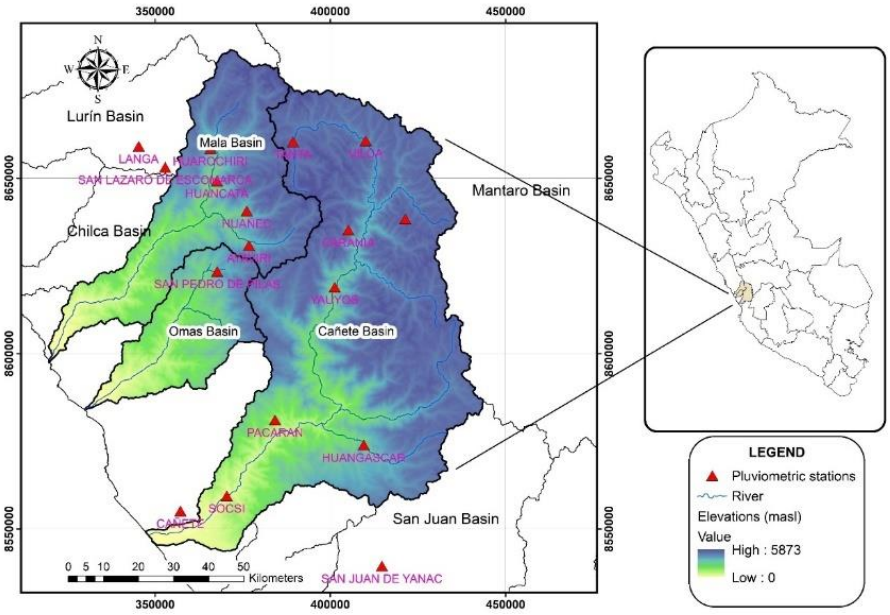


Figure 1. Elevation of the study area, main rivers, selected watershed boundaries, and location of rainfall stations.

2.2. Methods

The methodology has four stages; Figure 2 shows the methodological diagram. The first stage is the collection of available daily precipitation information from within and near the study area. The second stage is the exploratory analysis and homogenization of rainfall data. The third stage is the regionalization process, which includes the use of the Ward, K-means, and regional vector analysis methods (RVM). Finally, the fourth stage consists of the filling of gaps in daily precipitation series using the RM and ML methods. In addition, the performance of each model was evaluated using metrics.

2.2.1. Collection of available information

A total of 17 rainfall stations were selected, some with records since 1965, others since 1980, etc., all of which had periods with irregular records. The stations are part of the network managed by the National Meteorology and Hydrology Service of Peru (SENAMHI, https://www.senamhi.gob.pe/mapas/mapa-estaciones/mapadepesta1.php). In addition, stations located outside the study region and those inactive during the selected period were discarded. Similarly, there are rainfall stations with more than 10% missing (empty) data relative to the total length of the analyzed series. Figure 1 shows the spatial locations of the rainfall stations. In addition, Table 1 shows the geographic coordinates, quantity of observed data, and quantity of missing data.

2.2.2. Exploratory data analysis (EDA)

It is an essential requirement to guarantee reliable results using raw rainfall data, the application of quality control procedures, by means of graphs and the homogenization of time series, allowing the detection of observation and measurement errors, supported in recommended by Estévez et al. [31]. This process was carried out in two phases: first, a time series graph and boxplot, which allowed the identification of missing values and outliers, a process performed in Python. Second, in order to determine inconsistencies at the stations, which could stem from a change in instrument location, variations in the conditions at the measurement site, or an observer change, the data were analyzed using the standard normal homogeneity test (SNHT), described by [33-35].

The SNHT was developed by [36] and modified by [37, 38]; it uses Y to denote the candidate series and Y_i to denote a specific value (for example, cumulative annual precipitation or mean annual temperature) in the year (or other unit of time) i. In addition, X_j denotes one of the surrounding reference sites (j th of a total of k) and X_{ji} a specific value from that site. The following equations were used to detect the relative non-homogeneities (traditionally used in precipitation studies):

$$Q_i = \frac{Y_i}{\left\{\frac{\left[\sum_{j=1}^{k} \rho_j^2 X_{ji} \frac{\hat{Y}}{\overline{X_j}}\right]}{\sum_{j=1}^{k} \rho_j^2}\right\}} \quad (1)$$

and

$$Q_i = Y_i - \left\{\frac{\sum_{j=1}^{k} \rho_j^2 \left[X_{ji} - \widehat{X_j} + \hat{Y}\right]}{\sum_{j=1}^{k} \rho_j^2}\right\} \quad (2)$$

Where Q_i is the ratio in Equation (1) and the difference in Equation (2) in a specific year i; $\hat{Y}$ represents the multi-annual mean of the candidate time series; and ρ_j is the correlation coefficient between the test variable Y and the reference variable X_j [36, 38, 39]. This method is implemented in the Climatol package for R language [34]. Climatol has three normalization methods: division by mean values, subtraction of means, and complete standardization; here we opted for subtraction of means, as the minimum precipitation values can be zero [34, 40, 41]. On a preliminary basis Climatol was run for a monthly time step, identifying breaks; based on these breaks Climatol was run again for a daily time step. The results show graphs of absolute maximum autocorrelation (ACmx), SNHT, root mean square error (RMSE), and percentage of original data (POD).

2.2.3. Regionalization Process

This section describes the regionalization process, which was performed using three methods. In the first method Ward's hierarchical clustering analysis was applied. This method is also known as "minimum variance" grouping, where ward´s objective function, of the [42] algorithm minimizes the sum of squared deviations of the attribute vectors from the centroid of their respective groups; instead of merging samples or clusters as a function of distance, it starts by assigning "zero variance" to all clusters. This method was applied to ascertain the preliminary clustering

of the stations [43]. This process was carried out by programming in R language.

Table 1. Rainfall stations of MOC watersheds, 2001-2019 period.

ID	Stations	Coordinates		Altitude	Observed data		Missing data	
		Latitude	Longitude	(masl)	No of data	(%)	No of data	(%)
1	Ayaviri	-12.38	-76.13	3228	6881	99.2	58	0.8
2	Cañete	-13.07	-76.32	158	3830	55.2	3109	44.8
3	Carania	-12.34	-75.87	3875	6939	100	0	0
4	Huancata	-12.22	-76.22	2700	6939	100	0	0
5	Huangascar	-12.9	-75.83	2533	6908	99.6	31	0.4
6	Huañec	-12.29	-76.14	3205	6939	100	0	0
7	Huarochiri	-12.13	-76.23	3154	6787	97.8	152	2.2
8	Langa	-12.13	-76.42	2863	6484	93.4	455	6.6
9	Pacaran	-12.83	-76.07	700	5132	74	1807	26
10	San Juan de Yanac	-13.21	-75.79	2550	6482	93.4	457	6.6
11	San Lazaro de Escomarca	-12.18	-76.35	3758	6486	93.5	453	6.5
12	San Pedro de Pilas	-12.45	-76.22	2600	6909	99.6	30	0.4
13	Socsi	-13.03	-76.19	500	4687	67.5	2252	32.5
14	Tanta	-12.12	-76.02	4323	6819	98.3	120	1.7
15	Vilca	-12.11	-75.83	3864	6297	90.7	642	9.3
16	Yauricocha	-12.32	-75.72	4675	6818	98.3	121	1.7
17	Yauyos	-12.49	-75.91	2327	6878	99.1	61	0.9

In the second method non-hierarchical K-means clustering (KM) was applied, a statistical technique designed to assign objects to a fixed number of clusters according to a set of specified variables [11, 44]. It consists of obtaining a partition that minimizes intraclass inertia. This is achieved locally (it depends on the initial points) using the Euclidian distance between individuals and the moving centers used for aggregation. The KM algorithm is an iterative procedure in which the attribute vectors move from one group to another to minimize the value of the objective function, F, defined in Eq. (3).

$$F = \sum_{k=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{N_k} d^2\left(y_{ij}^k - y_{\bullet j}^k\right) \tag{3}$$

In Eq. (3), k indicates the number of groups, N_k represents the number of attribute vectors in group k; $y^k{}_{ij}$ denotes the rescaled value of attribute j in attribute vector i assigned to group k; and $y^k{}_{\bullet j}$ is the mean value of attribute j for group k (Eq. (4))[43, 45].

$$y_{\bullet j}^k = \frac{\sum_{i=1}^{N_k} y_{ij}^k}{N_K} \tag{4}$$

However, one of the problems encountered when applying the KM method lies inchoosing the number of clusters. Although there is no single

criterion for choosing the number of clusters, here we used the elbow method, implementing it by programming in R language.

Finally, the regional vector method (RVM), described by [10, 11, 44], was the third to be applied, in order to corroborate the previously obtained results. It consists of creating a fictitious station (vector), with average values from all stations in the zone. This method is aimed at the homogenization and completion-extension of precipitation data [46, 47] and is based on the creation of an "average value" "vector" station. This concept refers to the calculation of a weighted average of rainfall anomalies for each station, overcoming the effects of stations with extreme and low rainfall values and problems associated with the weight of the rainiest stations relative to the least rainy ones.

This method applies the least squares method to find annual regional rainfall indices Z_i and extended mean precipitation P_j, which is achieved by minimizing the expression [10, 11, 45]:

$$S = \sum_{I=1}^{N} \sum_{J=1}^{M} \left(\frac{P_{ij}}{P_j} - Z_i \right)^2 \quad (5)$$

Where: i is the index of the year; j is the index of the station; N is the number of years; M is the number of stations; P_{ij} is annual precipitation at station j in year i; P_j is mean precipitation extended to a period of N years; and, finally, Z_i is the regional rainfall index of year i. This process was carried out using the Hydracces program [48].

2.2.4. Gap-filling model

In this stage of the study, the results from the regionalization process were used. The RM and ML techniques were applied for each homogenous region. The daily precipitation series were graphed for each homogenous region, allowing the dates with missing data to be identified. In addition, the intensity of the relationships between stations was analyzed using Pearson coefficient correlations [29, 30].

To apply the LRM and MRM techniques, in both cases target stations (Y) and variables to predict were identified. Predictor stations (X) were identified for LRM and multiple predictor stations (X_m) for MRM. LRM is a computing procedure based on the alternate least squares algorithm (ALS) [49]. It has two steps: first, estimating the relationship between predictors and missing values, and then using the trend equation to fill the gaps [50], in accordance with Eq. (6):

$$P_i(t) = a + b.P_i(t) \quad (6)$$

The values of a and b can be estimated using Equations (7) and (8), respectively.

$$a = \bar{y} - b\bar{x} \quad (7)$$

$$b = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)\, x^2}{n}} \quad (8)$$

Where $\bar{y}$ and $\bar{x}$ are mean values of the data series of the reference and similarity sta- tions, respectively [50, 51].

Meanwhile, MRM is a statistical technique that consists of finding a linear relationship between a dependent variable and more than one independent variable. It can be represented using the following equation:

$$Y_i = a + b_1X_1 + b_2X_2 + \cdots + b_mX_m + C \quad (9)$$

Where Y_i is the dependent variable; X_1, X_2, ...X_m are the independent variables; a is the intersection; b_1, b_2, ...b_m are the multiple regression coefficients, estimated using the method of least squares; and *C* is the error term [50, 51]. ML is a scientific discipline in the artificial intelligence field that creates systems that learn automatically [8, 14]. For gap filling using this technique, the data available at each station were divided randomly to generate a training dataset (train) and test dataset (test), in proportions of 75 % and 25 %, respectively [8]. The algorithms implemented were MRM, K-nearest neighbors (KNN), gradient boosting trees (GBT), and random forest (RF). In addition, an optimization process was carried out, generating OML-MRM, OML-KNN, OML-GBT, and OML-RF models. These algorithms were implemented using Python programming language. KNN is a non-parametric method that can be used for both classification and regression.

The result is calculated based on the weighting of a number of nearest neighbors in the attribute space based on a distance function; the most common is Euclidian distance for continuous data [8]. GBT is a method in which multiple decision trees are iteratively fit to the data, and each tree is based on the previous tree to reduce losses and improve performance. It is based on the boosting principle, that is, on the creation of a set of weak learners to improve prediction precision [8, 52]. This method has three advantages: first, it does not require the application of a direct physical model, second, it serves as a computationally feasible method of capturing complex non-linear interactions between variables and a response [52, 53], and finally, it presents almost no overfitting problems, an important advantage, as many models over- or underestimate results [14, 52, 53]. RF was proposed by [54]. It is a semi-unsupervised non-parametric algorithm in the decision tree family that consists of a set of uncorrelated trees to produce predictions for classification and regression tasks [55].

2.2.5. Bayesian Optimization

One of the critical aspects of machine learning models' efficiency is hyperparameter selection. It is very important to establish the correct values; performance can change drastically from excellent to very poor. A common practice in the scientific community uses a trial and error technique, where different values, ranging from tens to thousands of possibilities, are evaluated [23, 31]. Therefore, efficiently setting the hyperparameter space is essential, because if the hyperparameter space is ample, the algorithm wastes significant time in non-promising configurations, (apart from being very slow). On the other hand, when the hyperparameter space is small, an accurate hyperparameter configuration set may be missing, even though it is fast [23, 31].

Bayesian optimization was used to estimate the hyperparameters, due to its great popularity in machine learning models and its good performance in optimization [56, 57]. The procedure consists of four steps, as described by [23]: (1) define the hyperparameter space; (2) the algorithm considers previous evaluations to choose the next set of values to be evaluated (acquisition function); (3) to assess the new hyperparameter configuration using an objective function y (4) If the optimization process has not finished yet, it goes to the second point. In this work, this algorithm was implemented using Python.

2.2.6. Evaluation Metrics

To assess the efficiency of the developed models, coefficient of determination (R^2), root mean square error (RMSE), Nash Sutcliffe coefficient (NSE) and percentage bias (PBIAS) were used [8, 51, 58]. All of them are mathematically expressed as Equations (10) - (13), respectively:

$$R^2 = \frac{\left[\sum_{t=1}^{n}(P_{obs} - \bar{P}_{obs})(P_{ped} - \bar{P}_{pred})\right]^2}{\sum_{t=1}^{n}(P_{obs} - \bar{P}_{obs})^2 \sum_{t=1}^{n}(P_{pred} - \bar{P}_{pred})^2} \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_{obs,i} - P_{pred,i})^2}{n}} \quad (11)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(P_{pred,i} - P_{obs,i})^2}{\sum_{i=1}^{n}(P_{pred,i} - \bar{P}_{pred,i})^2} \quad (12)$$

$$PB = \frac{\sum_{i=1}^{n}(P_{obs,i} - P_{pred,i}) \times 100}{\sum_{i=1}^{n} P_{pred,i}} \quad (13)$$

where n represents the number of prediction days, P_{obs} corresponds to the measured value for a specific day, P_{pred} is the predicted value, i represents measurement on a specific day, $\bar{P}_{obs}$ and $\bar{P}_{pred}$ correspond to the average measured and predicted values, respectively.

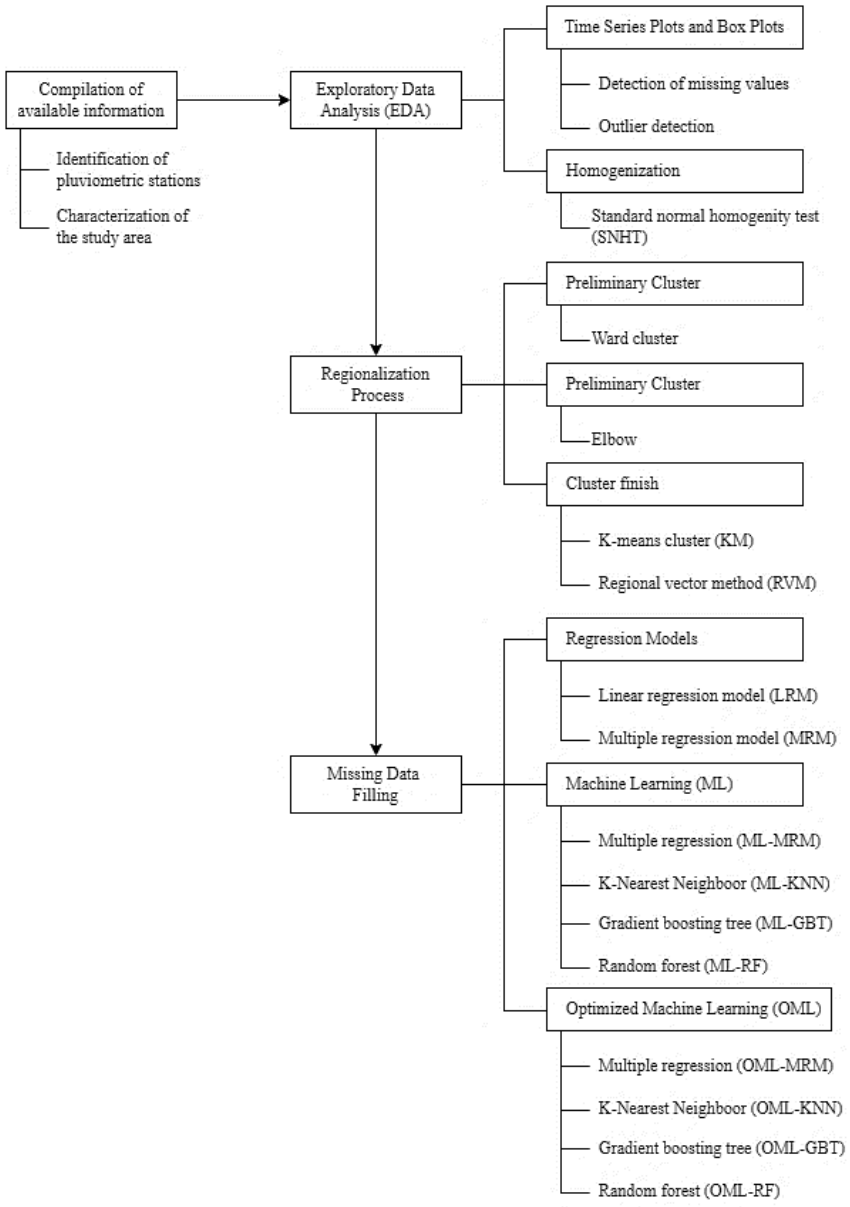


Figure 2. Methodological diagram for daily precipitation series gap filling.

3. Results

3.1. Analysis of missing data, outliers, and homogenization

In Figure 3a, the bar graph shows the quantity of unavailable precipitation data by station; there are 3 stations with more than 10 % missing data (Cañete, Socsi, and Pacaran), while the remaining stations present less than 10% missing data. The Cañete, Socsi and Pacaran rainfall stations are located in the lower part of the basin, which is characterized by being dry almost all year round (less than 20 mm/year).

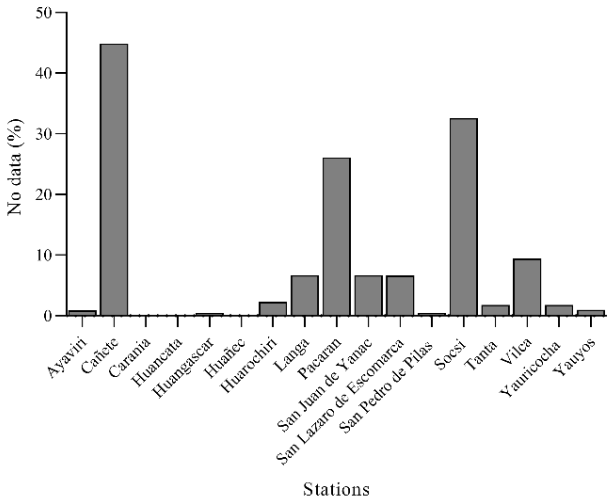
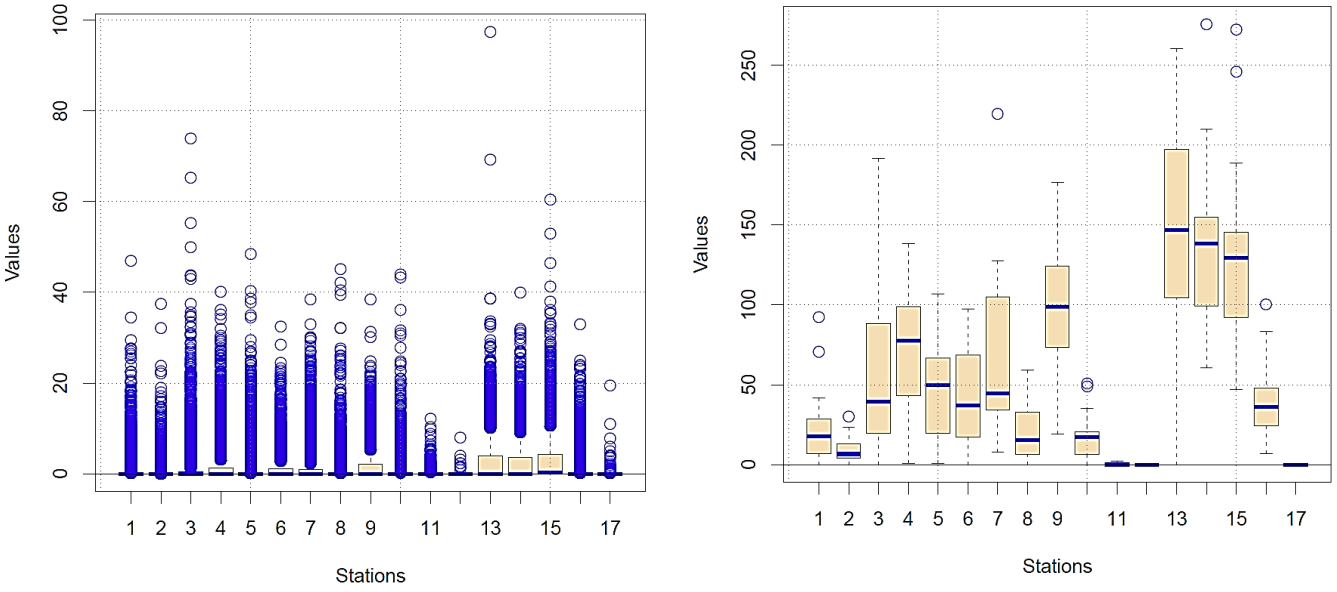


Figure 3. Missing daily precipitation data: Quantity of unavailable daily precipitation data as a percentage by station.

Figure 4a, boxplot for daily precipitation series of each station, shows a large number of scattered values, which initially could be considered outliers. However, it should be taken into account that daily precipitation shows high temporal and spatial variability patterns. Figure 4b shows the boxplot at a monthly scale, showing smaller dispersions, probably lower outliers, reflecting less spatial and temporal variability.



(**a**) (**b**)

Figure 4. Exploratory analysis of outliers: (**a**) Daily series and (**b**) Monthly series.

Figure 5a shows the results of Pearson coefficient correlations ŕ; high spatial and temporal variability on a daily scale are observed, complicating the detection of homogeneities. The high variability of daily records compared to that of monthly or annual values makes it very difficult to

directly apply methods for identifying inhomogeneities at the daily scale, in accordance with the recommendations of [34], the homogenization process was performed at a monthly scale, at which it is possible to detect cutoffs or breakpoints. Once the breakpoints were identified, the homogenization process was carried out on a daily scale using the Climatol package in R (https://cran.r-project.org/web/packages/climatol/index.html) [34, 59]. The results in Figures 5a and 5b show the correlation between the original normalized series and the reference series obtained based on the other stations. The reference series was constructed based on the average value of the nearest stations, weighted by the inverse of the distance from the analysis station [34, 39, 41]. The daily-scale correlation results present a maximum value of 0.40 and a minimum below zero (Figure 5a); the monthly-scale results reach values close to 1.0 (Figure 5b). This analysis was carried out for all the stations.

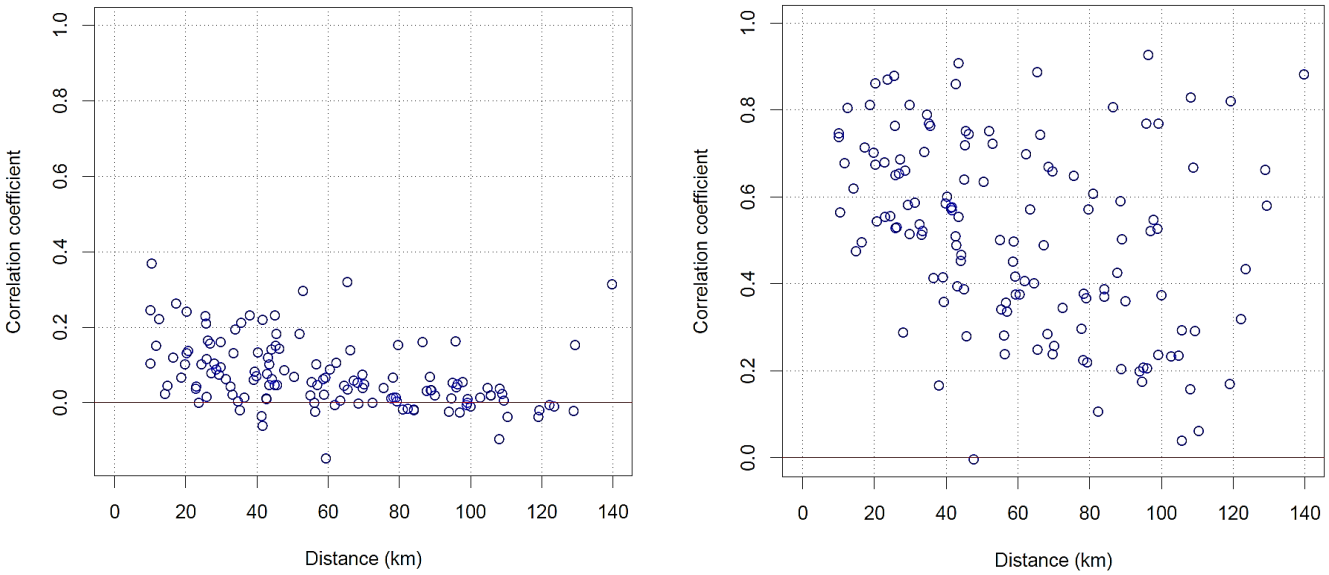


Figure 5. Correlogram between stations: (**a**) Daily precipitation series and (**b**) Monthly precipitation series.

Climatol provided overall absolute maximum autocorrelation (ACmx), SNHT, root mean square error (RMSE), and percentage of original data (POD) results. The ACmx values are not significant until the third quartile of the series (0.34); the values are below 60% autocorrelation, which indicates that the series are non-seasonal (Figure 6a). The series present anomalies in SNHT values between the original and homogenized series; the values range from 9.10 to 80.90, with the exception of the Cañete station, which reaches a maximum of 228, creating a rather wide variation spectrum (Figure 6b). RMSE presents high variation, with a minimum value of 1.26 and a maximum of 4.79 (Figure 6c). Finally, POD, which compares the original and homogenized data series, presents high values, meaning that the original data available is of good quality (Figure 6d). In addition, results of the analysis of homogeneity by station were obtained (Table 2). The Cañete, Socsi, and Pacaran stations presented ACmx values

above 0.60, SNHT values above 90.0, and POD values above 10%. Only RMSE presented low values.

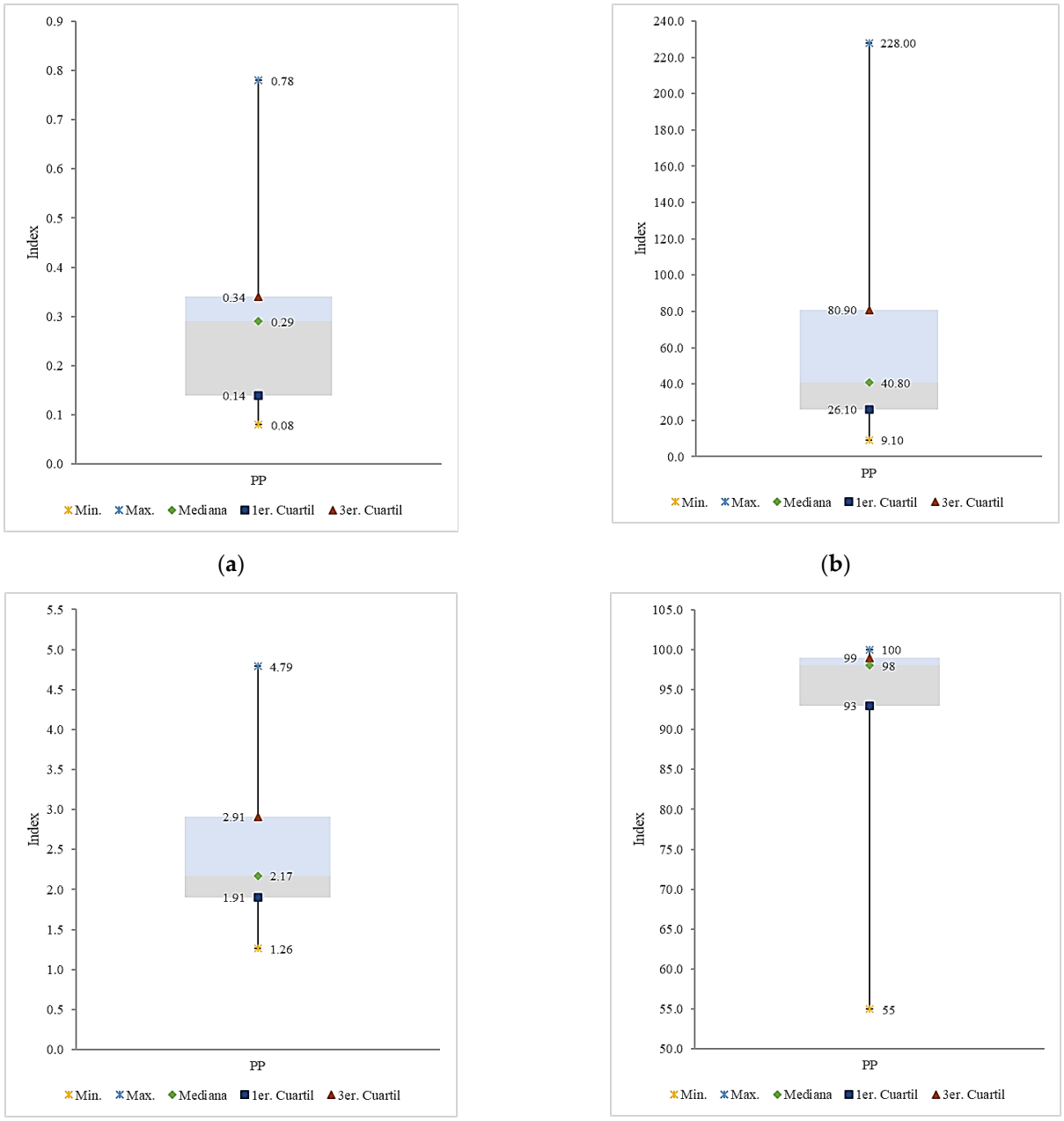


Figure 6. Homogeneity analysis statistics: (**a**) Station maximum absolute autocorrelation (ACmx), (**b**) Standard normal homogeneity test (SNHT), (**c**) Root mean squared error (RMSE) and (**d**) Percentage of original data (POD).

3.2. *Regionalization* analysis

The results provided by ward showed 4 groups of regions for the 17 stations (Figure 7). This process, implemented based on R code, allowed the initial clustering to be ascertained. Clustering analysis with KM is a method that creates the most heterogenous clusters possible; that is, the

objects in the k-clusters must be as similar as possible to those that belong to their cluster and completely unlike the objects in other clusters [11]. A fundamental point in the application of KM is to ascertain the optimum number of clusters. There are many criteria for choosing the optimal number of clusters, however; for this study the elbow method (EM) was used due to its extensive application in diverse hydrological studies with good results. The optimal cluster or region value is shown in Figure 8. According to EM analysis, the optimal number of regions was 4. In addition, KM was used to define the stations belonging to each homogenous region. Table S3 (Supplementary Material) details the number and names of the stations in each region.

Table 2. Homogeneity analysis statistics for each station.

Stations	ACmx	SNHT	RMSE	POD
Ayaviri	0.19	47.4	2.9	99
Cañete	0.65	228.0	1.3	55
Carania	0.20	26.1	2.6	100
Huancata	0.33	95.0	2.4	100
Huangascar	0.14	35.9	1.9	99
Huañec	0.29	68.7	2.2	100
Huarochiri	0.13	55.0	2.7	97
Langa	0.08	80.9	2.0	93
Pacaran	0.73	166.1	1.3	73
San Juan de Yanac	0.10	21.5	1.5	93
San Lazaro de Escomarca	0.32	20.9	3.4	93
San Pedro de Pilas	0.15	13.4	2.0	99
Socsi	0.78	40.8	1.4	67
Tanta	0.34	155.6	4.8	98
Vilca	0.34	30.5	3.9	90
Yauricocha	0.36	38.6	4.6	98
Yauyos	0.08	9.1	1.9	99

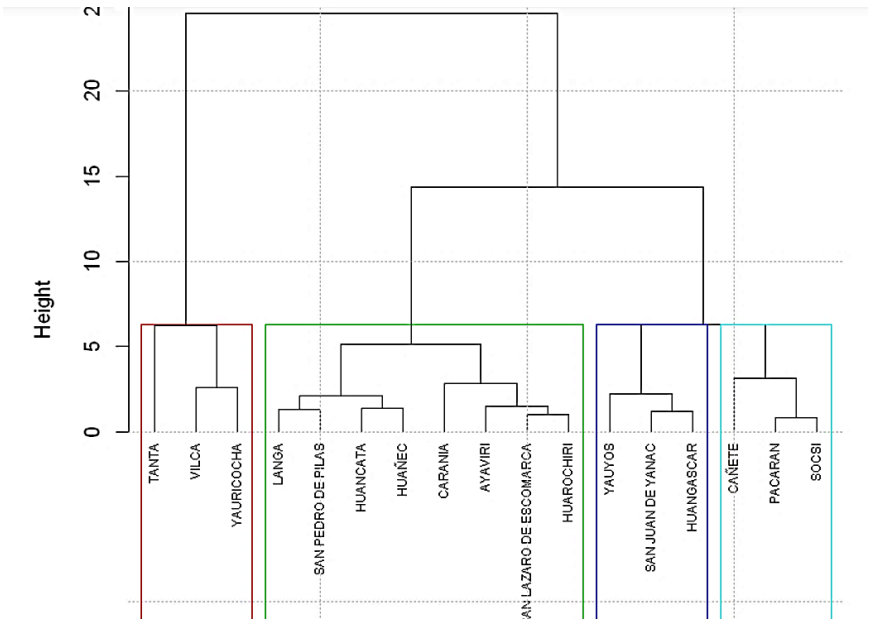


Figure 7. Ward method clustering: Dendrogram (2001-2019 period).

The results obtained with ward and KM indicate that precipitation during the evaluated period was not similar at every station throughout the watersheds. The application of the ward and KM methods was performed using code written in R and for EM, code written in Python.

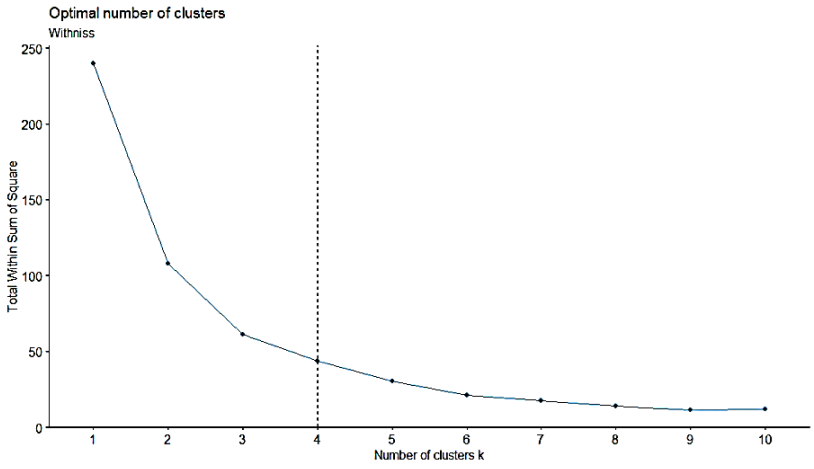


Figure 8. Optimal number of clusters according to the elbow method.

Finally, the RVM method was applied to validate the results obtained based on the described models. The Hydraccess program was used to apply RVM (https://hybam.obsmip.fr/es/hydraccess-3). The results show clustering of stations with similar behaviors in terms of interannual precipitation variation, taking the standard deviation and correlation coefficient /vector as indicators. The regions are considered homogenous if the values of the standard deviation (SD) are lower than 0.4 and the correlation coefficient/vector values are above 0.7 [11]. The final results show the clustering of rainfall stations into homogenous regions.

The RVM method was used to obtain three final clusters that, in accordance with their statistics and analysis of the results, included the stations that are shown in Table 3 and Tables S2 and S3 (Supplementary Material), and Figure 9, and Figures S1 and S2 (Supplementary Material). It was not possible to analyze cluster 3, as its stations presented a high percentage of missing data.

Table 3. Annual regional vector indices – Region 1.

Station	N° Years	Standard deviation	Station/vector correlation
Langa	16	0.252	0.882
San Lazaro de Escomarca	16	0.263	0.664
Ayaviri	17	0.116	0.904
Huancata	19	0.341	0.863
Huañec	19	0.187	0.751
Huarochiri	14	0.159	0.851
Carania	19	0.191	0.679

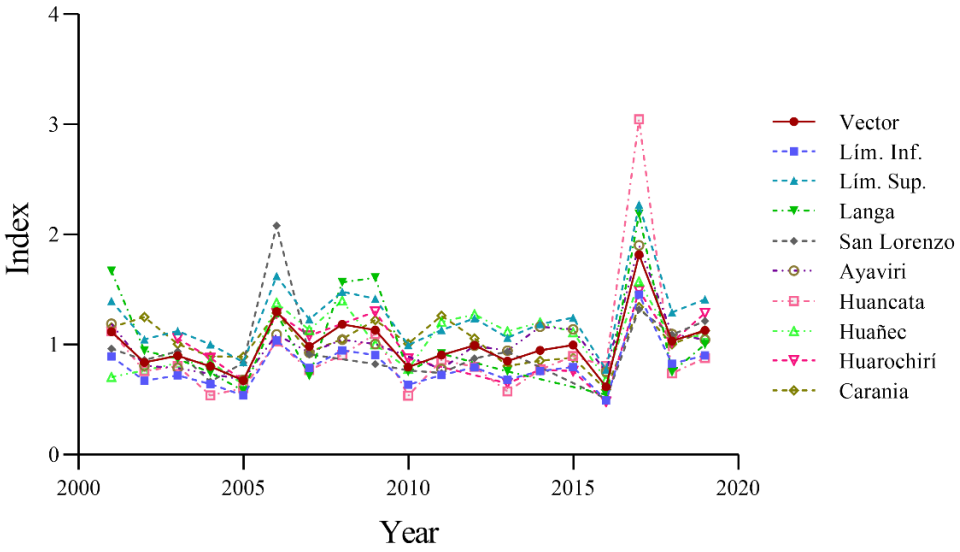


Figure 9. Annual indices of the regional vector and stations in Region 1.

The results obtained with the ward and KM methods are consistent in the number of homogenous regions. However, the number of stations in Regions 1 and 2 presented a slight discrepancy between the results obtained with ward and KM; therefore, the results obtained with RVM were used for verification, showing accord between the KM and RVM results. Tables S3 (Supplementary Material) shows the final results of the homogenous region clustering. In addition, Figure 10 shows the regionalization of rain gauge stations based on the KM and RVM results.

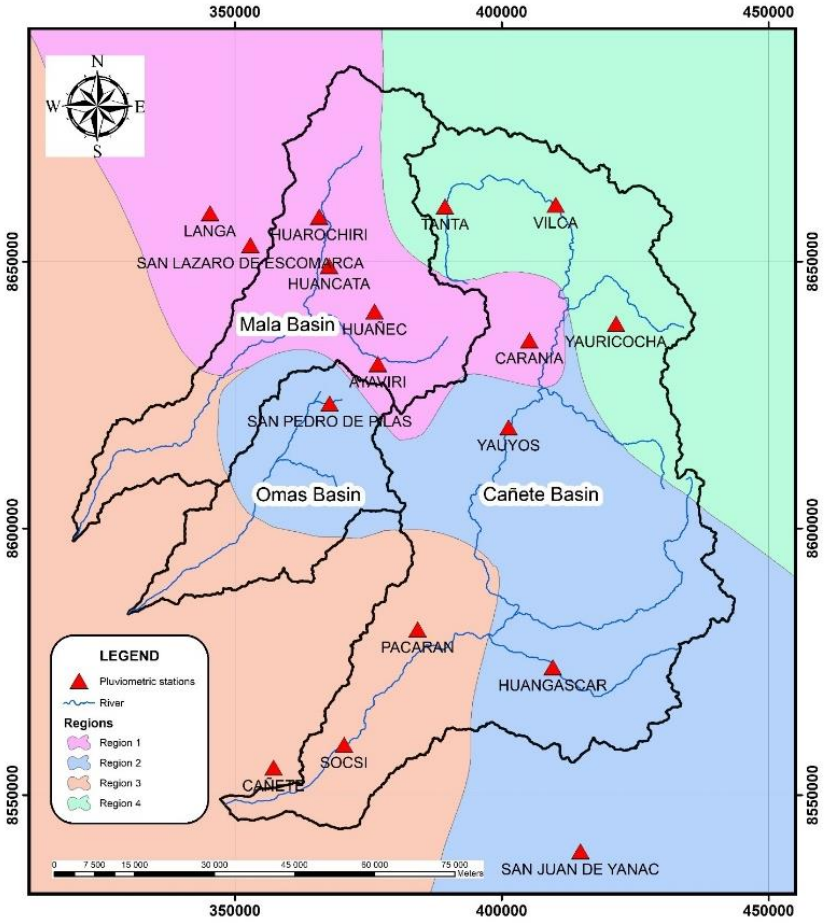


Figure 10. Regionalization of the stations according to the KM and MVR methods.

3.3. Analysis of the series gap-filling process

In Table 4 and Tables S4 and S5 (Supplementary Material), the correlation values corresponding to the stations clustered by homogenous region are shown. The correlations were below 0.60 and above 0.38; below 0.58 and above 0.32, and below 0.45 and above 0.37 in Regions 1, 2, and 4, respectively. These coefficients are considered acceptable given the dry conditions, with more than 90% of the rain gauge records close to zero throughout the year due to the hydroclimatic conditions, with any value greater than zero causing high variability [11]. Therefore, this analysis allowed the level of representation using Pearson coefficient correlations within a region to be highlighted.

In the application of RM for filling missing precipitation data, the models were gener- ated based on the homogenous regions. For the LRM algorithm, the Ayaviri station was designated as variable Y and the Huancata station as variable *X*, based on the greater Pearson coefficient correlations value. The other stations with missing data were selected in a similar manner. Table 5 shows the *Y* and *X* variables for each region. Meanwhile, for MRM, the procedure was similar to that of the previous case; the Ayaviri station was identified as the *Y* variable and all the remaining stations (Huancata, Langa, San Lazaro de Escomarca, Huañec, Huarochiri, and Carania) as X_m (see Table 5).

For the application of the different gap-filling algorithms, this process was carried out independently in each homogenous region. The Y stations with missing data in each homogenous region were identified, as were the X_m stations corresponding to each target station. The Y and X_m stations for each homogenous region are shown in Table 5.

Table 4. Correlation coefficient - Region 1.

	Ayaviri	Carania	Huancata	Huañec	Huarochiri	San Lazaro de Escomarca	Langa
Ayaviri	1						
Carania	0.48	1					
Huancata	0.60	0.47	1				
Huañec	0.51	0.43	0.49	1			
Huarochiri	0.56	0.55	0.58	0.46	1		
San Lazaro de Escomarca	0.45	0.40	0.43	0.39	0.44	1	
Langa	0.45	0.38	0.46	0.38	0.46	0.55	1

For the filling of missing precipitation data with ML, the analysis was also carried out independently in each homogenous region. First, the available data were divided, with one portion for training and another for testing (75% and 25% respectively); this division was performed randomly. Then the ML-MRM, ML-KNN, ML-GBT and ML-RF were selected along with their respective parameters (see Table 6). In addition, considering that many models contain parameters that cannot learn from training data, it was necessary to carry out an optimization process. To this end, it was important to ascertain the hyperparameter values using the Bayesian Optimization method. The results of a model can depend largely on the values taken by its hyperparameters; however, it cannot be known beforehand what values are suitable. The most common means of finding optimal values is testing different possibilities; in this study optimization processes were carried out for the OML-MRM, OML-KNN, OML-GBT, and OML-RF algorithms (Table 7).

Based on the Y and X_m variables, ML was first applied for default parameter values using the ML-MRM, ML-KNN, ML-GBT, and ML-RF models. It was also applied using parameters called hyperparameters, generating the OML-MRM, OML-KNN, OML-GBT, and OML-RF models. This process allowed the model parameters to be optimized. The parameter and hyperparameter values used in the algorithms created in Python are shown in Table 6.

Table 6 describes the parameter and hyperparameter values used in each algorithm in ML. It is observed that only one parameter value was assigned when using the default algorithm. However, for the algorithm optimization process, a wide range of values was defined and using the Bayesian optimization method the optimal hyperparameters were estimated.

Table 5. Identification of target stations (Y) and predictor stations by homogenous region.

Regions	Target Station (Y)	Predictor Station (X)	Multiple Predictor Stations (Xm)
Region 1	Ayaviri	Huancata	Huancata, Langa, San Lazaro de Escomarca, Huañec, Huarochiri, Carania
	Huarochiri	Huancata	Huancata, Langa, San Lazaro de Escomarca, Ayaviri, Huañec, Carania
	San Lazaro de Escomarca	Langa	Langa, Ayaviri, Huancata, Huañec, Huarochiri, Carania
	Langa	San Lazaro de Escomarca	San Lazaro de Escomarca, Ayaviri, Huancata, Huañec, Huarochiri, Carania
Region 2	San Pedro de Pilas	Huangascar	Huangascar, San Juan de Yanac, Yauyos
	Huangascar	San Pedro de Pilas	San Pedro de Pilas, Yauyos, San Juan de Yanac
	Yauyos	San Pedro de Pilas	San Pedro de Pilas, Huangascar, San Juan de Yanac
	San Juan de Yanac	San Pedro de Pilas	San Pedro de Pilas, Huangascar, Yauyos
Region 4	Tanta	Vilca	Vilca and Yauricocha
	Yauricocha	Vilca	Vilca and Tanta
	Vilca	Yauricocha	Yauricocha and Tanta

3.4. Assessment of model performance

To assess the performance of the models, different statistical metrics – R^2, RMSE, NSE and PBIAS – were used for both datasets (training and test). The obtained results are presented in Table 7 and Tables S6, and S7 (Supplementary Material). These statistics were calculated for the 2001-2019 period; periods with missing data were not considered.

Many linear models, among them LRM, contain parameters that cannot learn from training data, making it necessary for the modeler to establish them. In addition, to establish the predictive capacity of ML, which consists of testing how close its predictions are to the actual values of the response variable, a set of observations is needed, with its corresponding response variables, but that the model has not "seen", that is, which have not participated in its initial fitting. Finally, to assess the performance of models by comparing predicted and actual precipitation values, the use of statistical metrics is important.

Table 6. Parameter and hyperparameter values for the ML algorithms.

Algorithm	Parameters [values]	Hyperparameters [values]
Multiple Regression	alpha [1] solver ['auto'] modelo[Ridge]	alpha [logspace(-5, 5, 500)] solver ['auto'] modelo[Ridge]
K-nearest neighbors	n_neighbors [5] leaf_size [30] algoritm ['auto'] modelo [KNeighborsRegressor]	n_neighbours [linspace(1, 100, 500] leaf_size [1, 3] algoritm ['auto'] modelo [KNeighborsRegressor]
Gradient boosting tree	n_estimators [100] max_feature ['none'] max_depth [3] subsample [1] modelo[GradientBoostingRegressor]	n_estimators [50, 100, 1000, 2000] max_feature ['auto', 3, 5, 7] max_depth ['None', 3, 5, 10, 20] subsample [0.5, 0.7, 1] modelo[GradientBoostingRegressor]
Random forest	n_estimators [100] max_feature ['auto'] max_depth ['None'] modelo[RandomForestRegressor]	n_estimators [50, 100, 1000, 2000] max_feature ['auto', 3, 5, 7] max_depth ['None', 3, 5, 10, 20] modelo[RandomForestRegressor]

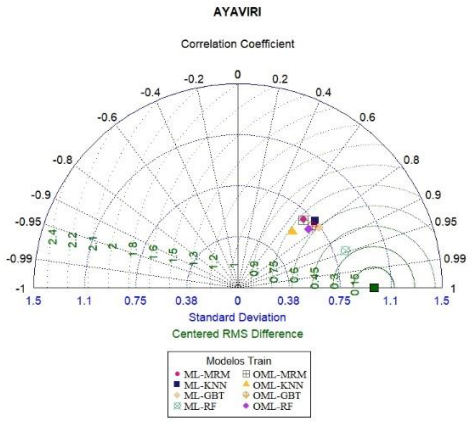
The R^2 values for the dataset (training and test) present a correlation between the Y and X variables in each model. For the Ayaviri station, which belongs to homogenous region 1 (see Table 7), the ML-RF model gives the best R^2 value (0.89) for the training data; however, for the test dataset, this value is reduced by nearly half (R^2 = 0.45). For optimized ML, the training and test R2 values are close to each other, and in some cases the R^2 values are better for the test datasets than the training datasets. RMSE is a measure of the variance of residuals, which allows the magnitude of deviation of simulated values from observed values to be quantified; the LRM model presents the greatest RMSE (3.15) for the Ayaviri station. It was also observed that the test dataset generally presents a lower RMSE, particularly with the optimized ML models (OML-GBT and OML-RF).

The NSE is a tool that measures the predictive capacity of a model, which can take values between -∞ and 1.0, with 1.0 being the optimal value. Values between 0.0 and 1.0 are generally seen as acceptable performance levels, while values equal to or less than 0.0 indicate that the mean of the observed values is a better predictor than the simulated value, indicating inadequate performance [60].In accordance with the results shown in Table 7, values for the Ayaviri station are between 0.36 and 0.88 for both datasets (test and training). However, the ML models present values very close to 1.0 (ML-RF, NSE = 0.88) for the training dataset, indicating an acceptable level of performance.

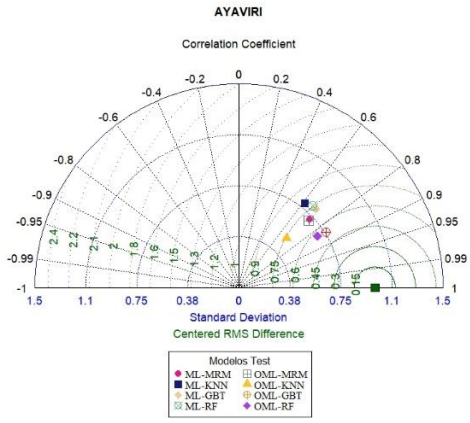
Table 7. Model efficiency according to fit statistics - Region 1.

Stations	Samples	Statistics	LRM	MRM	Machine Learning				Optimized Machine Learning			
					MRM	KNN	GBT	RF	MRM	KNN	GBT	RF
Ayaviri	Train	R^2	0.36	0.49	0.48	0.57	0.64	0.89	0.48	0.49	0.59	0.59
	Train	RMSE	3.15	2.81	2.87	2.61	2.39	1.36	2.87	2.89	2.55	2.58
	Train	NSE	0.36	0.49	0.48	0.57	0.64	0.88	0.48	0.47	0.59	0.58
	Train	PBIAS	0.00	0.00	0.00	3.92	0.00	-1.73	0.00	0.45	0.00	0.38
	Test	R^2			0.52	0.38	0.49	0.45	0.52	0.48	0.71	0.70
	Test	RMSE			2.62	3.03	2.75	2.86	2.62	2.83	2.05	2.14
	Test	NSE			0.52	0.36	0.47	0.43	0.52	0.44	0.71	0.68
	Test	PBIAS			0.00	0.67	-8.46	-10.65	0.00	21.64	0.00	1.01
Huarochiri	Train	R^2	0.34	0.49	0.49	0.60	0.65	0.92	0.49	0.51	0.60	0.61
	Train	RMSE	3.12	2.74	2.80	2.47	2.32	1.19	2.80	2.76	2.49	2.48
	Train	NSE	0.34	0.49	0.49	0.60	0.65	0.91	0.49	0.50	0.60	0.60
	Train	PBIAS	0.00	0.00	0.00	6.48	0.00	-1.39	0.00	4.38	0.00	0.47
	Test	R^2			0.52	0.41	0.51	0.49	0.53	0.53	0.73	0.73
	Test	RMSE			2.54	2.83	2.58	2.64	2.51	2.57	1.93	1.96
	Test	NSE			0.52	0.40	0.50	0.48	0.53	0.51	0.72	0.71
	Test	PBIAS			-1.72	7.12	-5.63	-9.30	0.00	18.36	0.00	0.95
San Lazaro de Escomarca	Train	R^2	0.30	0.38	0.38	0.49	0.65	0.90	0.38	0.41	0.54	0.45
	Train	RMSE	3.44	3.22	3.16	2.87	2.42	1.41	3.17	3.11	2.75	3.03
	Train	NSE	0.30	0.38	0.38	0.49	0.64	0.88	0.38	0.40	0.53	0.43
	Train	PBIAS	0.00	0.00	0.00	7.01	0.00	-1.96	0.00	10.88	0.00	0.14
	Test	R^2			0.42	0.28	0.34	0.37	0.41	0.43	0.73	0.56
	Test	RMSE			3.33	3.73	3.55	3.46	3.34	3.35	2.31	2.98
	Test	NSE			0.42	0.27	0.33	0.37	0.41	0.41	0.72	0.53
	Test	PBIAS			0.00	16.25	9.41	1.78	0.00	14.86	0.00	-0.05
Langa	Train	R^2	0.30	0.39	0.40	0.53	0.68	0.93	0.40	0.45	0.59	0.60
	Train	RMSE	1.98	1.85	1.85	1.64	1.37	0.71	1.85	1.80	1.55	1.55
	Train	NSE	0.30	0.39	0.40	0.53	0.67	0.91	0.40	0.43	0.58	0.58
	Train	PBIAS	0.00	0.00	0.00	5.79	0.00	-3.09	0.00	10.61	0.00	0.60
	Test	R^2			0.36	0.24	0.32	0.31	0.37	0.36	0.70	0.70
	Test	RMSE			1.85	2.09	1.94	1.98	1.83	1.87	1.28	1.33
	Test	NSE			0.35	0.17	0.28	0.26	0.37	0.34	0.69	0.67
	Test	PBIAS			-4.32	0.76	-7.22	-16.36	0.00	17.93	0.00	1.23

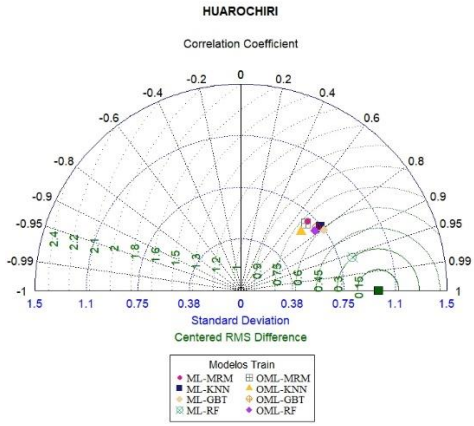
PBIAS measures the tendency of simulated data to be larger or smaller than their observed counterparts; its optimal value is 0. Positive values indicate a model with an underestimation bias and negative values indicate an overestimation bias. For the Ayaviri station, the OML-KNN presents high underestimation (PBIAS = 21.64), while the ML-RF model presents high overestimation (PBIAS = -10.65). However, the LRM, MRM, ML-MRM, OML-MRM, and OML-GBT models present an optimal PBIAS value for both datasets (training and test).


AYAVIRI
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Train
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

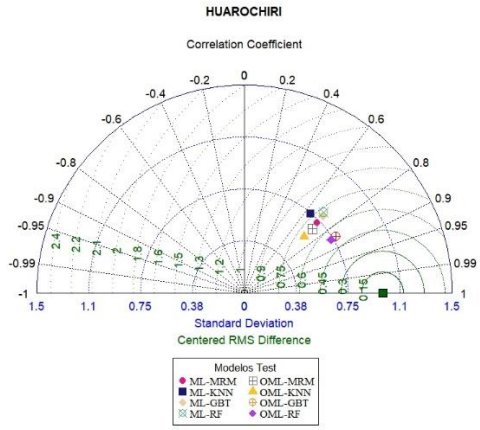
(**a**)


AYAVIRI
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Test
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

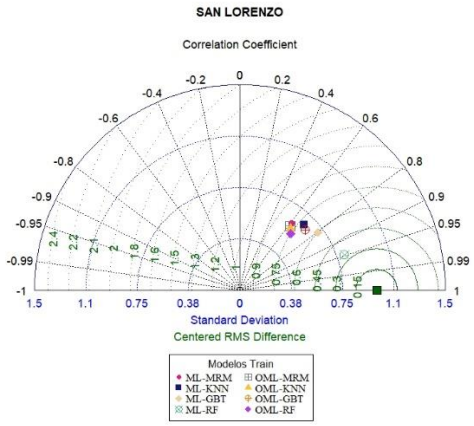
(**b**)


HUAROCHIRI
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Train
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

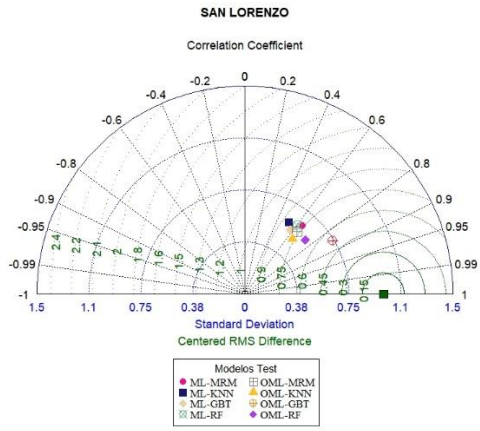
(**c**)


HUAROCHIRI
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Test
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

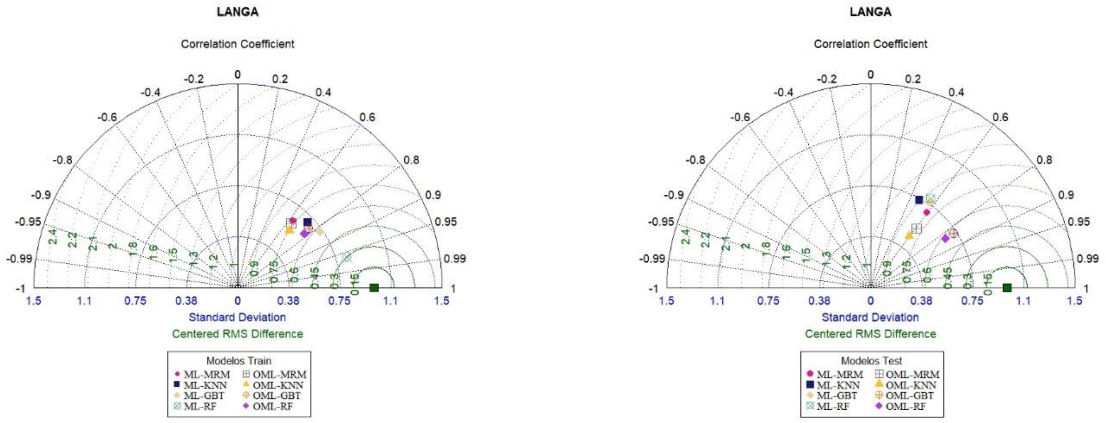
(**d**)


SAN LORENZO
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Train
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

(**e**)


SAN LORENZO
Correlation Coefficient
Standard Deviation
Centered RMS Difference
Modelos Test
ML-MRM
ML-KNN
ML-GBT
ML-RF
OML-MRM
OML-KNN
OML-GBT
OML-RF

(**f**)



(g) (h)

Figure 11. Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and modeled precipitation based on precipitation datasets (training and test) for four stations: (**a**) Ayaviri (training), (**b**) Ayaviri (test), (**c**) Huarochiri (training), (**d**) Huarochiri (test), (**e**) San Lazaro de Escomarca (training), (**f**) San Lazaro de Escomarca (test), (**g**) Langa (training), and (**h**) Langa (test).

4. Discussion

Based on the results obtained in the exploratory analysis, the Cañete, Socsi and Pacaran stations presented large quantities of missing values (over 10%); they also failed the homogeneity test. Therefore, the initial number of stations (17) was reduced to 14. In this study the RM and ML methods were used to fill gaps in daily precipitation series at stations located in the MOC watersheds on the Peruvian Pacific Slope and coast. The procedure was carried out in four stages: collection of information on daily precipitation series; exploratory analysis and homogenization. Therefore, it is essential to implement quality control procedures for raw rainfall data to ensure their reliability for use. In addition, preliminary ward cluster analysis, followed by KM and RVM analysis, through which three homogenous regions that concisely represent the relationship between precipitation variability and altitude were identified; and, finally, the application of RM and ML as a method of filling gaps in precipitation series.

RM and ML are customizable and easy-to-implement techniques that seek the best performance for a given problem among numerous algorithms. ML analyses with hyperparameter values (OML-MRM, OML-KNN, OML-GBT, and OML-RF) presented the best data recovery performance, demonstrating that ML models can extract additional information from data that by nature present noisy characteristics due to their high spatial and temporal variability [8, 34, 39]. In general terms, the decision tree methods (OML-GBT and OML-RF) perform the regression task well; however, some variations are observed that demonstrate that not all ML algorithms are equal in datasets that are superficially similar and can vary widely in terms of their prediction power. This also underlines the variation in the mechanisms of ML algorithms, even though

all of them are capable of extracting information from non-linear and noisy datasets.

Figure 11a shows the Taylor diagram for the Ayaviri station for the training dataset; the ML-RF model presents the best results, with prediction precipitation the most consistent with observed precipitation (R^2= 0.89, RMSE = 1.36, NSE = 0.88, and PBIAS = -1.73). Figure 11b shows the Taylor diagram for the Ayaviri station for the test dataset; the OML-GBT and OML-RF present the best results (R^2 = 0.71, RMSE = 2.05, NSE = 0.71, and PBIAS = 0.00 and R^2 = 0.70, RMSE = 2.14, NSE = 0.68 y PBIAS = 1.01, respectively). The analyses of the other stations (Huarochiri, San Lazaro de Escomarca, and Langa), are shown in Figures 11c, 11d, 11e, 11f, 11g, and 11h; all these stations are located in homogenous region 1 and the values of the results obtained for them are similar to those of the Ayaviri station. Likewise, it is observed that in terms of the statistical metrics for the training and test datasets, the optimized ML models present the best results, particularly the OML-GBT and OML-RF models. The results of the analysis of the statistical metrics are shown in the figures. For the Ayaviri station, the OML-RF model presents a slight underestimation, while the results of the OML-GBT model are more efficient. Finally, in regions 2 and 4, the OML-GBT and OML-RF present the best results in terms of statistical metrics.

5. Conclusions

This study has demonstrated the performance advantages of ML techniques for filling gaps in daily precipitation series, as well as the potential of ML models in the optimization process using hyperparameter values for training (75%) and test datasets (25%), based on the efficiencies of the statistical metrics. However, it is important to note that a quality control raw rainfall data and regionalization process are necessary, which allows homogenous regions to be identified. Precipitation along the Peruvian Pacific Slope is highly influenced by El Niño, with marked positive asymmetry of strong events, and La Niña, with non-Gaussian distribution of precipitation data, which limits to a certain extent the linear analysis approach [9]. Finally, the results obtained in this study showed that the OML-GBT and OML-RF models presented the least variability in estimation errors and the best approximation to the actual data, efficiently interpreting the spatiotemporal variability of precipitation, as demonstrated by the analyzed statistical metrics.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: Annual indices of the regional vector and stations in Region 2; Figure S2: Annual indices of the regional vector and stations in Region 4; Figure S3: Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and modeled precipitation based on precipitation datasets (training and test) for four stations: (**a**) San Pedro de Pilas (training), (**b**) San Pedro de Pilas (test), (**c**) Huangascar (training), (**d**) Huangascar (test), (**e**) Yayos (training), (**f**) Yayos (test), (**g**) San Juan de Yanac (training), and (**h**) San Juan de Yanac (test); Figure S4: Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and

modeled precipitation based on precipitation datasets (training and test) for four stations: (**a**) Tanta (training), (**b**) Tanta (test), (**c**) Yauricocha (training), (**d**) Yauricocha (test), (**e**) Vilca (training), (**f**) Vilca (test); Table S1: K-means clustering (2001-2019 period); Table S2: Annual regional vector indices – Region 2; Table S3: Annual regional vector indices – Region 4; Table S4: Correlation coefficient - Region 2; Table S5: Correlation coefficient - Region 4; Table S6: Model efficiency according to fit statistics - Region 2; Table S7: Model efficiency according to fit statistics - Region 4.

Author Contributions: Conceptualization, M.P-M. and J.L.A.; methodology, M.P-M.; software, M.P.M.; validation, M.P-M.and J.L.A.; formal analysis, M.P-M.; investigation, M.P-M., J.L.A., O.L., A.S. and N.M.A.; writing—original draft preparation, M.P-M.; writing—review and editing, M.P-M., J.L.A., O.L. A.S. and N.M.A.; visualization, M.P-M. and N.M.A.; supervision, J.L.A. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the project ANID/FONDAP/15130015.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://www.mdpi.com/article, time series, algorithms and other.

Acknowledgments: CRHIAM Water Research Center, Project ANID/FONDAP/15130015, Universidad Nacional Agraria La Molina, Eliana Contreras López and Ricardo León Ochoa.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, D.; Christakos, G.; Ding, X.;Wu, J., Adequacy of TRMM satellite rainfall data in driving the SWAT modeling of Tiaoxi catchment (Taihu lake basin, China). Journal of Hydrology, 2018. 556: 1139-1152. doi: https://doi.org/10.1016/j.jhydrol.2017.01.006.
2. Santos, L.O.F.d.; Querino, C.A.S.; Querino, J.K.A.d.S.; Pedreira Junior, A.L.; Moura, A.R.d.M.; Machado, N.G.; Biudes, M.S.., Validation of rainfall data estimated by GPM satellite on Southern Amazon region. Revista Ambiente & Água, 2019. 14.
3. Zambrano-Bigiarini, M.; Nauditt, A.; Birkel, C.; Verbist, K.; Ribbe, L., Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile. Hydrology and Earth System Sciences, 2017. 21(2): 1295-1295. doi:10.5194/hess-21-1295-2017
4. Jiang, L. and J. Wu. Hybrid PSO and GA for Neural Network Evolutionary in Monthly Rainfall Forecasting. 2013. Berlin, Heidelberg: Springer.
5. Cramer, S., Kampouridis, M., Freitas, A.A., Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. Expert Systems with Applications, 2017. 85: 169-181. doi:10.1016/j.eswa.2017.05.029
6. Chen, F.; Gao, Y.;Wang, Y.; Qin, F.; Li, X., Downscaling satellite-derived daily precipitation products with an integrated framework. International Journal of Climatology, 2019. 39(3): 1287-1304. [https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5879]. doi:10.1002/joc.5879.
7. Bai, P. and X. Liu, Evaluation of five satellite-based precipitation products in two gauge-scarce basins on the Tibetan Plateau. Remote Sensing, 2018. 10(8): 1316-1316. doi:10.3390/rs10081316.
8. Chivers, B.D.; Wallbank, J.; Cole, S.J.; Sebek, O.; Stanley, S.; Fry, M.; Leontidis, G., Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. Journal of Hydrology, 2020. 588: 125126-125126. doi:https://doi.org/10.1016/j.jhydrol.2020.125126.
9. Lavado Casimiro, W.S. ; Ronchail, J.; Labat, D.; Espinoza, J.C.; Guyot, J.L., Basin-scale analysis of rainfall and runoff in Perú (1969-2004): Pacific, Titicaca and Amazonas drainages. Hydrological Sciences Journal, 2012. 57(4): 625–642-625–642. doi:10.1080/02626667.2012.672985.

10. Espinoza Villar, J.C.; Ronchail, J.; Guyot, J.L.; Cochonneau, G.; Naziano, F.; Lavado, W.; De Oliveira, E.; Pombosa, R.; Vauchel, P., Spatio-temporal rainfall variability in the Amazon basin countries (Brazil, Peru, Bolivia, Colombia, and Ecuador). International Journal of Climatology, 2009. 29(11): 1574-1594. https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1791.
11. Rau, P.; Bourrel, L.; Labat, D.; Melo, P.; Dewitte, B.; Frappart, F.; Lavado, W.; Felipe, O. Regionalization of rainfall over the Peruvian Pacific slope and coast. International Journal of Climatology, 2017. 37(1): 143–158-143–158. doi:10.1002/joc.4693.
12. Körner, P.; Kronenberg, R.; Genzel, S.; Bernhofer, C. I, Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. Meteorologische Zeitschrift, 2018. 27(5): 369-376. doi:10.1127/metz/2018/0908.
13. Lavado Casimiro, W., Espinoza J.C. Impactos de El Niño y La Niña en las lluvias del Perú (1965-2007). Revista Brasileira de Meteorologia, 2014. 29: 171-182. doi:10.1590/S0102-77862014000200003.
14. Bertsimas, D., C. Pawlowski,. Zhuo, Y.D, From Predictive Methods to Missing Data Imputation: An Optimization Approach. J. Mach. Learn. Res., 2017. 18(1): 7133-7171. doi:https://dl.acm.org/doi/abs/10.5555/3122009.324205.
15. Teegavarapu, R.S.V. and Chandramouli V., Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 2005. 312: 191-206.
16. Barrios, A., G. Trincado, and R. Garreaud, Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. Forest Ecosystems, 2018. 5(1): 28.
17. Xia, Y., Fabian, P., Stohl, A., Winterhalter, Forest climatology: estimation of missing values for Bavaria, Germany. Agricultural and Forest Meteorology, 1999. 96(1): 131-144.
18. Bostan, P.A., Heuvelink, G.B.M.and Akyurek, S.Z., Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. International Journal of Applied Earth Observation and Geoinformation, 2012. 19: 115-126.
19. Mair, A. and Fares, A., Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island. Journal of Hydrologic Engineering, 2011. 16(4): 371-383.
20. Simolo, C., Brunetti, M., Maugeri, M., Nanni. T, Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. International Journal of Climatology, 2010. 30(10): 1564-1576.
21. Huang, M., Lin R., Huang S., Xing T., A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. Advanced Engineering Informatics, 2017. 33: 89-95.
22. Gorshenin, A., Lebedeva, M., Lukina, S., Yakovleva, A. Application of Machine Learning Algorithms to Handle Missing Values in Precipitation Data. In: Vishnevskiy, V., Samouylov, K., Kozyrev, D. (eds) Distributed Computer and Communication Networks. DCCN 2019. Lecture Notes in Computer Science, 2019. 11965. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36614-8_43.
23. Bellido-Jiménez, J.A., Gualda, J.E. and García-Marín A.P., Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. Atmosphere, 2021. 12(9): 1158.
24. Devi, U., Shekhar, M.S., Singh, G.P., Rao, N.N., Bhatt, U.S., Methodological application of quantile mapping to generate precipitation data over Northwest Himalaya. International Journal of Climatology, 2019. 39(7): 3160-3170.
25. Estévez, J., Bellido-Jiménez, J.A., Liu, X., García-Marín, A.P, Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. Water, 2020. 12(7): 1909.
26. Sattari, M.T., Rezazadeh-Joudi, A. and Kusiak, A. Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research, 2016. 48(4): 1032-1044.
27. Tang, G., Clark, M. P., Newman, A. J., Wood, A. W., Papalexiou, S. M., Vionnet, V., & Whitfield, P. H., SCDNA: a serially complete precipitation and temperature dataset for North America from 1979 to 2018. Earth Syst. Sci. Data, 2020. 12(4): 2381-2409.
28. Tang, G., Clark, M.P. and Papalexiou S.M., SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019. Journal of Climate, 2021. 34(16): 6493-6511.
29. Carrera-Villacrés, D.V., et al., Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media. Idesia (Arica), 2016. 34: 81-90.

30. Luna Romero, A.E. and Lavado Casimiro, W.S., Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenta Jetepeque, Perú. Revista Tecnológica-ESPOL, 2015. 28(3).
31. Estévez, J., Gavilán, P. and Giráldez J.V., Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology, 2011. 402(1): 144-154.
32. Portuguez Maurtua, D.M., Aplicación de la geoestadística a modelos hidrológicos en la cuenca del río Cañete. Master's thesis. Universidad Nacional Agraria La Molina. Lima, Peru, 2017.
33. Guevara Ochoa, C., Briceño, N.; Zimmermann, E.D.; Vives, L.S.; Blanco, M.; Cazenave, G.; Ares, M.G. Relleno de series de precipitación diaria para largos periodos de tiempo en zonas de llanura: Caso de estudio cuenca superior del arroyo del Azul. Geoacta. 2017. 42 (1) http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1852-77442017000100004&lng=es&nrm=iso>. ISSN 1852-7744.
34. Guijarro, J., Homogenization of climatic series with Climatol. Reporte técnico State Meteorological Agency (AEMET). 2018. https://www.climatol.eu/homog_climatol-en.pdf
35. Toreti, A., Kuglitsch, F.G.; Xoplaki, E.; Della-Marta, P.; Aguilar, E.; Prohom, M.; Luterbacher, J. A note on the use of the standard normal homogeneity test (SNHT) to detect inhomogeneities in climatic time series. International Journal of Climatology, 2011. 31: 630-632. doi:10.1002/joc.2088.
36. Alexandersson, H., A homogeneity test applied to precipitation data. Journal of Climatology, 1986. 6(6): 661-675. [https://rmets.onlinelibra4r4y7.wiley. doi:https://doi.org/10.1002/joc.3370060607.
37. Alexandersson, H., Moberg, A. Homogenization of swedish temperature data. Part I: Homogeneity test for linear trends. International Journal of Climatology, 1997. 17(1): 25-34.
38. Moberg, A., Alexandersson, H. Homogenization of swedish temperature data. Part ii: homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. International Journal of Climatology, 1997. 17(1): 35-54.
39. Pandzic, K., Kobold, M.; Oskorus, D.; Biondic, B.; Biondic, R.; Bonacci, O.; Likso, T.; Curic, O. Standard normal homogeneity test as a tool to detect change points in climate-related river discharge variation: case study of the Kupa River Basin. Hydrological Sciences Journal, 2020. 65(2): 227-241. doi:10.1080/02626667.2019.1686507.
40. Ahmad, N.H. and S.M. Deni, Homogeneity test on daily rainfall series for Malaysia. Matematika: Malaysian Journal of Industrial and Applied Mathematics, 2013. 29: 141–150-141–150.
41. Marcolini, G., A. Bellin, Chiogna, G. Performance of the Standard Normal Homogeneity Test for the homogenization of mean seasonal snow depth time series. International Journal of Climatology, 2017. 37(S1): 1267-1277. [https://rmets.onlinelibrary.wiley.com/doi/pdf/46100.1002/doi:https://doi.org/10.1002/joc.4977.
42. Ward, J.H., Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 1963. 58(301): 236-244. doi:10.1080/01621459.1963.10500845.
43. Yashwant, S. Sananse S. Comparisons of Different Methods of Cluster Analysis with Application to Rainfall Data. IJIRSET 2015. 4(11) 10861. doi:10.15680/IJIRSET.2015.0411087.
44. Luna Vera, J.A. Domínguez Mora, R.T. Un método para el análisis de frecuencia regional de lluvias máximas diarias: aplicación en los Andes bolivianos. Ingeniare. Revista chilena de ingeniería, 2013. 21: 111-124. doi:10.4067/S0718-33052013000100010.
45. Ilbay, M.L., Barragán R.Z., Lavado-Casimiro,W. Regionalization of precipitation, its aggressiveness and concentration in the Guayas river basin, Ecuador. La Granja, 2019. 30(2): 57. doi:https://doi.org/10.17163/lgr.n30.2019.06.
46. Hiez, G. and others, L'homogénéité des données pluviométriques. Cahiers ORSTOM, série Hydrologie, 1977. 14(02): 29–173.
47. Brunet-Moret, Y., Homogénéisation des précipitations. Bureau Central Hydrologique de l'ORSTOM á Paris, 1979.
48. Vauchel, P. Hydraccess: progiciel de gestion et d'exploitation de bases de données hydrologiques. in HYDROMED: séminaire international les petits barrages dans le monde méditerranéen : recueil des résumés. Paris (FRA) ; Tunis : IRD ; INRGREF, 1 p. multigr. Les Petits Barrages dans le Monde Méditerranéen : Séminaire International, Tunis (TUN), 2001/05/28-31.
49. Wang, J.-H., Hopke, P.K.; Hancewicz, T.M.; Zhang, S.L. Application of modified alternating least squares regression to spectroscopic image analysis. Analytica Chimica Acta, 2003. 476(1): 93-109. doi:https://doi.org/10.1016/S0003-2670(02)01369-7.

50. Bárdossy, A., Pegram, G. Infilling missing precipitation records - A comparison of a new copula-based method with other techniques. Journal of Hydrology, 2014. 519: 1162-1170. doi:https://doi.org/10.1016/j.jhydrol.2014.08.025.
51. Khosravi, G., Nafarzadegan, A.R.; Nohegar, A.; Fathizadeh, H.; Malekian, A. A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. Theoretical and Applied Climatology, 2015. 119(1): 33–42.
52. Natekin, A. Knoll, A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 2013. 7: 21-21. doi:10.3389/fnbot.2013.00021.
53. Ma, L., Zhang, G., Lu, E. Using the Gradient Boosting Decision Tree to Improve the Delineation of Hourly Rain Areas during the Summer from Advanced Himawari Imager Data. Journal of Hydrometeorology, 2018. 19(5):761-776. doi:10.1175/JHM-D-17-0109.1..
54. Breiman, L., Random Forests. Machine Learning, 2001. 45(1):5–32. doi:10.1175/JHM-D-17-0109.1.
55. James, G., Witten, D.; Hastie, T.; Tibshirani, R. An introduction to statistical learning. 2013. 112.: Springer.
56. Bellido-Jiménez, J.A., Estévez J., and García-Marín A.P., New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. Agricultural Water Management, 2021. 245: p. 106558.
57. Bellido-Jiménez, J.A., Estévez, J. and García-Marín A.P., Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. Environmental Sciences Proceedings, 2021. 4(1): 19.
58. Gómez Guerrero, J.S., Aguayo Arias, M.I. Evaluación de desempeño de métodos de relleno de datos pluviométricos en dos zonas morfoestructurales del Centro Sur de Chile. Investigaciones geográficas, 2019. doi:10.14350/rig.59837.
59. Guijarro, J. A., and Guijarro, M. J. Package 'climatol'. 2019, Online: doi:http://ftp5.gwdg.de/pub/misc/cran/web/packages/climatol/climatol.pdf (retrieved 20.04. 2020).
60. Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L., Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE, 2007. 50(3): 885–900-885–900.

CAPITULO III: Mapping of Areas Vulnerable to Flash Floods by Means of Morphometric Analysis with Weighting Criteria Applied

Marcelo Portuguez-Maurtua, José Luis Arumi, Alejandra Stehr, Octavio Lagos, Eduardo Chavarri-Velarde and Daniela Rivera-Ruiz.

3.1 Resultado clave

- Se realizó el análisis morfométrico en 11 subcuencas, considerándose 15 parámetros para cada subcuenca, las cuales fueron agrupadas en 3 aspectos; lineales, áreas y relieve
- Se asignaron ranking preliminar de prioridades para cada subcuenca, en función de su capacidad de influir directa o indirectamente a la susceptibilidad a las inundaciones repentinas
- Determinación de ranking final que fueron calculados en función del modelo de suma ponderada, identificando zonas vulnerables a inundaciones repentinas ante eventos extremos

3.2 Resumen en extenso

Uno de los fenómenos naturales más devastadores, capaces de causar una gran destrucción en muy poco tiempo son las inundaciones repentinas, producidas por las fuertes lluvias estacionales, que se caracterizan por su alta velocidad y poder destructivo. La ocurrencia de este desastre natural se presenta con mayor frecuencia provocada por eventos pluviométricos extremos (cambio climático), que se producen durante un corto periodo de tiempo en un área relativamente pequeña, dando lugar a una descarga excesiva, deslizamientos de tierra y flujos de lodo. Este fenómeno se mide más a menudo por los cambios en las variables climáticas primarias, como la temperatura y la precipitación

La predicción y el control de estos eventos han sido muy difíciles, debido a la naturaleza altamente dinámica del clima y a su repentina aparición, en estas circunstancias, es necesario poner mayor énfasis en los estudios de prevención y protección en zonas vulnerables a desastres naturales. En ausencia de datos hidrológicos, la caracterización morfométrica puede proporcionar información significativa sobre medidas preventivas contra las inundaciones repentinas. Estudios de priorización de las subunidades hidrográficas, realizada en base al

análisis morfométrico ha permitiendo a los planificadores y responsables políticos en la preparación de planes de gestión.

Además, los sensores remotos y los Sistemas de Información Geográfica (SIG), actualmente se presentan como herramientas ideales para el análisis morfométrico a partir del tratamiento y cuantificación de datos topográficos. Diversos estudios han evaluado la susceptibilidad a las inundaciones repentinas en cuencas hidrográficas. Sin embargo, las diversas metodologías aplicadas se diferencian entre sí, en la cantidad y el tipo de parámetro usado en su análisis. El objetivo fue identificar zonas vulnerables a inundaciones repentinas en base a la caracterización de los parámetros morfométricos (lineales, áreas y relieve), empleando el análisis de suma ponderada (WSA) basado en una matriz de correlación estadística y fijar categoría de prioridad para cada unidad hidrográfica.

La cuenca en estudio se subdividió en 11 subcuencas y 15 parámetros morfométricos fueron seleccionados. La categoría de priorización (muy alta, alta y moderada) de cada subcuenca se asignaron en función del valor del factor compuesto y este obtenido mediante WSA. Los resultados de este análisis mostraron que el 26.08% de la superficie total se encuentra bajo riesgo de inundación repentina muy alta (subcuencas 3, 9 y 11), 38.46% se encuentran bajo riesgo de inundación repentina alta (subcuencas 5, 7, 8 y 10) y 35.45% se encuentran bajo riesgo de inundación repentina moderada. Además, se muestra como alternativa en ausencia de datos hidrológicos para identificar zonas de riesgo a inundaciones repentinas analizando las características morfométricas.

water

MDPI

Article

Mapping of Areas Vulnerable to Flash Floods by Means of Morphometric Analysis with Weighting Criteria Applied

Marcelo Portuguez-Maurtua [1,2,3,*], Jose Luis Arumi [2,4], Alejandra Stehr [5], Octavio Lagos [2,4], Eduardo Chávarri-Velarde [3] and Daniela Rivera-Ruiz [4]

[1] Doctoral Program in Water Resources and Energy for Agriculture, Universidad de Concepcion, Av. Vicente Mendez 595, Chillan 3812120, Chile
[2] CRHIAM Water Research Center, Universidad de Concepción, Victoria 1295, Concepción 4070386, Chile.
[3] Water Resources Department, College of Agricultural Engineering, Universidad Nacional Agraria La Molina Av. La Molina s/n, Lima 15024, Peru.
[4] Water Resources Department, College of Agriculture Engineering, Universidad de Concepción, Av. Vicente Mendez 595, Chillan 3812120.
[5] Facultad de Ingeniería, Departamento de Ingeniería Civil, Universidad de Concepción, Concepción 4070386, Chile.
* Correspondence: mportuguez@lamolina.edu.pe; Tel.: +51-1-949-377-610

Abstract: Flash floods, produced by heavy seasonal rainfall and characterized by high speeds and destructive power, are among the most devastating natural phenomena and are capable of causing great destruction in very little time. In the absence of hydrological data, morphometric characterization can provide important information on preventive measures against flash floods. A priority categorization of hydrographic units in the Cañete River basin was carried out using morphometric analysis together with a weighted sum analysis (WSA) based on a statistical correlation matrix. The delineation of the drainage network was performed based on Digital Elevation Model (DEM) data from the Shuttle Radar Topography Mission (SRTM). The Cañete River basin was subdivided into 11 sub-basins, and 15 morphometric parameters were selected. The priority category (very high, high, and moderate) of each sub-basin was assigned according to the value of the composite factor obtained through WSA. The results of this analysis showed that 26.08% of the total area is under a very high flash flood risk (sub-basins 3, 9, and 11), 38.46% is under a high flash flood risk (sub-basins 5, 7, 8, and 10), and 35.45% is under a moderate flash flood risk. This study concludes that flash floods predominate in sub-basin 3 and that downstream areas present characteristics of river flooding (sub-basins 9 and 11).

Keywords: flash flood; morphometric parameter; morphometric characterization; weighted sum analysis; basin prioritization

Citation: Portuguez-Maurtua, M.; Arumi, J.L.; Stehr, A.; Lagos, O.; Chávarri-Velarde, E; Rivera-Ruiz, D. Mapping of Areas Vulnerable to Flash Floods by Means of Morphometric Analysis with Weighting Criteria Applied. *Water* **2023**, *15*, 1053 https://doi.org/10.3390/w15061053

Academic Editor: Gwo-Fong Lin

Received: 27 January 2023
Revised: 7 March 2023
Accepted: 7 March 2023
Published: 10 March 2023

1. Introduction

Flash floods are among the most devastating natural phenomena and are capable of causing great destruction in very little time [1]. These natural disasters are most often caused by extreme rainfall events (climate change), which occur during a short time in a relatively small area, resulting in excessive discharge, landslides, and mudflows [2]. Climate change is one of the greatest threats to the entire world, affecting Earth's natural balance and ecosystems and disrupting all life [3]. This

phenomenon is often measured by changes in primary variables such as temperature and precipitation [4]. The prediction and control of flash floods have been very difficult due to the highly dynamic nature of the climate and their sudden appearance; under these circumstances, a greater emphasis on prevention and protection studies in areas vulnerable to natural disasters is necessary [2].

In the absence of hydrological data, morphometric analysis can provide important information on the hydrological characteristics of a basin [5,6]. Quantitative morphometric assessment of a basin is a very effective method for interpreting various aspects of its drainage network and evaluating its hydrological behavior [7–9]. Studies on basin sub-unit prioritization that have been conducted based on morphometric analysis have aided planners and decision-makers in the development of management plans [10–15]. Thus, to achieve good drainage basin management, it is necessary to study their morphometry [5,7].

Various studies have evaluated susceptibility to flash floods in drainage basins based on an analysis of morphometric parameters, and quantitative assessment of the morphometric characteristics of a basin has allowed its hydrological response behavior to be defined [1,9,16–20]. Remote sensing and Geographic Information Systems (GIS) are ideal tools for morphometric analysis based on the treatment and quantification of topographic data [7,9,10,21–25]. Recent publications on basin prioritization have used various methodologies such as principal component analysis (PCA) [26–29], weighted sum analysis (WSA), also known as multicriteria analysis [9,16,18,22–24,28], simple additive weighting (SAW), and the technique for order of performance by similarity to ideal solution (TOPSIS) [30]. All these methodologies, together with statistical techniques, have allowed ranked priority scores to be assigned to each sub-basin according to its relationship with flash flood risk [20]. However, there are differences among the various applied methodologies in the quantity and types of parameters used in their analyses.

This investigation aims to identify zones vulnerable to flash floods based on the characterization of morphometric parameters (linear aspects, area, and relief), use weighted sum analysis (WSA) based on a statistical correlation matrix, and set a priority category for each hydrographic unit. In addition, in the absence of hydrological data, the alternative is to use topographic data as an input to identify flash flood risk zones by analyzing morphometric characteristics. The methodology adopted in this study is to use GIS tools, programming, and statistical analysis in mapping flash flood risk zones. The results show morphometric characteristics of 11 fourth-order sub-basins with very high, high, and moderate susceptibility to flash floods. The article is organized as follows: Section 2 presents information on the location of the study area, the morphometric parameters considered, the dataset, and the methodology. Section 3 presents and discusses the results. Finally, the conclusions of this study are described in Section 4.

2. Materials and Methods

2.1. Study Area

The Cañete River basin is part of the Peruvian Pacific slope, located between geographic coordinates 11°58′ and 13°10′ south and 76°25′ and 75°30′ west, in the Cañete and Yauyos provinces, in the Department of Lima, Perú. The river basin covers an area of 6192 km². It has a maximum altitude of 5800 masl in the central Andes, and the river flows into the Pacific Ocean at 0 masl. (Figure 1). The 79.5% of the basin located above 2500 masl is classified as wet. The Cañete River originates at Ticllacocha Lake and is located in the foothills of the Ticlla and Pichahuarco ranges at an altitude of 4429 masl. After flowing 235.67 km, with an average slope of 1.85%, it empties into the Pacific Ocean. Along its path, the Cañete River receives the contributions of numerous tributaries, including, on its right bank, the Miraflores, Yauyos, Huantuya (Carania), and Aucampi rivers and, on its left bank, the Alis, Laraos, Huantán, Tupe, Cacra, and Huangascar rivers, primarily [31,32]

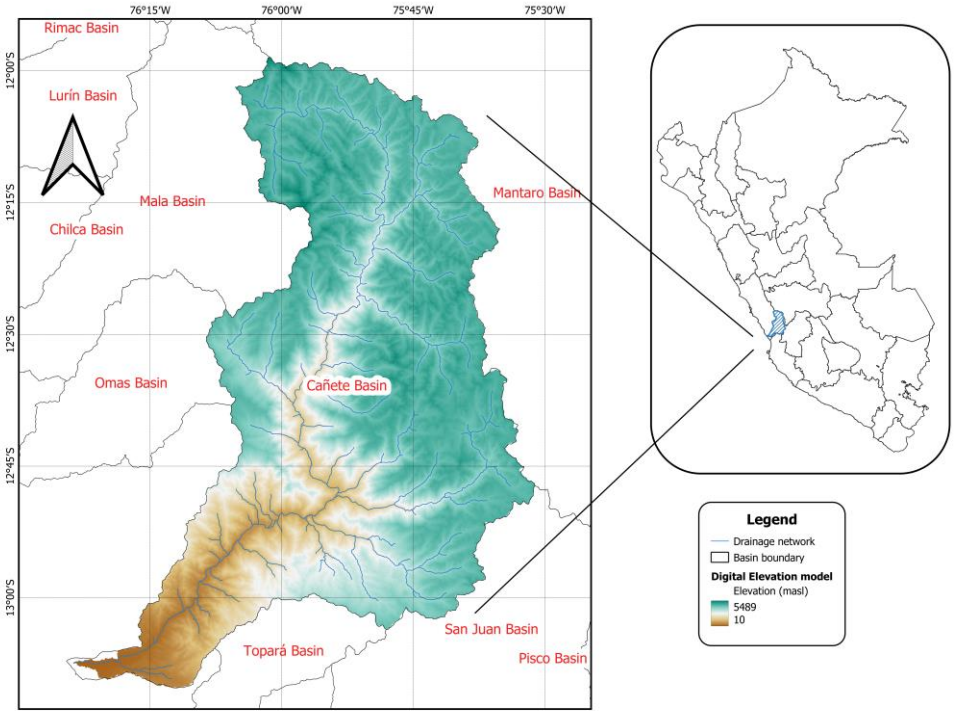


Figure 1. Map showing the location of the Cañete River drainage system.

2.2. Morphometric Parameters

These indicators are necessary to understand and estimate the hydrological and morphological characteristics of a basin; they also allow an interpretation of linear aspects, area, and relief of a drainage basin [1,14,15,20,33–35]. Remote sensing and Geographic Information Systems (GIS) are effective tools for delineating and understanding the morphometry of any drainage basin [7,9,35,36].

The morphometric parameters used in the study are presented in Tables 1–3.

Table 1. Linear morphometric parameters and formulas with references.

Item	Morphometric Parameter	Unit	Formula and Definition	Reference
1	Stream order (u)	Dimensionless	Hierarchical range	[37]
2	Stream length (L_u)	Km	Stream length	[38]
3	Basin length (L_b)	Km	$L_b = 1.312 \times A^{0.568}$ L_b = basin length (km) A = basin area (km²).	[39,40]

Table 2. Morphometric parameters related to area and formulas with references.

Item	Morphometric Parameter	Unit	Formula and Definition	Reference
1	Basin area (A)	Km²	Estimated in GIS	
2	Basin perimeter *(P)*	Km	Estimated in GIS	[41]
3	Stream frequency or flow frequency (F_s)	(Km^{-2})	$F_s = \Sigma N_u/A$ N_u = total number of stream segments of order *"u" and* A = basin area (km²)	[42]
4	Drainage density *(D_d)*	(Km^{-1})	$D_d = \Sigma L/A$ L = total stream length; A = basin area	[42]
5	Form factor *(F_f)*	Dimensionless	$F_f = A/L_b^2$ A = basin area L_b = basin length	[42]
6	Circularity ratio *(C_r)*	Dimensionless	$C_r = 4\pi A/P^2$ A = basin area (km²), P = basin perimeter (km)	[43]
7	Texture ratio *(T_r)*	Dimensionless	$Tr = N_l/P$ N_l = total number of first order streams P = basin perimeter	[38]
8	Elongation ratio *(E_r)*	Dimensionless	$E_r = 2\sqrt{(A/\pi)}/L_b$ A = basin area $\pi = 3.14$ L_b = basin length	[41]
9	Shape factor *(S_f)*	Dimensionless	$S_f = L_b^2/A$	[42]

Table 3. Morphometric parameters related to relief and formulas with references.

Item	Morphometric Parameter	Unit	Formula and Definition	Reference
1	Basin relief *(R)*	Meters	$R = H - h$ R = basin relief, H = maximum elevation in meters H = minimum elevation in meters	[41]
2	Relief ratio *(R_r)*	Dimensionless	$R_r = H/L_b$ R_r = relief ratio, H = basin relief, L_b = basin length	[41]
3	Average slope *(A_s)*	Degrees	Estimated in GIS	

2.3. Data Used

In this study, Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) data were used to delineate the basin, extract the drainage/stream network, and subdivide the catchment area into sub-

basins. The SRTM V3 (SRTM Plus) product is provided by the National Aeronautics and Space Administration (NASA) and has a resolution of 1 arc-second (approximately 30 m) [44].

The DEMs were downloaded from Google Earth Engine (GEE), a cloud-based platform that facilitates access to a catalog of petabytes of publicly available geospatial data, including satellite and aerial images, environmental, meteorological, climate, and topographic variables, land cover, etc. [45–47]. GEE is accessed using an application programming interface (API) accessible by the internet and an associated web-based integrated development environment (IDE) that allows the quick creation of prototypes and visualization of results [45,47–49]. In this study GEE was accessed using the Kaggle web platform, which offers a customizable Jupyter Notebooks interface, using Python programming language [50]. A rectangular region that covers the entire Cañete River basin was downloaded. Data Availability shows the codes that were used.

2.4. Methodology

The methodology adopted in this study is described by the following steps: (1) Extraction of drainage network and sub-basin delineation. (2) Morphometric analysis using the QGIS tool. (3) Assignment of preliminary ranking of sub-basin priority. (4) Weighted sum analysis and final ranking, y. (5) Mapping of areas vulnerable to flash floods. Figure 2 presents the methodological diagram of this research.

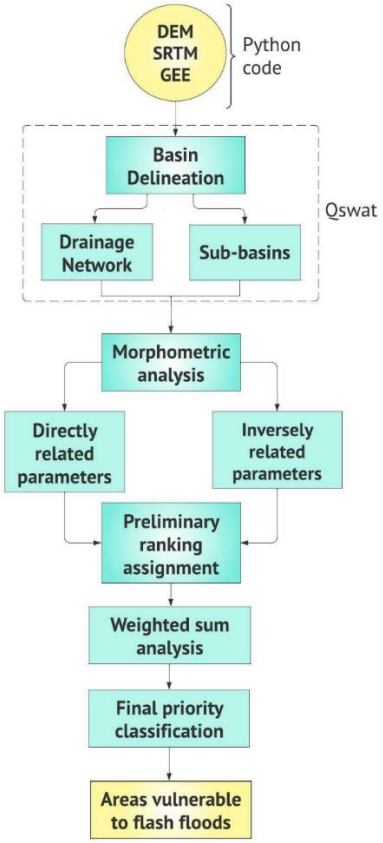


Figure 2. Methodological diagram for mapping of areas vulnerable to flash floods.

2.4.1. Extraction of Drainage Network and Sub-Basin Delineation

The drainage network extraction and basin delineation were carried out following DEM processing methods (cell filling, flow direction, flow accumulation, *stream order* definition, stream segmentation, and basin delineation) using the QSWAT complement in QGIS 3.16 [51,52]. To define the streams in the study area, a threshold value corresponding to 1% (192,777 cells) of the DEM cell count was used. The Cañete River basin was subdivided into 11 sub-basins, designated as SC-1 to SC-11, and the entire drainage network was classified as a fourth-order catchment basin using the Strahler method [37].

2.4.2. Morphometric Analysis

A morphometric analysis is necessary to ascertain the sub-basin prioritization. This process is advantageous as the derived basin variables are in the form of ratios or dimensionless numbers, providing an effective comparison independent of scale [1,14,14,20,24,34,35]. Morphometric analysis offers a complete representation of the drainage network, geometry, and topography of the basin, allowing an interpretation of linear aspects, area, and relief of the basin, respectively [9].

Each of these parameters can directly or indirectly influence the surface runoff of the drainage basin [9], facilitating the understanding of how they influence flash flood risk. Tables 1–3 show a list of morphometric parameters that describe the dimensionless and multidimensional characteristics of the basin, which are grouped into 3 aspects. First, linear morphometric characterization: *stream order (u)*, *stream length* (L_u), and *basin length* (L_b), shown in Table 1 [37–39]. Second, morphometric characterization of area: *basin area (A)*, *basin perimeter* (*P*), stream frequency, also known as *flow frequency* (F_s), *drainage density* (D_d), *form factor* (F_f), *circularity ratio* (C_r), drainage texture (T_r), *elongation ratio* (E_r), and *shape factor* (S_f), shown in Table 2 [38,41–43]. Finally, *basin relief (R)*, *relief ratio* (R_r), and *average slope* (A_s) are morphometric relief parameters, shown in Table 3 [41].

2.4.3. Preliminary Ranking of Sub-Basin Priority

The preliminary ranking (PR) approach was used, in which morphometric parameters were divided into two groups according to their ability to (directly or indirectly) influence the adversity of the morphometric parameter conditions or degree of susceptibility to flash floods [16,18,19,22,24]. The first group of parameters (capable of direct influence) consisted of: *basin relief (R)*, *relief ratio* (R_r), *drainage density* (D_d), stream frequency (F_s), *circularity ratio* (C_r), drainage texture (T_r), and *average slope* (A_s). The second group of parameters (capable of indirect influence) consisted of: *elongation ratio* (E_r), *form factor* (F_f), and *shape factor* (S_f).

The preliminary rankings of the parameters in the first group were assigned in such a way that for each parameter the sub-basin with the highest value was classified as 1, that with the next highest value was classified as 2, and so on for the remaining sub-basins. The opposite was done for the parameters in the second group, assigning rankings in such a way that for each parameter the sub-basin with the lowest value was placed in position 1, that with the next lowest value was placed in position 2, and so on for the remaining sub-basins [16,18,22].

2.4.4. Weighted Sum Analysis and Final Ranking

Weighted sum analysis (WSA) is a well-known method that provides consistency in addressing complicated problems to compare land surface processes in related entities such as drainage basins [26]. WAS, also known as multi-criteria decision making, is widely used to select the best alternatives among multiple options. Based on preliminary rankings, the correlation matrix and correlation coefficients were calculated [16,18]. The composite parameters (WSA_{cp}) were calculated using the following equation (Equation (1)) [9,16,22,26]:

$$WSA_{cp} = PR_{p1} \times W_{p1} + PR_{p2} \times W_{p2} + \cdots + PR_{pn} \times W_{pn} \quad (1)$$

where WSA_{cp} = composite parameter used for weighted sum analysis; PR = preliminary priority ranking of each morphometric parameter ($p1$, $p2$, … pn); and W = weights of the morphometric parameters obtained by means of the correlation matrix, which was calculated using the following equation (Equation (2)):

$$Parameter\ weights\ (W) = \frac{Correlation\ coefficient\ sum}{Correlation\ total} \quad (2)$$

Therefore, sub-basin priority was assigned by taking the $WSA_{cp}(+)$ values (capable of direct influence) corresponding to the first group and subtracting the $WSA_{cp}(-)$ values (capable of direct influence) corresponding to the second group, following the equation (Priority = $WSA_{cp}(+) - WSA_{cp}(-)$). Subsequently, final ranking values were assigned, one for the lowest priority value, two for the following value, and so on. Finally, to categorize the priority type, the maximum category (very high) corresponds to the lowest priority levels [9,18,22]. Finally, a model was formulated to assess the final priority using the WSA_{cp} value of each parameter, as shown below:

$$Priorizacion = (R + R_r + D_d + F_s + C_r + T_r + A_s) - (E_r + F_f + S_f) \quad (3)$$

3. Results and Discussion

The identification of areas in drainage basins vulnerable to flash floods using morphometric parameters is considered one of the most effective methods to characterize various geohydrological properties of a basin [14,24]. Therefore, this study has used different morphometric parameters that govern the hydrological response of a basin to prioritize the sub-basins of the Cañete River in terms of their vulnerability to flash floods.

3.1. Morphometric Analysis of the Basin

The morphometric analysis was carried out for 11 sub-basins, and the drainage network of the entire study area was considered a fourth-order basin (Figure 3). The morphometric analysis covered 15 parameters for each sub-basin, which was necessary to determine its dimensions, shape and area, and the characteristics of the drainage network. The results reveal that SB-9 is the smallest, with an area of 31.49 km^2, while SB-11 is the largest, with an area of 1175.47 km^2 (Table 4). More than half of the sub-basins have areas greater than 400 km^2, with only SB-9 having an area smaller than 50 km^2. Most of the sub-basins are large, with areas greater

than 300 km². The area of a sub-basin directly affects its susceptibility to flash floods [16,19].

Table 4. Area, perimeter, and length of each sub-basin.

Sub-Basin	Perimeter (km)	Area (km²)	Length (km)
SB-1	197.09	943.30	64.20
SB-2	158.45	448.58	42.09
SB-3	137.14	364.68	37.42
SB-4	145.79	419.96	40.54
SB-5	162.71	586.76	49.03
SB-6	118.83	324.56	35.02
SB-7	183.65	596.71	49.50
SB-8	204.97	616.61	50.43
SB-9	30.01	31.49	9.31
SB-10	167.51	517.64	45.66
SB-11	281.88	1175.47	72.75

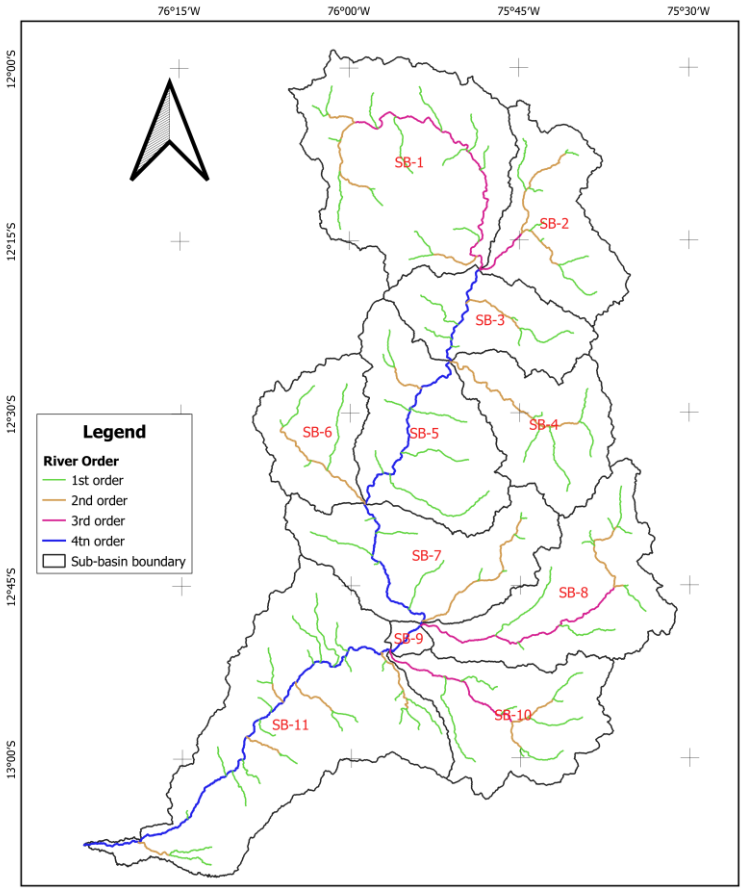


Figure 3. Sub-basins and drainage network of the Cañete River.

Basin perimeter (P) values for the study area vary from 30.01 km (SB-9) to 281.88 km (SB-11). *Basin length (L_b)* is understood as the maximum length and is calculated from the farthest point on the basin border to the confluence point; the values vary from 9.31 km (SB-9) to 72.75 km (SB-11). The values are presented in Table 4.

Figure 3 and Table 5 show that there are 97 first-order streams, 41 second-order streams, 18 third-order streams, and 37 fourth-order streams. Table 5 also shows that the greatest number of streams was found in SB-11 (49), and the smallest number of streams was found in SB-9 (1).

Table 5. Number of streams in each sub-basin by stream order.

Stream Order (u)	SB-1	SB-2	SB-3	SB-4	SB-5	SB-6	SB-7	SB-8	SB-9	SB-10	SB-11
1st order	15	8	6	5	7	4	10	10	0	8	24
2nd order	5	6	2	4	1	3	4	5	0	3	8
3rd order	9	1	0	0	0	0	0	4	0	4	0
4th order	0	0	5	0	7	0	7	0	1	0	17
Total	29	15	13	9	15	7	21	19	1	15	49

Table 6 shows the lengths of the first- to fourth-order streams of the 11 sub-basins, determined using the GIS technique [51], which indicates that the total length of all streams varies from a minimum of 7.68 km (SB-9) to a maximum of 247.19 km (SB-11), and that the total length of all streams in the basin is 1113.75 km. The drainage length values show a significant interrelationship with drainage flow discharge and the erosion phase of the river basin [18]. In addition, it is seen that the five sub-basins with fourth-order streams (SB-3, SB-5, SB-7, SB-9, and SB-11) are possibly susceptible to flash floods [16,19]. Drainage length is a parameter that is directly related to the hydrological reaction of a watershed and its importance is approximately equal to half the reciprocal of the *drainage density* [19].

Table 6. Stream length in each sub-basin by stream order.

Stream Order (u)	SB-1 (km)	SB-2 (km)	SB-3 (km)	SB-4 (km)	SB-5 (km)	SB-6 (km)	SB-7 (km)	SB-8 (km)	SB-9 (km)	SB-10 (km)	SB-11 (km)
1st order	88.01	35.04	32.84	36.10	66.93	29.11	41.07	54.75	0.00	59.13	128.42
2nd order	31.80	29.39	11.92	30.52	7.66	22.31	29.79	17.16	0.00	14.11	39.91
3rd order	54.55	10.43	0.00	0.00	0.00	0.00	0.00	41.79	0.00	28.51	0.00
4th order	0.00	0.00	19.35	0.00	36.67	0.00	29.95	0.00	7.68	0.00	78.86
Total	174.36	74.86	64.11	66.62	111.26	51.43	100.80	113.70	7.68	101.74	247.19

Drainage density (D_d) and stream frequency (F_s) were calculated for all sub-basins (Table 7). In the analysis of the sub-basin and drainage network map (Figure 3), the Cañete River basin was found to be a fourth-order basin, with a dendritic drainage pattern. D_d and F_s are significant parameters that contribute to the hydrological responses in each sub-basin [19]. The F_s and D_d values in each sub-basin are directly proportional; therefore, a greater stream number corresponds to a greater stream length. The high value of F_s in sub-basins 3, 7, and 11 indicates that they produce more runoff in comparison to the other sub-basins. Meanwhile, high D_d values are observed in sub-basins 9 and 11, indicating a well-developed network, which is conducive to high runoff concentrations that give rise to flash floods. The low value of D_d in sub-basins 4 and 6 indicates that they have highly permeable subsoil material, with dense plant cover and a low

relief [53]. Therefore, high values of D_d and F_s are likely to increase susceptibility to high surface runoff and flash floods [19].

Table 7. Morphometric parameters that represent area and form aspects of the sub-basins.

Sub-Basin	Drainage Density (D_d)	Stream Frequency (F_s)	Circularity Ratio (C_r)	Texture Ratio (T_r)	Elongation Ratio (E_r)	Form Factor (F_f)	Shape Factor (S_f)
SB-1	0.185	0.031	0.305	0.076	0.540	0.229	4.369
SB-2	0.167	0.033	0.225	0.050	0.568	0.253	3.949
SB-3	0.176	0.036	0.244	0.044	0.576	0.260	3.840
SB-4	0.159	0.021	0.248	0.034	0.570	0.255	3.914
SB-5	0.190	0.026	0.279	0.043	0.558	0.244	4.096
SB-6	0.158	0.022	0.289	0.034	0.580	0.265	3.779
SB-7	0.169	0.035	0.222	0.054	0.557	0.244	4.106
SB-8	0.184	0.031	0.184	0.049	0.556	0.242	4.124
SB-9	0.244	0.032	0.439	0.000	0.680	0.363	2.752
SB-10	0.197	0.029	0.232	0.048	0.562	0.248	4.027
SB-11	0.210	0.042	0.186	0.085	0.532	0.222	4.502

The *circularity ratio* (C_r) is influenced by *stream length* and frequency, geological structures, land use and cover, climate, relief, and basin slope [35]. For a perfectly circular basin, the value of the *circularity ratio* is 1 [43]. In this study, the C_r values of the sub-basins range from 0.184 to 0.439 (Table 7); all the values are less than 1, which indicates that the area is characterized by an elevated relief and that the drainage system is structurally controlled.

Infiltration capacity is the only important factor that controls the *texture ratio* (T_r) [9]. The texture ratio in the sub-basins range from 0.00 to 0.085 (Table 7). The highest relief values are attributed to the pronounced slopes of the sub-basins, which makes them more prone to flood risk [16,24]. Meanwhile, the lowest Tr values indicate that the basin is flat, with slope values close to zero degrees [35].

The *elongation ratio* (E_r) is the measure of basin dimensions or basin shape [41]. According to [37], a sub-basin with an Er value above 0.9 is classified as circular, those with Er values between 0.9 and 0.8 are classified as oval, those with Er values between 0.7 and 0.8 are classified as less elongated, those with Er values between 0.5 and 0.7 are classified as elongated, and those with Er values below 0.5 are classified as more elongated. In this study, the Er values vary between 0.532 and 0.680, indicating that the sub-basins are classified as elongated.

The *form factor* (F_f) is a dimensionless ratio of the area of a basin to the square of its length [42] and can be effectively related to flood occurrence, erosion intensity, and sediment transport capacity in a basin [1]. The lower the *form factor* value, the longer the basin. Basins with a high *form factor* have high, short-duration maximum flows, while those with a low *form factor* have lower, longer-duration maximum flows [16,35]. In this study the sub-basins have Ff values ranging from 0.222 to 0.363 (Table 7), indicating that they have elongated shapes and suggesting lower, longer-duration maximum flows.

The *shape factor* (S_f) is a dimensionless ratio of the square of the length of a basin to its area [42]. Its values indicate the opposite of those of Ff,

with a maximum Sf value corresponding to a minimum Ff value. The Sf values for each sub-basin vary between 2.752 (SB-9) and 4.502 (SB-11) (Table 7).

Basin relief (R) is defined as the change between its highest- and lowest-elevation points [9,18]. Table S1 (Supplementary Material) details the minimum and maximum height values and other statistics for each sub-basin. R shows the potential energy of a drainage basin, which significantly influences the channel gradient and aspect and landform evolution; therefore, it directly affects surface runoff, flood patterns, and sediment transport [9,18,24]. The highest and lowest R values are found in SB-11 (4420.00 m) and SB-2 (2061.00 m) (Table 8). Maximum R values are indicative of the potential energy of a given sub-basin to move water and sediment along the slope.

Table 8. Morphometric parameters that represent sub-basin relief aspects.

Sub-Basin	Basin Relief (*R*) (m)	Relief Ratio (R_r)	Average Slope (A_s) (°)
SB-1	2906.00	0.045	22.590
SB-2	2061.00	0.049	20.816
SB-3	3287.00	0.088	27.833
SB-4	2844.00	0.070	22.800
SB-5	4127.00	0.084	27.959
SB-6	3309.00	0.094	19.485
SB-7	4063.00	0.082	27.023
SB-8	4156.00	0.082	22.011
SB-9	2405.00	0.258	34.042
SB-10	3963.00	0.087	21.225
SB-11	4420.00	0.061	23.243

The *relief ratio* (R_r) is estimated as the ratio of *basin relief* to *basin length*. According to [41], there is a correlation between the hydrological characteristics of a basin and the *relief ratio*. Therefore, it is presented as an indicator of the intensity of the erosion process in the basin [9,16]. High R values are characteristic of mountainous regions. The R values for all the sub-basins (Table 8) are between 0.045 (SB-1) and 0.258 (SB-9). *Average slope* refers to the amount of inclination of the physical feature or the topographic form with respect to the horizontal. Table S2 (Supplementary Material) details the slope raster statistics in degrees for each sub-basin. Slope analysis is very important in morphometric studies. Slope elements are, in turn, controlled by climate-morphogenic processes in areas with rocks of varying resistance [1,54]. The average slopes of the sub-basins vary from 19.485° (SB-6) to 34.042° (SB-9) (Table 8). The pronounced slopes also favor a faster movement of surface runoff.

3.2. Assignment of Preliminary Sub-Basin Priority Rankings

The seven morphometric parameters (R, R_r, D_d, F_s, C_r, T_r, and A_s) are directly proportional to soil degradation and water factors. The rankings were assigned from greatest to lowest priority, i.e., rank 1 for the sub-basin with the maximum parameter value and rank 11 for the sub-basin with the minimum parameter value. For example, parameter R (Table 8) with the maximum value of 4420.0 m (SB-11) was assigned the highest priority

(rank 1), the next descending value was assigned rank 2, and this went up to the minimum parameter value of 2061.0 m (SB-2), which was assigned the lowest priority (rank 11). The assignment of the rankings for the six parameters R_r, D_d, F_s, C_r, T_r, and A_s was assigned in a similar way as explained above, the results are shown in Table 9.

Table 9. Preliminary priority rankings.

Sub-Basin	R	R_r	D_d	F_s	C_r	T_r	A_s	E_r	F_f	S_f
SB-1	8	11	5	7	2	2	7	2	2	10
SB-2	11	10	9	4	8	4	10	7	7	5
SB-3	7	3	7	2	6	7	3	9	9	3
SB-4	9	8	10	11	5	9	6	8	8	4
SB-5	3	5	4	9	4	8	2	5	5	7
SB-6	6	2	11	10	3	10	11	10	10	2
SB-7	4	7	8	3	9	3	4	4	4	8
SB-8	2	6	6	6	11	5	8	3	3	9
SB-9	10	1	1	5	1	11	1	11	11	1
SB-10	5	4	3	8	7	6	9	6	6	6
SB-11	1	9	2	1	10	1	5	1	1	11

The three remaining parameters (E_r, F_f, and S_f) have an inverse relationship with soil degradation and water factors. Rankings were assigned from lowest to highest priority, i.e., rank 1 for the sub-basin with the lowest parameter value and rank 11 for the sub-basin with the highest parameter value. For example, parameter E_r (Table 7) with a minimum value of 0.532 (SB-11) was assigned the highest priority (rank one), and the next higher value was assigned rank two; the maximum parameter value of 0.680 (SB-9) was assigned the lowest priority (rank 11). The assignment of the rankings for the next two parameters (F_f and S_f) were similarly assigned, the results are shown in Table 9.

A correlation matrix of the 10 morphometric parameters capable of directly or indirectly influencing the vulnerability to flash floods is presented in Table 10. It was estimated based on the preliminary priority rankings (Table 9), and the correlation coefficients were obtained in a process carried out using Python code (Data Availability shows the codes used). The statistical correlation matrix shows that *basin relief* has a positive correlation with *relief ratio*, *drainage density*, stream frequency, *texture ratio*, *average slope*, *elongation ratio*, and *form factor* and an inverse correlation with *circularity ratio* and *shape factor*. The *relief ratio* has a positive correlation with *basin relief*, *drainage density*, *circularity ratio*, *average slope*, and *shape factor* and an inverse correlation with stream frequency, *texture ratio*, *elongation ratio*, and *form factor*. The *drainage density*, stream frequency, *circularity ratio*, *texture ratio*, *average slope*, *elongation ratio*, *form factor*, and *shape factor* correlations were also calculated and are shown in Table 10.

Table 10. Correlation matrix of the morphometric parameters.

	R	R_r	D_d	F_s	C_r	T_r	A_s	E_r	F_f	S_f
R	1.000	0.009	0.282	0.173	−0.564	0.373	0.082	0.636	0.636	−0.636
R_r	0.009	1.000	0.136	−0.155	0.345	−0.782	0.255	−0.709	−0.709	0.709
D_d	0.282	0.136	1.000	0.327	0.064	0.145	0.518	0.282	0.282	−0.282
F_s	0.173	−0.155	0.327	1.000	−0.482	0.564	0.318	0.282	0.282	−0.282
C_r	−0.564	0.345	0.064	−0.482	1.000	−0.600	0.227	−0.545	−0.545	0.545
T_r	0.373	−0.782	0.145	0.564	−0.600	1.000	−0.173	0.873	0.873	−0.873
A_s	0.082	0.255	0.518	0.318	0.227	−0.173	1.000	−0.064	−0.064	0.064
E_r	0.636	−0.709	0.282	0.282	−0.545	0.873	−0.064	1.000	1.000	−1.000
F_f	0.636	−0.709	0.282	0.282	−0.545	0.873	−0.064	1.000	1.000	−1.000
S_f	−0.636	0.709	−0.282	−0.282	0.545	−0.873	0.064	−1.000	−1.000	1.000
Sum	1.991	0.100	2.755	2.027	−0.555	1.400	2.164	1.755	1.755	−1.755
Weight (w)	0.171	0.009	0.237	0.174	−0.048	0.120	0.186	0.151	0.151	−0.151

The final weights for each parameter were calculated by dividing the correlation coefficient sum of each parameter by the overall correlation total (Equation (2)). The final weight of R (weight = 0.171) was obtained though division (1.991/11.636). The final weights of the other morphometric parameters were obtained in a similar manner; the results are shown at the bottom of Table 10. The value 11.636 was obtained by summing the sum row (penultimate row Table 10).

3.3. Final Ranking Using Weighted Sum Analysis

The WSA_{cp} values of the morphometric parameters (Equation (1)) were calculated according to their importance using the weighted sum model. This process used the preliminary priority values (Table 9) and final weights of each morphometric parameter (Table 10). For example, the WSA_{cp} value of R (SB-1) was obtained by multiplying 8 × 0.171 = 1.37, and similarly for the following WSA_{cp} values of R of the other sub-basins; WSA_{cp} of R_r was multiplied 11 × 0.009 = 0.09. Therefore, for each parameter there is a corresponding weighting (w). The WSA_{cp} values are shown in Table 11.

Table 11. WSA_{cp} values in each sub-basin.

Sub-Basin	R	R_r	D_d	F_s	C_r	T_r	A_s	E_r	F_f	S_f
SB-1	1.37	0.09	1.18	1.22	−0.10	0.24	1.30	0.30	0.30	−1.51
SB-2	1.88	0.09	2.13	0.70	−0.38	0.48	1.86	1.06	1.06	−0.75
SB-3	1.20	0.03	1.66	0.35	−0.29	0.84	0.56	1.36	1.36	−0.45
SB-4	1.54	0.07	2.37	1.92	−0.24	1.08	1.12	1.21	1.21	−0.60
SB-5	0.51	0.04	0.95	1.57	−0.19	0.96	0.37	0.75	0.75	−1.06
SB-6	1.03	0.02	2.60	1.74	−0.14	1.20	2.05	1.51	1.51	−0.30
SB-7	0.68	0.06	1.89	0.52	−0.43	0.36	0.74	0.60	0.60	−1.21
SB-8	0.34	0.05	1.42	1.05	−0.52	0.60	1.49	0.45	0.45	−1.36
SB-9	1.71	0.01	0.24	0.87	−0.05	1.32	0.19	1.66	1.66	−0.15
SB-10	0.86	0.03	0.71	1.39	−0.33	0.72	1.67	0.90	0.90	−0.90
SB-11	0.17	0.08	0.47	0.17	−0.48	0.12	0.93	0.15	0.15	−1.66

Based on the results of the composite weighted sum WSA_{cp} of the different parameters (Table 11), a model was formulated to evaluate the

final priority ranking (equation 3). The parameters were categorized into two groups: first group WSA_{cp} $(+)$, parameters that have the ability to directly influence flash floods; second group, WSA_{cp} $(-)$, parameters that have the ability to indirectly influence flash floods. The priorities ranking (prioritization) is obtained by subtracting the two groups WSA_{cp} $(+)$ $-$ WSA_{cp} $(-)$; the sub-basin with the lowest composite value receives the highest priority (one), that with the next lowest composite value receives the second rank, and so on for the classification of each hydrological unit.

The final classification of priorities was carried out in such a way that the lowest composite factor value received priority rank 1, the next lowest value received priority rank 2, and so on for the 11 sub-basins. As observed in Table 12, the highest priority rank (1) was assigned to SB-9, followed by SB-3, SB-11, SB-5, SB-7, SB-10, SB-8, SB-2, SB-6, SB-4, and SB-1. Figure 4 shows the final map of priority classifications of the 11 studied sub-basins.

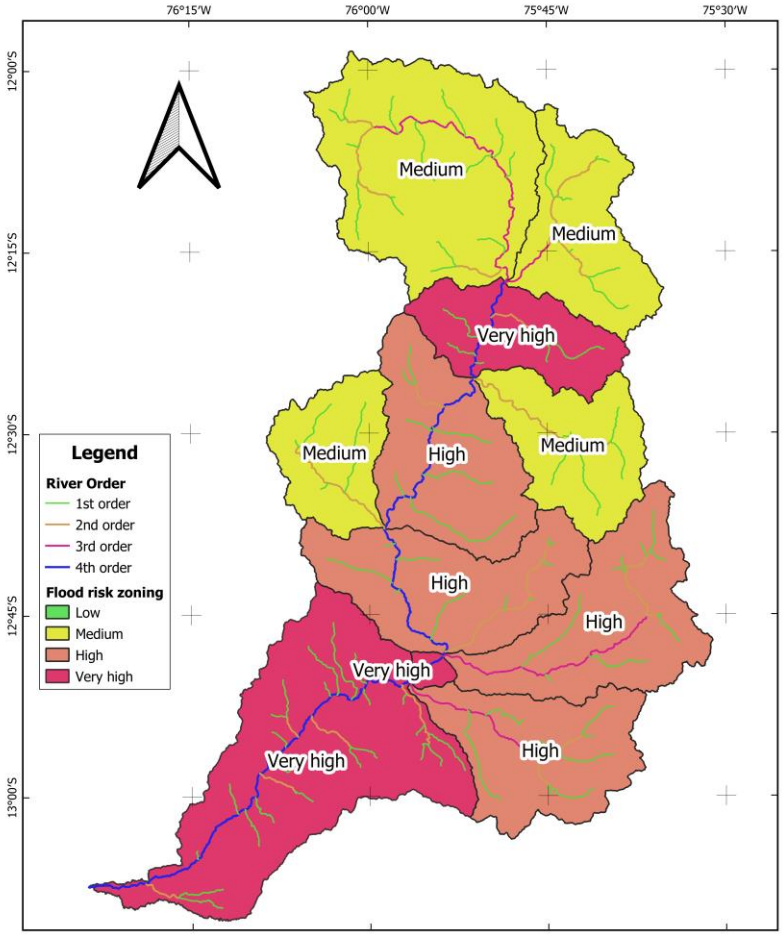


Figure 4. Map of zones vulnerable to flash floods in the Cañete basin.

Table 12. Final ranking and sub-basin priority areas.

Sub-Basin	WSA_{cp} (+)	WSA_{cp} (-)	Priority	Priority Level	Priority Type	Area (%)
SB-1	5.31	−0.90	6.22	11	Medium	15.65
SB-2	6.75	1.36	5.40	8	Medium	7.44
SB-3	4.34	2.26	2.08	2	Very high	6.05
SB-4	7.85	1.81	6.04	10	Medium	6.97
SB-5	4.21	0.45	3.76	4	High	9.74
SB-6	8.50	2.71	5.78	9	Medium	5.39
SB-7	3.84	0.00	3.84	5	High	9.90
SB-8	4.42	−0.45	4.88	7	High	10.23
SB-9	4.29	3.17	1.12	1	Very high	0.52
SB-10	5.06	0.90	4.15	6	High	8.59
SB-11	1.47	−1.36	2.83	3	Very high	19.51

From the results shown in Table 12, there are composite parameter values WSA_{cp} that have the ability to directly or indirectly influence the degree of susceptibility to flash floods. Based on the composite factor value, the 11 sub-basins of the Cañete River basin were classified into three priority categories: very high, high, and medium [9]. In Table 12 it can be seen that three sub-basins (SB-9, SB-3, and SB-11) are in the "very high" category, four sub-basins (SB-5, SB-7, SB-10, and SB-8) are in the "high" category, and four sub-basins (SB-2, SB-6, SB-4, and SB-1) are in the "medium" category. The final priority category map of the 11 sub-basins (Figure 4) shows that the "very high" category accounts for 26.08% of the total area, the "high" category for 38.46%, and the "medium" category for 35.45%.

These results show that 64.55% of the sub-basins (seven sub-basins) are in zones with a very high or high propensity to soil erosion, indicating that there are potential areas to carry out soil protection measures for efficient basin management and development. In addition, with the final priority classification of the 11 sub-basins (Figure 4), and considering the continuity of the flow wave or flood that starts in SB-3 and moves downstream and is produced by a maximum rainfall event over the central region of the sub-basin, SB-5 and SB-7 could also be considered very high categories, as they follow the continuity of SB-3. Therefore, these results will be of great help to policy-makers, planners, and managers to address vulnerable areas through specific action plans for flood risk reduction.

4. Conclusions

Through the morphometric characterization carried out by interpreting the linear, area, and relief aspects of a hydrological basin, it was possible to identify areas vulnerable to flash floods in the occurrence of extreme events, as a methodological technique in basins with an absence of hydrological data and to know the hydrological behavior in a basin. In addition, the important role of GIS tools and statistical approaches in developing research was shown.

The results show that SB-3, SB-9, and SB-11 are susceptible to floods and soil loss. Flash floods predominate in the upstream sub-basin (SB-3), while the downstream sub-basin (SB-9 and SB-11) present characteristics of river floods. Both are destructive in the study area, affecting the human

population and their residential units, farmland with standing crops, and infrastructure (highways, bridges, sewers, and water supply systems).

In addition, the sub-basins located in the middle part of the basin (SB-5, SB-7, SB-8, and SB-10) are categorized as high risk and the remaining sub-basins (SB-1, SB-2, SB-4, and SB-6) as medium risk. Thus, the Cañete River basin has sub-basins that present a very high, high, and moderate susceptibility to flash floods. Finally, it can be affirmed that the hydrological response and specifically the risk of flash floods and extreme events in a river basin depends on its morphometric characteristics. Therefore, this methodology is presented as an alternative to decision-makers in the application of suitable drainage basin management techniques in terms of soil and water conservation measures, allowing them to safeguard the study area and mitigate its degradation.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1; Table S1: DEM statistics for each sub-basin; Table S2: raster statistics of slope in grade for each sub-basin.

Author Contributions: Conceptualization, M.P.-M. and J.L.A.; methodology, M.P.-M.; software, M.P.-M.; validation, M.P.-M. and J.L.A.; formal analysis, M.P.-M. and D.R.-R.; investigation, M.P.-M., J.L.A., A.S., O.L., E.C.-V., and D.R.-R.; writing—original draft preparation, M.P.-M. and D.R.-R.; writing—review and editing, M.P.-M., J.L.A., A.S., O.L., E.C.-V., and D.R.-R.; visualization, M.P.-M., E.C.-V., and D.R.-R.; supervision, J.L.A. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the CRHIAM Water Research Center: ANID/FONDAP/15130015

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: CRHIAM Water Research Center: Project ANID/FONDAP/15130015, Josselyn Portuguez Contreras, and Mariam De la Cruz.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bisht, S.; Chaudhry, S.; Sharma, S.; Soni, S. Assessment of flash flood vulnerability zonation through Geospatial technique in high altitude Himalayan watershed, Himachal Pradesh India. Remote Sensing Applications: Soc. Environ. 2018, 12, 35–47. https://doi.org/10.1016/j.rsase.2018.09.001.
2. Cahyono, C.; Adidarma, W.K. Influence analysis of peak rate factor in the flood events' calibration process using HEC-HMS. Model. Earth Syst. Environ. 2019, 5, 1705–1722. https://doi.org/10.1007/s40808-019-00625-8.
3. Teng, F.; Huang, W.; Ginis, I. Hydrological modeling of storm runoff and snowmelt in Taunton River Basin by applications of HEC-HMS and PRMS models. Nat. Hazards 2018, 91, 179–199. https://doi.org/10.1007/s11069-017-3121-y.
4. Wang, N.; Lombardo, L.; Gariano, S.L.; Cheng, W.; Liu, C.; Xiong, J.; Wang, R. Using satellite rainfall products to assess the triggering conditions for hydro-morphological processes in different geomorphological settings in China. International Journal of Applied Earth Observation and Geoinformation 2021, 102, 102350. https://doi.org/10.1016/j.jag.2021.102350.
5. Prakash, K.; Rawat, D.; Singh, S.; Chaubey, K.; Kanhaiya, S.; Mohanty, T. Morphometric analysis using SRTM and GIS in synergy with depiction: A case study of the Karmanasa River basin, North central India. Appl. Water Sci. 2019, 9, 13. https://doi.org/10.1007/s13201-018-0887-3.

6. Perucca, L.P.; Esper Angilieri, Y. Morphometric characterization of del Molle Basin applied to the evaluation of flash floods hazard, Iglesia Department, San Juan, Argentina. Quat. Int. 2011, 233, 81–86. https://doi.org/10.1016/j.quaint.2010.08.007.
7. Rai, P.K.; Singh, P.; Mishra, V.N.; Singh, A.; Sajan, V.; Shahi, A.P. Geospatial approach for quantitative drainage morphometric analysis of Varuna river basin, India. J. Landsc. Ecol. 2019, 12, 1–25. https://doi.org/10.2478/jlecol-2019-0007.
8. Rai, P.K.; Chandel, R.S.; Mishra, V.N.; Singh, P.; Hydrological inferences through morphometric analysis of lower Kosi river basin of India for water resource management based on remote sensing data. Appl. Water Sci. 2018, 8, 15. https://doi.org/10.1007/s13201-018-0660-7.
9. Aher, P.; Adinarayana, J.; Gorantiwar, S. Quantification of morphometric characterization and prioritization for management planning in semi-arid tropics of India: A remote sensing and GIS approach. J. Hydrol. 2014, 511, 850–860. https://doi.org/10.1016/j.jhydrol.2014.02.028.
10. Shivhare, V.; Gupta, C.; Mallick, J.; Singh, C.K. Geospatial modelling for sub-watershed prioritization in Western Himalayan Basin using morphometric parameters. Nat. Hazards 2022, 110, 545–561. https://doi.org/10.1007/s11069-021-04957-6.
11. Abdeta, G.C.; Tesemma, A.B.; Tura, A.L.; Atlabachew, G.H. Morphometric analysis for prioritizing sub-watersheds and management planning and practices in Gidabo Basin, Southern Rift Valley of Ethiopia. Appl. Water Sci. 2020, 10, 158. https://doi.org/10.1007/s13201-020-01239-7.
12. Waiyasusri, K.; Chotpantarat, S. Watershed Prioritization of Kaeng Lawa Sub-Watershed, Khon Kaen Province Using the Morphometric and Land-Use Analysis: A Case Study of Heavy Flooding Caused by Tropical Storm Podul. Water 2020, 12, 94, 515–524.
13. Ahirwar, R.; Malik, M.S.; Shukla, J.P. Prioritization of Sub-Watersheds for Soil and Water Conservation in Parts of Narmada River through Morphometric Analysis Using Remote Sensing and GIS. J. Geol. Soc. India 2019, 94, 515–524.
14. Anees, M.T.; Abdullah, K.; Nawawi, M.; Rahman, N.N.N.A.; Ismail, A.Z.; Syakir, M.; Abdul Kadir, V. Prioritization of Flood Vulnerability Zones Using Remote Sensing and GIS for Hydrological Modelling. Irrig. Drain. 2019, 68, 176–190. https://doi.org/10.1002/ird.2293.
15. Chauhan, P.; Chauniyal, D.D.; Singh, N.; Tiwari, R.K. Quantitative geo-morphometric and land cover-based micro-watershed prioritization in the Tons river basin of the lesser Himalaya. Environ. Earth Sci. 2016, 75, 498. https://doi.org/10.1007/s12665-016-5342-x.
16. Singh, G.; Pandey, A. Morphometric Characterization and Flash Flood Zonation of a Mountainous Catchment Using Weighted Sum Approach. Geospatial Technologies for Land and Water Resources Management; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 409–428. https://doi.org/10.1007/978-3-030-90479-1_23.
17. El-Fakharany, M.A.; Hegazy, M.N.; Mansour, N.M.; Abdo, A.M. Flash flood hazard assessment and prioritization of sub-watersheds in Heliopolis basin, East Cairo, Egypt. Arab. J. Geosci. 2021, 14, 1693. https://doi.org/10.1007/s12517-021-07991-7.
18. Jothimani, M.; Abebe, V.; Dawit, Z. Mapping of soil erosion-prone sub-watersheds through drainage morphometric analysis and weighted sum approach: A case study of the Kulfo River basin, Rift valley, Arba Minch, Southern Ethiopia. Model. Earth Syst. Environ. 2020, 6, 2377–2389. https://doi.org/10.1007/s40808-020-00820-y.
19. Mahmood, S.; Rahman, A. Flash flood susceptibility modeling using geo-morphometric and hydrological approaches in Panjkora Basin, Eastern Hindu Kush, Pakistan. Environ. Earth Sci. 2019, 78, 43. https://doi.org/10.1007/s12665-018-8041-y.
20. Mahmood, S.; Rahman, A. Flash flood susceptibility modelling using geomorphometric approach in the Ushairy Basin, eastern Hindu Kush. J. Earth Syst. Sci. 2019, 128, 97. https://doi.org/10.1007/s12040-019-1111-z.
21. Khan, I.; Bali, R.; Agarwal, K.K.; Kumar, D.; Singh, S.K. Morphometric Analysis of Parvati Basin, NW Himalaya: A Remote Sensing and GIS Based Approach. J. Geol. Soc. India 2021, 97, 165–172. https://doi.org/10.1007/s12594-021-1648-8.
22. Malik, A.; Kumar, A.; Kandpal, H. Morphometric analysis and prioritization of sub-watersheds in a hilly watershed using weighted sum approach. Arab. J. Geosci. 2019, 12, 118. https://doi.org/10.1007/s12517-019-4310-7.

23. Sakthivel, R.; Jawahar Raj, N.; Sivasankar, V.; Akhila, P.; Omine, K. Geo-spatial technique-based approach on drainage morphometric analysis at Kalrayan Hills, Tamil Nadu, India. Appl. Water Sci. 2019, 9, 24. https://doi.org/10.1007/s13201-019-0899-7.
24. Prasad, R.N.; Pani, P. Geo-hydrological analysis and sub watershed prioritization for flash flood risk using weighted sum model and snyder's synthetic unit hydrograph. Model. Earth Syst. Environ. 2017, 3, 1491–1502. https://doi.org/10.1007/s40808-017-0354-4.
25. Prakash, K.; Mohanty, T.; Singh, S.; Chaubey, K.; Prakash, P. Drainage morphometry of the Dhasan river basin, Bundelkhand craton, central India using remote sensing and GIS techniques. J. Geomat. 2016, 10, 122–132.
26. Kumar, A.; Singh, S.; Pramanik, M.; Chaudhary, S.; Maurya, A.K.; Kumar, M. Watershed prioritization for soil erosion mapping in the Lesser Himalayan Indian basin using PCA and WSA methods in conjunction with morphometric parameters and GIS-based approach. Environ. Dev. Sustain. 2022, 24, 3723–3761. https://doi.org/10.1007/s10668-021-01586-8.
27. Rahman, M.M.; Zaman, M.N.; Biswas, P.K. Optimization of significant morphometric parameters and sub-watershed prioritization using PCA and PCA-WSM for soil conservation: A case study in dharla River watershed, Bangladesh. Model. Earth Syst. Environ. 2022, 8, 2661–2674. https://doi.org/10.1007/s40808-021-01255-9.
28. Malik, A.; Kumar, A.; Kushwaha, D.P.; Kisi, O.; Salih, S.Q.; Al-Ansari, N.; Yaseen, Z.M. The Implementation of a Hybrid Model for Hilly Sub-Watershed Prioritization Using Morphometric Variables: Case Study in India. Water 2019, 11, 1505–1519.
29. Meshram, S.G.; Sharma, S.K. Prioritization of watershed through morphometric parameters: a PCA-based approach. Appl. Water Sci. 2017, 7, 1505–1519.
30. Meshram, S.G.; Alvandi, E.; Meshram, C.; Kahya, E.; Fadhil Al-Quraishi, A.M. Application of SAW and TOPSIS in Prioritizing Watersheds. Water Resour. Manag. 2020, 34, 715–732. https://doi.org/10.1007/s11269-019-02470-x
31. CARE-Perú. Modelización Hidrológica de la Cuenca Cañete y Evaluación del Impacto del Cambio Climático; CARE-Perú: Lima, Peru, 2018.
32. Portuguez-Maurtua, M. Aplicación de la geoestadística a modelos hidrológicos en la cuenca del río Cañete. Master's Thesis, Universidad Nacional Agraria La Molina, Lima, Peru, 2017.
33. Arulbalaji, P.; Padmalal, D. Sub-watershed Prioritization Based on Drainage Morphometric Analysis: A Case Study of Cauvery River Basin in South India. J. Geol. Soc. India 2020, 95, 25–35. https://doi.org/10.1007/s12594-020-1383-6.
34. Meraj, G.; Romshoo, S.A.; Yousuf, A.R.; Altaf, S.; Altaf, F. Assessing the influence of watershed characteristics on the flood vulnerability of Jhelum basin in Kashmir Himalaya. Nat. Hazards 2015, 77, 153–175.
35. Gajbhiye, S.; Mishra, S.K.; Pandey, A. Prioritizing erosion-prone area through morphometric analysis: An RS and GIS perspective. Appl. Water Sci. 2014, 4, 51–61. https://doi.org/10.1007/s13201-013-0129-7.
36. Odiji, C.A.; Aderoju, O.M.; Eta, J.B.; Shehu, I.; Mai-Bukar, A.; Onuoha, H. Morphometric analysis and prioritization of upper benue river watershed. Appl. Water Sci. 2021, 11, 41. https://doi.org/10.1007/s13201-021-01364-x.
37. Strahler, A. Quantitative Geomorphology of Drainage Basin and Channel Networks. In Handbook of Applied Hydrology; McGraw-Hill: New York, NY, USA, 1964.
38. Horton, R.E. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. Geol. Soc. Am. Bull. 1945, 56, 275–370.
39. Nooka Ratnam, K.; Srivastava, Y.; Venkateswara Rao, V.; Amminedu, E.; Murthy, K. Check dam positioning by prioritization of micro-watersheds using SYI model and morphometric analysis remote sensing and GIS perspective. J. Indian Soc. Remote Sens. 2005, 33, 25–38. Available online: https://sci-hub.wf/10.1007/bf02989988 (accessed on 25 November 2022).
40. Sreedevi, P.D.; Sreekanth, P.D.; Khan, H.H.; Ahmed, S. Drainage morphometry and its influence on hydrology in an semi arid region: Using SRTM data and GIS. Environ. Earth Sci. 2013, 70, 839–848. https://doi.org/10.1007/s12665-012-2172-3.
41. Schumm, S.A. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. Geol. Soc. Am. Bull. 1956, 67, 597–646. https://doi.org/10.1130/0016-7606.
42. Horton, R.E. Drainage-basin characteristics. Trans. Am. Geophys. Union 1932, 13, 350–361.

43. Miller, V. A Quantitative Geomorphic Study of Drainage Basin Characteristics in the Clinch Mountain Area Virginia and Tennessee. The Journal of Geology. 1953, 65, 112-113.New York-USA. https://doi.org/10.1086/626413.
44. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. Rev. Geophys. 2007, 45, 13, 5326–5350. 10.1109/JSTARS.2020.3021052.
45. Amani, M.; Ghorbanian, A.; Ahmadi, S.A.; Kakooei, M.; Moghimi, A.; Mirmazloumi, S.M.; Moghaddam, S.H.A.; Mahdavi, S.; Ghahremanloo, M.; Parsian, S.; et al. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2020, 13, 5326–5350. https://doi.org/10.1109/JSTARS.2020.3021052.
46. Mutanga, O.; Kumar, L. Google Earth Engine Applications. Remote Sens. 2019, 11, 18–27.
47. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 2017, 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031
48. Zhao, Q.; Yu, L.; Li, X.; Peng, D.; Zhang, Y.; Gong, P. Progress and Trends in the Application of Google Earth and Google Earth Engine. Remote Sens. 2021, 13, 3778.
49. Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 2020, 164, 152–170. https://doi.org/10.1016/j.isprsjprs.2020.04.001.
50. Konrad Banachewicz, L.M. The Kaggle Book; Packt: Birmingham, UK, 2022. ISBN 9781801817479.
51. Ismail, M.; Singh, H.; Farooq, I.; Yousuf, N. Quantitative morphometric analysis of Veshav and Rembi Ara watersheds, India, using quantum GIS. Appl. Geomat. 2022, 14, 119–134. https://doi.org/10.1007/s12518-022-00417-3.
52. Basnet, K.; Paudel, R.C.; Sherchan, B. Analysis of watersheds in Gandaki province, Nepal using QGIS. Tech. J. 2019, 1, 16–28.
53. Alencar da Silva Alves, K.M.; Parodi Dávila, M.C.; Zimmermann García, E.D.; Rodrigues de Lira, D.; De Araujo Monteiro, K. Caracterización morfométrica de la cuenca del Salado Bajo, Región de Atacama, Chile. Investig. Geográficas 2021, 62, 90–105. https://doi.org/10.5354/0719-5370.2021.64574.
54. Gayen, S.; Bhunia, G.S.; Shit, P.K. Morphometric Analysis of Kangshabati-Darkeswar Interfluves Area in West Bengal, India using ASTER DEM and GIS Techniques, 2013. Available online: http://111.93.204.14:8080/xmlui/handle/123456789/582 (accessed on 30 November 2022).

CAPITULO IV: Modelación hidrológica en base a precipitaciones extremas grilladas

1. Introducción

La precipitación es un componente fundamental del ciclo hidrológico global que gobierna la distribución de los recursos hídricos (Li et al., 2018). El entendimiento de su comportamiento temporal y espacial es de gran interés, especialmente en los estudios de riesgos climáticos, donde la disponibilidad de información de alta resolución y de buena calidad es esencial (Chen et al., 2019). Además, los desastres naturales se están presentando con mayor frecuencia debido a la ocurrencia de fenómenos climáticos como consecuencia del cambio climático, alterando los cambios en las variables climáticas primarias, como la temperatura y la precipitación, que son las principales impulsoras de los cambios climáticos (Pérez et al., 2021, Kropp, 2015).

Las precipitaciones picos, como las series de datos máximos anuales de cierta duración, se utilizan generalmente para estimar las precipitaciones extremas mediante el análisis de frecuencia (Das, 2021, Ashkar and Ba, 2017). El análisis de la frecuencia permite relacionar la magnitud de los eventos extremos con su ocurrencia mediante el uso de distribuciones de probabilidad (Alam et al., 2018). La frecuencia de los fenómenos extremos, como las precipitaciones y las inundaciones, puede expresarse en términos de período de retorno (Tr), permitiendo estimar el valor de la magnitud del fenómeno correspondiente a un Tr determinado (Salhi, 2022, Alam et al., 2018, Hromadka et al., 2010).

Para abordar la estimación de precipitaciones en lugares donde no se dispone de series de datos confiables, se tiene como alternativa la interpolación espacial en la generación de la distribución espacial de las precipitaciones extremas (Ali et al., 2021, Das, 2021). Existen diferentes métodos de interpolación espacial de precipitaciones a partir de estaciones situadas dentro o en las cercanías a zona de estudio (Salhi, 2022, Zou et al., 2021, Amini et al., 2019, Adhikary et al., 2017). La elección del método de interpolación es crucial para la generación de isoyetas a diferente Tr, debido que las precipitaciones extremas suelen tener una gran variabilidad espacial, especialmente en duraciones cortas (Zou et al., 2021). Los métodos de interpolación espacial son clasificadas en métodos determinísticos y geoestadísticos, comúnmente utilizados en muchos estudios (Ali et al., 2021, Bárdossy et al., 2021, Amini et al., 2019, Foehn et al., 2018). Los métodos geoestadísticos en la actualidad, se presentan como alternativa eficiente en la

representación espacial de precipitaciones extremas (Ali et al., 2021, Amini et al., 2019, Adhikary et al., 2017).

Además, la representación espacial y temporal de la precipitación es vital para la generación de escorrentía mediante modelación hidrológica (Belayneh et al., 2020). No obstante, los productos de la precipitación grillada sirvieron de datos entrada en el modelo precipitación - escorrentía en zona con escasez de datos de campo. En este estudio evaluamos la respuesta hidrológica para Tr de 5, 10, 20, 50 y 100 años utilizando precipitaciones grilladas en la Cuenca del río Cañete, mediante el programa HEC-HMS. El rendimiento del modelo hidrológico fue evaluado mediante métricas estadísticas, entre datos precipitación terrestre y precipitación grillada. Los resultados mostraron que el modelo HEC-HMS predijo bien la escorrentía de la cuenca con producto de precipitación grillada. En general, nuestro estudio demostró que la simulación basada en precipitación grillada superó a la basada en estaciones terrestre en el área de estudio. Finalmente, esta metodología puede utilizarse en futuras investigaciones para evaluar el rendimiento de los datos de precipitación grilladas obtenidas de otras fuentes (sensores satelitales) y evaluar su rendimiento en un modelo hidrológico.

2. Materiales y métodos

2.1 Método de interpolación espacial de precipitaciones

La metodología adoptada en este ítem consta de cuatro etapas; la figura 2 muestra el diagrama metodológico. La primera etapa consiste en colección de información disponible y evaluación de calidad de la precipitación máxima en 24 horas. La segunda etapa corresponde a la selección de la distribución de probabilidad que mejor se ajuste a una determinada estación, la elección se da mediante la prueba de bondad de ajuste. La tercera etapa corresponde al proceso de interpolación espacial aplicando kriging ordinario. Por último, la cuarta etapa consiste en la evaluación de los resultados mediante métricas estadísticas.

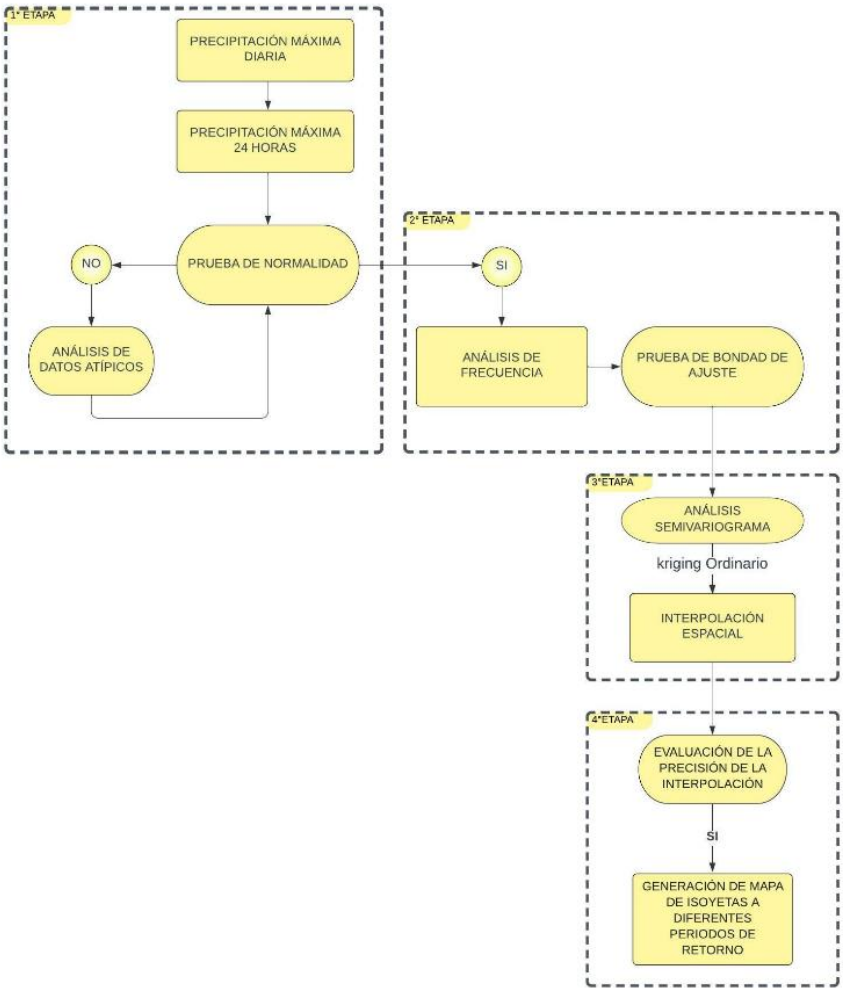


Figura 2: Esquema metodológico de interpolación espacial de precipitaciones extremas

2.1.1 Colección y evaluación de calidad de los datos

a Colección de datos disponibles

Se seleccionaron un total de 17 estaciones pluviométricas, algunas con registros desde 1965, otras desde 1970, etc., la mayoría con periodos de registros faltantes. Las estaciones forman parte de la red meteorológica gestionada por el Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI, https://www.senamhi.gob.pe/mapas/mapa-estaciones/mapadepesta1.php consultado el 20 de octubre de 2022). Desde las estaciones pluviométrica ubicadas dentro y alrededores de la región de estudio, se seleccionaron las precipitaciones máximas diarias anuales. La Figura 1 muestra la ubicación espacial de las estaciones pluviométricas y la Tabla 1 la localización en coordenadas geográficas, periodo de registro y cantidad de datos observados.

Tabla 1: Estaciones pluviométricas de las cuencas MOC, período 1965-2019

ID	Stations	Coordinates		Altitude	Period	Observed Data
		Latitude	Longitude	(masl)	From-To	No of Data
1	Ayaviri	−12.38	−76.13	3228	1965-2019	51
2	Cañete	−13.07	−76.32	158	1970-2019	42
3	Carania	−12.34	−75.87	3875	1965-2019	55
4	Huancata	−12.22	−76.22	2700	1980-2019	40
5	Huangascar	−12.9	−75.83	2533	1965-2019	55
6	Huañec	−12.29	−76.14	3205	1965-2019	55
7	Huarochiri	−12.13	−76.23	3154	1965-2019	55
8	Langa	−12.13	−76.42	2863	1980-2019	40
9	Pacaran	−12.83	−76.07	700	1964-2019	48
10	San Juan de Yanac	−13.21	−75.79	2550	1966-2019	53
11	San Lazaro de Escomarca	−12.18	−76.35	3758	1965-2019	55
12	San Pedro de Pilas	−12.45	−76.22	2600	1997-2019	37
13	Socsi	−13.03	−76.19	500	2004-2019	14
14	Tanta	−12.12	−76.02	4323	1965-2019	55
15	Vilca	−12.11	−75.83	3864	1965-2019	55
16	Yauricocha	−12.32	−75.72	4675	1986-2019	34
17	Yauyos	−12.49	−75.91	2327	1965-2019	48

b Estimación de Precipitación máxima en 24 horas

Las series máximas diarias anuales son valores de precipitación en 24 horas, las cuales son obtenidas de las estaciones en intervalos fijos de tiempo, generalmente de 7 am. a 7 am. Estos valores normalmente no coinciden con los valores máximos reales en 24 horas. (Hershfield, 1961) propuso un factor multiplicativo de 1.13 a la precipitación máxima diaria anual con intervalos fijos, para aproximarla a los valores reales de precipitación máxima en 24 horas. Sin embargo, la Organización Meteorológica Mundial (OMM) (WMO, 2009), para el ajuste recomienda para corregir el uso de intervalos de tiempo fijos de 24 horas, siempre que se disponga de series de datos largas (más de 15 años).

La base de la metodología para estimar la precipitación máxima en 24 horas se toma en consideración a partir de la ecuación de frecuencia de (Hershfield, 1965). En general, es una función de la media aritmética de la precipitación máxima diaria, $\overline{X}$, y la desviación estándar, S_X, de toda la serie temporal de precipitación máxima diaria a través del factor, K_m, de la desviación estándar como, ecuación (1) (Sen, 2018)

$$P_{max} = \overline{X} + K_m S_X \qquad (1)$$

Donde, K_m se denomina factor de frecuencia, que es una parte muy importante en la ecuación presentada ya que constituye el número de desviaciones estándar que se añaden al valor medio de la distribución para alcanzar el mayor valor posible de precipitación dentro de una serie.
Si el número anual de precipitaciones diarias máximas es n, la exclusión del máximo entre las precipitaciones diarias máximas conduce a otra serie de precipitaciones diarias máximas de longitud $n-1$. En tal situación, la cantidad de precipitaciones diarias máximas puede relacionarse con la media aritmética y la desviación estándar de esta nueva serie, de forma similar a la ecuación (1), del siguiente modo, ecuación (2) (Sen, 2018):

$$P_{max\ 24h} = \overline{X}_{\ n-1} + K_m S_{X(n-1)} \qquad (2)$$

Por lo tanto:

$$K_m = \frac{P_{max} - \overline{X}_{\ n-1}}{S_{X(n-1)}} \qquad (3)$$

c Prueba de normalidad precipitación máxima en 24 horas

Existen principalmente tres formas comunes de comprobar el supuesto de normalidad. La forma más sencilla es utilizar métodos gráficos, denominado el gráfico de cuantiles normales (gráfico Q-Q). Seguido de otros métodos gráficos comunes que pueden utilizarse para evaluar el supuesto normalidad son el histograma y diagrama de caja. Estos métodos gráficos pueden ser herramientas útiles para comprobar la normalidad de una muestra de n observaciones independientes, pero, no son suficientes para proporcionar pruebas concluyentes (Flores Tapia and Flores Cevallos, 2021, Razali et al., 2011). Por lo tanto, para respaldar los métodos gráficos, deben realizarse métodos más formales, como los métodos analíticos, los más comunes disponibles son la prueba de Anderson-Darling (AD), prueba de Shapiro-Wilk (SW) y prueba de Kolmogorov-Smirnov (KS).

Prueba de Anderson-Darling

Prueba de normalidad propuesto por (Anderson and Darling, 1954) para cualquier distribución completamente especificada, es decir, con parámetros conocidos, la ecuación (4) presenta el análisis (Bayoud, 2021, Flores Tapia and Flores Cevallos, 2021, Razali et al., 2011).

$$AD^2 = -n - \frac{\sum_{i=1}^{n} (\ 2i-1\ )\ (\ lnz_i + ln\ (\ 1 - z_{n+1-i}\ )\ )}{n} \qquad (4)$$

Donde z es la función de distribución acumulada de la distribución especificada, *i* son los datos ordenados y *n* es el tamaño de la muestra.

Para probar la normalidad con parámetros desconocidos, la versión modificada de AD es, (D'Agostino and Stephens, 1986)

$$AD^* = AD^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \qquad (5)$$

La hipótesis nula H_o se rechaza al nivel de significación 0.05 si $AD^* > 0.752$ para cualquier n

Prueba de Shapiro-Wilk

Prueba de normalidad propuesto por (Shapiro and Wilk, 1965) y presentado en la ecuación (6) por (Bayoud, 2021):

$$SW = \frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{(n-i+1)}\left(x_{(n-i+1)} - x_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2} \qquad (6)$$

El valor de SW deben estar entre cero y uno. Valores pequeños de SW conducen al rechazo de la normalidad, mientras que un valor de uno indica la normalidad de los datos (Razali et al., 2011). Es recomendable no aplicar cuando se dispone de muchos datos (más de 50) por su elevada sensibilidad a pequeñas desviaciones de la normal (Bayoud, 2021, Razali et al., 2011). La hipótesis nula H_o se rechaza cuando el SW es pequeño (Bayoud, 2021).

Prueba de Kolmogorov-Smirnov

La prueba de Kolmogórov-Smirnov es una prueba de bondad de ajuste ampliamente utilizada para probar la normalidad de los datos muestrales (Flores Tapia and Flores Cevallos, 2021). La ecuación (7) la representa y se muestra a continuación (Razali et al., 2011).

$$\mathrm{D} = \sup_x |F^*(X) - F_n(x)| \qquad (7)$$

Donde $'\sup'$ significa el mayor. $F^*(X)$ es la función de distribución hipotética, mientras que $F_n(x)$ es la función de distribución empírica (FDE) estimada a partir de la muestra aleatoria. Esta prueba considera que $F^*(X)$ es una distribución normal con media conocida, μ, y desviación estándar, σ.

Prueba modificada de Kolmogórov-Smirnov (Bayoud, 2021):

$$D^* = D\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) \qquad (8)$$

La hipótesis nula H_o se rechaza al nivel de significación 0.05 si D^* >0.895 para cualquier n (Bayoud, 2021, Stephens, 1974)

d Análisis de valores atípicos

Los valores atípicos son aquellos puntos de datos que se alejan de la mayoría de los puntos similares; estos causan problemas a la hora de construir modelos predictivos. Existe dos tipos de valores atípicos: univariantes y multivariantes. Los valores atípicos univariantes pueden encontrarse en distribuciones de una sola variable, mientras que los multivariantes pueden encontrarse en espacios n-dimensionales (Nathan, 2021, Navlani et al., 2021). Se aplicaron los siguientes métodos:

Método de Tukey:

Basado en rango interquartil (IQR).

$$IQR = 75\%\ Percentile - 25\%\ Percentile \quad (9)$$

Donde:

$$upper\ limit = 75\%\ Percentile + 1.5 * IQR \quad (10)$$

$$lower\ limit = 75\%\ Percentile - 1.5 * IQR \quad (11)$$

Método Z-score

$$z = \frac{x - u}{\sigma} \quad (12)$$

Donde x es un punto de datos (una observación), mu (u) es la media del conjunto de datos, y sigma (σ) es la desviación estándar del conjunto de datos (Atwan, 2022).

2.1.2 Selección de distribución de probabilidades y prueba de bondad de ajuste

Para seleccionar la distribución de probabilidad que mejor se ajuste a una muestra dada, es importante la elección de modelos de distribución de probabilidad. Se presentan los modelos de distribución que son utilizados habitualmente en análisis de precipitaciones extremas (Alam et al., 2018).

a Normal

La distribución normal o gaussiana es utilizada para explicar la conducta de una variable aleatoria continua que oscila de forma simétrica en torno a un valor esencial. Su función de distribución de probabilidad (PDF) es la siguiente (Alam et al., 2018, Naghettini, 2017).

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad (13)$$

Donde, μ es parámetros de escala y σ es la desviación estándar.

b Log-Normal

Su PDF está dado por (Alam et al., 2018, Naghettini, 2017):

$$f_x(x) = \frac{1}{x\sigma_{ln(x)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{ln(X) - \mu_{ln(X)}}{\sigma_{ln(X)}}\right]^2\right\} \quad (14)$$

Donde, los parámetros son μ_y y σ_y, $Y = \ln(X)$ es el argumento principal.

c Pearson III

Su función de densidad es la siguiente (Alam et al., 2018).

$$f_x(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x-\xi}{\alpha}\right)^{\beta-1} \exp\left(-\frac{x-\xi}{\alpha}\right) \quad (15)$$

Donde α: parámetros de escala, β: parámetros de forma y ξ localización.

d Log-Pearson III

Describe una variable aleatoria cuyo logaritmo sigue la distribución pearson 3 y su PDF es la siguiente (Alam et al., 2018):

$$f_x(x) = \frac{1}{\alpha x\Gamma(\beta)}\left[\frac{\ln(x)-\xi}{\alpha}\right]^{\beta-1} \exp\left[-\frac{\ln(x)-\xi}{\alpha}\right] \quad (16)$$

e EV1-max

La distribución de valores extremos de tipo 1, también llamada gumbel, se utiliza a menudo para representar un proceso máximo (Alam et al., 2018, Naghettini, 2017). El PDF de esta distribución está dado por:

$$F_Y(y) = \exp\left[-\exp\left(-\frac{y-\beta}{\alpha}\right)\right] \quad (17)$$

Donde α es parámetro de escala y β parámetro de localización.

f GEV-max

La distribución de valor extremo generalizado para máximos (GEV-max) es de forma paramétrica y su PDF es el siguiente (Das, 2021, Naghettini, 2017).

$$F_Y(y) = \exp\left\{-\left[1-\kappa\left(\frac{y-\beta}{\alpha}\right)\right]^{\frac{1}{\kappa}}\right\} \quad (18)$$

Donde, κ es parámetro de forma, α parámetro de escala y β parámetros de ubicación.

2.1.3 Interpolación espacial con kriging ordinario

La estructura espacial de las estaciones pluviométricas próximas entre sí, son más similares que las alejadas (autocorrelación espacial). El Semi-variograma es una herramienta descriptiva para especificar el patrón espacial de una característica en particular (Amini et al., 2019).

Semi-variograma

El semivariograma permite analizar la autocorrelación espacial de una variable, es crucial al utilizar los métodos geoestadísticos en obtener las estimaciones de puntos desconocidos, es estimado empíricamente a partir de los conjuntos de datos se siguiente manera (Salhi, 2022, Das, 2021, Zou et al., 2021).

$$\gamma(h) = \frac{1}{2 \cdot N(h)} \sum_{i=1}^{N(h)} \left(Z(x_i) - Z(x_i + h)\right)^2 \qquad (19)$$

Donde $\gamma(h)$ es el valor del semivariograma, h es la distancia de separación entre dos localizaciones, $N(h)$ representa el número de variables pares, que están situadas a un vector de distancia h y $\left(Z(x_i) - Z(x_i + h)\right)$ es la diferencia de valores entre localizaciones x_i y $(x_i + h)$.

Kriging Ordinario (KO)

Es la técnica de interpolación más común y frecuentemente aplicada en geoestadística (Das, 2021, Zou et al., 2021, Huang et al., 2018)

$$Z(x) = \Sigma_{i=1}^{n} \lambda_i Z(x_i) \qquad (20)$$

Donde $Z(x_i)$ es la precipitación máxima en 24 horas de cualquier nivel de retorno en el i^{th} sitio de la zona de estudio y λ_i es el coeficiente de ponderación, que representa la contribución de cada valor de muestra de Tr conocido $Z(x_i)$ al valor de Tr estimado por kriging $Z(x)$. La ponderación puede ser determinado mediante el modelo de semivariograma.

2.1.4 Métricas estadísticas

Para evaluar la eficacia de la interpolación espacial KO, se utilizaron las métricas estadísticas: coeficiente de Determinación (R^2), Raíz cuadrada del error cuadrático medio (RMSE), coeficiente de Nash-Sutcliffe (NSE) y el porcentaje de Sesgo (PBIAS). Las ecuaciones y las correspondientes clasificaciones se presentan en la Tabla 2.

Tabla 2: Criterio para evaluar el rendimiento de la interpolación espacial y sus clasificaciones.

Métricas estadísticas	Valores	Performance classification	Referencia
$R^2 = \frac{\left[\sum_{t=1}^{n}(P_{obs} - \bar{P}_{obs})(P_{ped} - \bar{P}_{pred})\right]^2}{\sum_{t=1}^{n}(P_{obs} - \bar{P}_{obs})^2 \sum_{t=1}^{n}(P_{pred} - \bar{P}_{pred})^2}$	r≥ 0.5 0 ≤ r < 0.5	Satisfactorio Malo	(Santhi et al., 2001)
$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_{obs,i} - P_{pred,i})^2}{n}}$	Valor inferior a la mitad de la desviación estándar	Satisfactorio	(Moriasi et al., 2007)
$NSE = 1 - \frac{\sum_{i=1}^{n}(P_{pred,i} - P_{obs,i})^2}{\sum_{i=1}^{n}(P_{pred,i} - \bar{P}_{pred,i})^2}$	0.75 < NSE ≤ 1.00 0.65 < NSE ≤ 0.75 0.50 < NSE ≤ 0.65 0.40 < NSE ≤ 0.50 NSE ≤ 0.40	Excelente Bueno Satisfactorio Aceptable Malo	(Moriasi et al., 2007)
$PB = \frac{\sum_{i=1}^{n}(P_{obs,i} - P_{pred,i}) \times 100}{\sum_{i=1}^{n} P_{pred,i}}$	PBIAS < ±10 ±10 ≤ PBIAS <±15 ±15 ≤ PBIAS <±25 PBIAS ≥ ±25	Excelente Bueno Satisfactorio Malo	(Moriasi et al., 2007)

Donde *n* representa el número de días de predicción, P_{obs} corresponde al valor de la precipitación para un Tr específico, P_{ped} es el valor de precipitación predicho para un Tr específico, *i* representa la medición en una estación específica, $\bar{P}_{obs}$ y $\bar{P}_{pred}$ corresponde a los valores promedios medidos y predichos, respectivamente.

2.2 Método de Modelación hidrológica

La metodología adoptada en este ítem consta de cinco etapas; la figura 3 muestra el diagrama metodológico. La primera etapa consiste en colección de información disponible. La segunda etapa corresponde a preprocesamiento de información espacial (Modelo Digital de Elevación, precipitación grillada y curva número grillada). La tercera etapa corresponde al desarrollo del modelo hidrológico en HEC-HMS. La cuarta etapa corresponde a la evaluación del rendimiento del modelo. Por último, la estimación de la escorrentía directa.

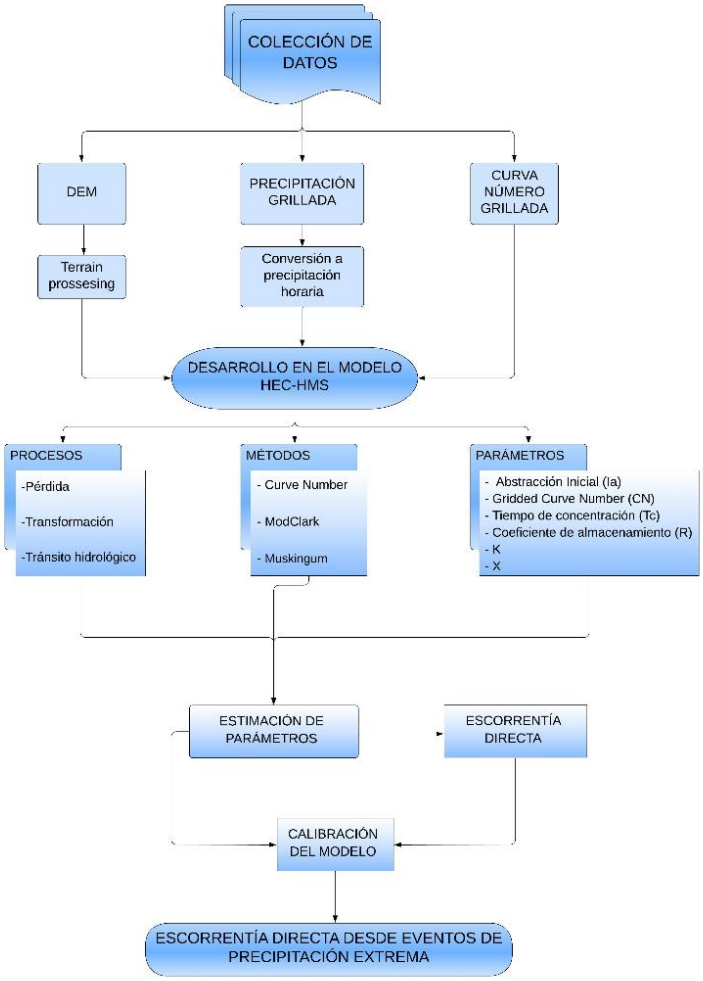


Figura 3: Esquema metodológico de modelación hidrológica con HEC-HMS

2.2.1 Colección de datos

Los conjuntos de datos, como Modelo digital de elevación (MDE), curva número grillada (CN), precipitaciones grilladas y serie de datos de descargas, son datos importantes y necesarios para la modelación hidrología. Estos conjuntos de datos se utilizaron para estimar los parámetros hidrológicos y las características de las subcuencas. Los tipos de datos utilizados y sus fuentes se detallan en la Tabla 3.

Tabla 3: Datos, tipos de estructuras y fuentes utilizadas

Datos	**Estructura**	**Fuente**
MDE	Raster	SRTM (Farr et al., 2007)
Curva Numero grillada	Rater	(Portuguez-Maurtua and Verano Zelada, 2016)
Precipitación grillada	Raster	Generada mediante interpolación Geoestadística
Datos de descarga	Serie de tiempo	SENAMHI

2.2.2 Preprocesamiento de información

a Modelo Digital de Elevación

Para el estudio se utilizaron MDE de Shuttle Radar Topography Mission (SRTM). El MDE, con una resolución espacial de 30 m (Farr et al., 2007), fue descargado desde Google Earth Engine (GEE), utilizando código en lenguaje Python en QGIS (PyQGIS). Para determinar y extraer parámetros hidrológicos como pendiente, dirección del flujo, acumulación de flujo, características del drenaje y la delineación de las cuencas hidrográficas. Además, se calculó la pendiente en porcentaje con un tamaño de píxel de 30, todos estos procesos se realizaron con el programa HEC-HMS. La Figura 4 muestra el MDE, las subcuencas de la cuenca Cañete y red de drenaje.

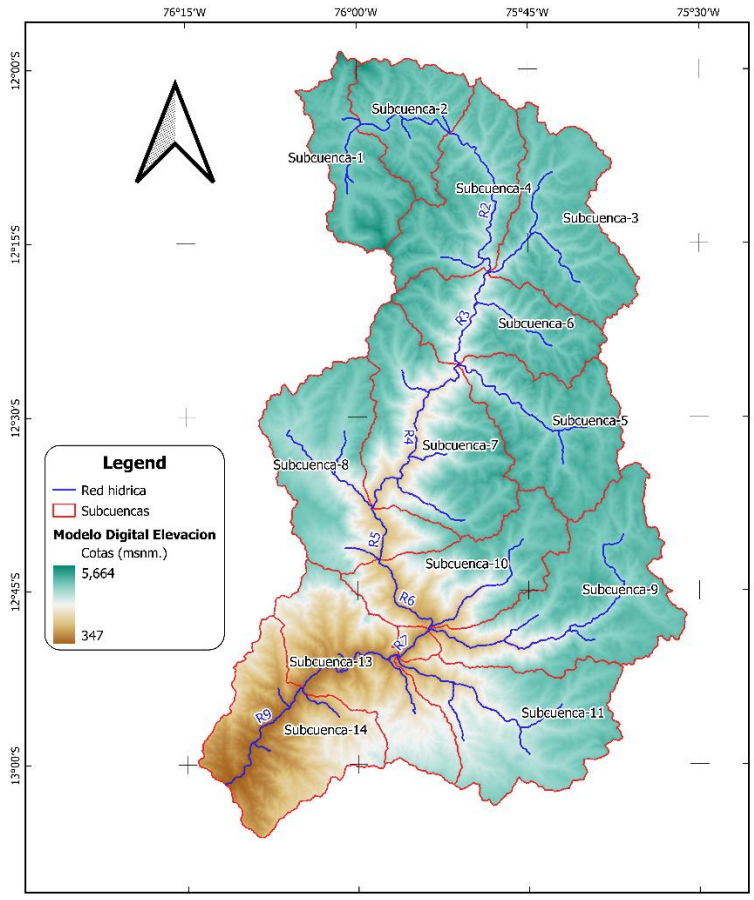


Figura 4: Subcuencas, red de drenaje y Modelo Digital de Elevación en Cuenca del río Cañete

b Curva Número Grillada

El método del Servicio de Conservación de Recursos Naturales de EE.UU. (siglas en ingles NRCS), es un método para calcular la cantidad de pérdida de agua causada por la infiltración de

la tierra. La cantidad de pérdida de agua en base al método NRCS depende en gran medida del parámetro del Curva Número (CN). CN es el número de curva SCS, que representa la combinación de un grupo hidrológico del suelo y las clases de uso del suelo (Natarajan and Radhakrishnan, 2021, Portuguez-Maurtua and Verano Zelada, 2016). Para la estimación del proceso de infiltración en este estudio, se ha utilizado el mapa de Curva Numero grillada, (Figura 5) disponible en la plataforma Google Earth Engine (https://code.earthengine.google.com/) publicado por (Portuguez-Maurtua and Verano Zelada, 2016).

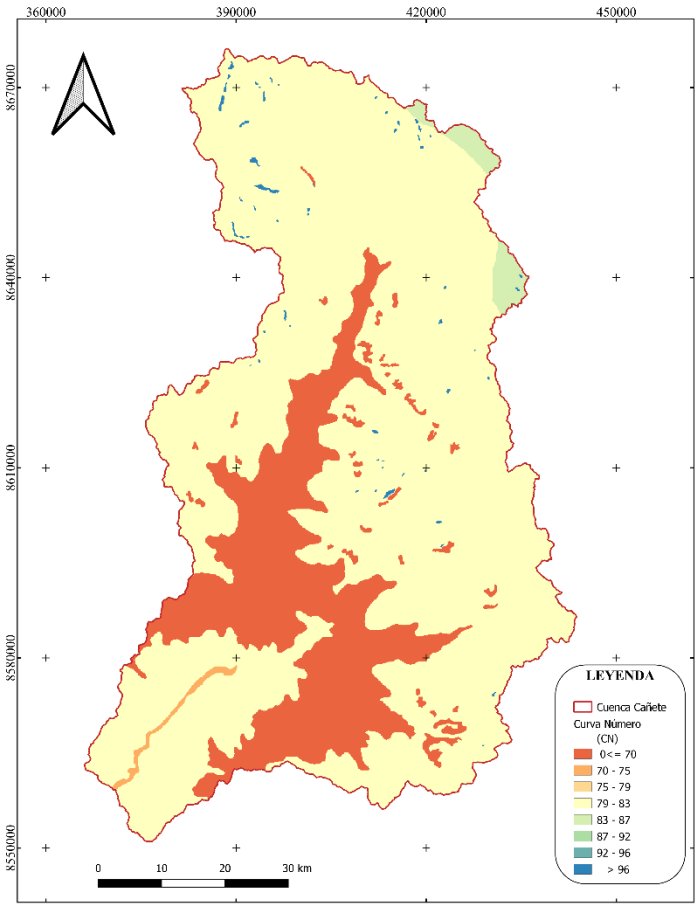


Figura 5: Mapa temático de Curva Numero – Cuenca rio Cañete

c Precipitación Grillada

Los datos de precipitación grillada fueron obtenidos a partir de interpolación espacial de las precipitaciones extremas para periodo de retorno de 5, 10, 20, 50 y 100 años, mediante el método geoestadístico kriging ordinario (KO), con resolución espacial de 1000 m. x 1000 m. (Figura 9). La conversión a precipitación horaria se realizó en base al método de hietograma sintético de tormentas con duraciones de 24 horas del tipo AI, desarrollado por NRCS (Chow et al., 1994).

d Datos de descarga

Los datos de descarga máxima fueron registrados en la estación hidrométrica Socsi, registros tomado para los periodos 1965-2019. Estación perteneciente a la red meteorológica del Servicio Nacional de Meteorología e Hidrología (SENAMHI) y Autoridad Nacional del Agua (ANA). La Tabla 4 muestra la localización en coordenadas geográficas, periodo de registro y cantidad de datos observados, información necesaria para la calibración de los parámetros en el modelo HEC-HMS.

Tabla 4: Estación hidrométrica ubicada cuenca Cañete, período 1965-2019

ID	Estación	Coordenadas		Altitud	Periodo	Datos Observados
		Latitud	Longitud	(msnm)	Desde-Hasta	No de Datos
1	Socsi	−13.03	−76.20	430	1965-2019	55

2.2.3 Modelo HEC-HMS

Para la modelación hidrológica con datos de precipitación grillada se utilizó el modelo HEC-HMS, que es un programa capaz de simular la precipitación-escorrentía distribuida (Cho, 2020, Paudel et al., 2009). Para simular las pérdidas por infiltración o perdida se utilizó Gridded SCS Curve Number, para transformar el exceso de precipitación en escorrentía superficial, se empleó el hidrograma unitario Clark modificado (ModClark) y para transito hidrológico en cauce se utilizó Muskingum. Además, en este modelo no se incluyó el flujo base, esto debido a la magnitud del evento. La lista de parámetros, métodos y procesos utilizados para configurar el modelo hidrológico en HEC-HMS se muestran en la Tabla 5.

Tabla 5: Proceso, métodos y parámetros usadas en HEC-HMS

Procesos	Métodos	Parámetros
Perdidas	Curva Numero	Abstracción Inicial (Ia) Gridded SCS Curve Number (CN)
Transformación	ModClark	Tiempo de Concentración (Tc) Coeficiente de almacenamiento (R)
Transito hidrológico	Muskingum	K X

Adaptado (Bastia et al., 2021, Eryani et al., 2021)

a Método de perdida

El método de pérdidas simula la pérdida de agua en la cuenca desde la lluvia hasta la escorrentía superficial (Cheah et al., 2019). El modelo simulará el proceso de pérdida de agua de lluvia debido al uso del suelo y al tipo de suelo a través de los procesos de infiltración y evapotranspiración antes de convertirse finalmente en escorrentía directa (Eryani et al., 2021). El método de pérdida adoptado en la presente investigación es Curva Número Grillada, en base a este raster se determinaron los parámetros iniciales de Abstracción, CN e impermeabilidad. El valor de la abstracción inicial (Ia) es estimada en función de CN, como se ilustra en la ecuación (21) (Jabbar et al., 2021).

$$I_a = 0.2 * \frac{25400 - 254 * CN}{CN} \qquad (21)$$

Donde Ia = abstracción inicial (mm) y CN = valor de numero de curva.

b Método de Transformación

El método de transformación en la subcuenca simula el proceso de lluvia a hidrograma de caudal. Este estudio se empleó el método Clark modificado (ModClark), es un método de hidrograma unitario lineal cuasi-distribuido que puede utilizarse con datos meteorológicos grilladas para la simulación en HEC-HMS, como método básico para las simulaciones distribuidas de precipitación-escorrentía (Bastia et al., 2021, Cho, 2020, Teng et al., 2018, Paudel et al., 2009, Kull and Feldman, 1998). Un modelo de parámetros distribuidos es aquel en el que la variabilidad espacial de las características y los procesos se consideran explícitamente (Bastia et al., 2021, Kull and Feldman, 1998). Este método requiere la estimación del tiempo de concentración (Tc) que se define como el tiempo que tarda el agua en desplazarse desde el punto hidráulicamente más alejado de la cuenca hasta la desembocadura (Paudel et al., 2009), y coeficiente de almacenamiento (R), es una medida del almacenamiento temporal del exceso de precipitaciones en la cuenca antes de que pueda drenar hacia el punto de salida (Sabol, 1988). En general, Tc puede estimarse conociendo la longitud, la pendiente y las propiedades superficiales de la trayectoria de flujo más larga, mientras que R puede estimarse con ecuaciones empíricas como algún múltiplo del Tc y luego ajustarse mediante calibración (Paudel et al., 2009, Kull and Feldman, 1998). Para estimar Tc y R en este procedimiento se utilizan las siguientes ecuaciones.

$$T_c = 0.0663 * S^{-0.385} L^{0.77} \qquad (22)$$

Donde T_c = tiempo de concentración (h) desarrollado por Kirpich; S = pendiente media del cauce (m/m) y L = longitud de cauce principal (km) (Alamri et al., 2023).

R puede calcularse como el caudal en el punto de inflexión en la curva descendente del hidrograma dividido por la derivada temporal del caudal, de acuerdo a la siguiente ecuación (Sabol, 1988).

$$\frac{A_t}{A} = \begin{cases} 1.414\left(\frac{t}{T_c}\right)^{1.5} & para\ t \le \frac{T_c}{2} \\ 1 - 1.414\left(1 - \frac{t}{T_c}\right)^{1.5} & para\ t \ge \frac{T_c}{2} \end{cases} \qquad (23)$$

Donde A_t = área acumulada de la cuenca que contribuye en el momento t; A = área total de la cuenca; T_c = tiempo de concentración de la cuenca.

c Método de Transito hidrológico

En este estudio se adopta el método Muskingum para el transito hidrológico en cauce (Chu and Chang, 2009, Natarajan and Radhakrishnan, 2021), y la ecuación (24).

$$Q_2 = (c_1 - c_2)\, I_1 + (1 - c_1) Q_1 + c_2 I_2 \qquad (24)$$

$$c_1 = \frac{2 * \Delta t}{2 * K\,(1 - X) + \Delta t} \qquad (25)$$

$$c_2 = \frac{\Delta t - 2 * K * x}{2 * K\,(1 - X) + \Delta t} \qquad (26)$$

I_1 e I_2 representan la entrada al cauce de inicio y final de los intervalos de cálculo, y Q_1 y Q_2 representan la salida del cauce al inicio y final de los intervalos de cálculo, respectivamente. K representa el tiempo de viaje a través del tramo, X es el factor de ponderación Muskingum que oscila entre (0 y 0.5), Δt es la longitud del intervalo de cálculo, y c_1 y c_2 son coeficientes dados en las ecuaciones (25 y 26) (Natarajan and Radhakrishnan, 2021).

El modelo desarrollado para la zona de estudio, se realizó siguiendo de acuerdo al esquema metodológico mostrado en la Figura 3. La cuenca fue subdivida en 14 subcuencas, como muestra la Figura 4. Los parámetros hidrológicos generados para todas las subcuencas de la cuenca Cañete se muestran en la Tabla 6. Además, se generaron 9 tramos de ríos, donde se realizaron transito hidrológico en cauce, el resumen de los parámetros de cada tramo de rio se muestra en la Tabla 7.

Tabla 6: Parámetros generales para cada subcuenca en Cuenca Cañete

Subbasin	Area (Km2)	Longest Flowpath Length (KM)	Longest Flowpath Slope (M/M)	Slope Subbasin (M/M)	Time of Concentration (HR)	Storage Coefficient (HR)
Subcuenca-1	275.14	29.111	0.057	0.383	2.680	1.340
Subcuenca-2	271.07	32.496	0.055	0.452	2.950	1.480
Subcuenca-3	448.29	39.722	0.046	0.446	3.710	1.860
Subcuenca-4	396.56	42.011	0.037	0.513	4.190	2.090
Subcuenca-5	419.69	44.227	0.058	0.486	3.680	1.840
Subcuenca-6	364.45	41.692	0.060	0.592	3.460	1.730
Subcuenca-7	586.43	49.257	0.068	0.566	3.750	1.880
Subcuenca-8	514.05	49.848	0.063	0.444	3.890	1.950
Subcuenca-9	616.20	71.373	0.055	0.468	5.430	2.710
Subcuenca-10	406.72	39.855	0.096	0.586	2.790	1.400
Subcuenca-11	517.31	51.772	0.074	0.406	3.780	1.890
Subcuenca-12	31.47	11.089	0.200	0.714	0.786	0.500
Subcuenca-13	450.50	48.966	0.072	0.576	3.660	1.830
Subcuenca-14	497.18	47.332	0.074	0.571	3.520	1.760

Tabla 7: Parámetros de cada tramo de rio para subcuenca en Cuenca Cañete

Reach	Length (KM)	Slope (M/M)	Muskingum K (HR)	Muskingum X
R1	19.998	0.015	1.689	0.250
R2	34.535	0.024	1.885	0.250
R3	19.349	0.023	1.399	0.300
R4	36.673	0.025	2.222	0.200
R5	10.669	0.014	1.078	0.300
R6	19.278	0.024	1.254	0.300
R7	7.679	0.019	0.715	0.300
R8	22.594	0.014	1.685	0.250
R9	24.799	0.013	1.830	0.200

2.2.4 Evaluación del rendimiento del modelo HEC-HMS

Varios estudiosos recomiendan el uso de funciones objetivo múltiples para medir los errores estadísticos entre los valores de escorrentía simulados y observados (Bhusal et al., 2022, Fanta and Tadesse, 2022, Nharo et al., 2019, Kumarasamy and Belmont, 2018). Par este estudio, se utilizaron las métricas estadísticas: coeficiente de Determinación (R^2), Raíz cuadrada del error cuadrático medio (RMSE), coeficiente de Nash-Sutcliffe (NSE), Porcentaje de Sesgo (PBIAS), Error porcentual en Volumen simulado (PEV) y Error en Pico simulado (PEP). Las ecuaciones y las correspondientes clasificaciones se presentan en la Tabla 8.

Tabla 8: Criterio para evaluar el rendimiento del Modelo HEC-HMS

Métricas estadísticas	Valores	Clasificación del rendimiento	Referencia
$R^2 = \frac{\left[\sum_{t=1}^{n}(Q_t^{obs} - \bar{Q}^{obs})(Q_t^{sim} - \bar{Q}^{sim})\right]^2}{\sum_{t=1}^{n}(Q_t^{obs} - \bar{Q}^{obs})^2 \sum_{t=1}^{n}(Q_t^{sim} - \bar{Q}^{sim})^2}$	r≥ 0.5 0 ≤ r < 0.5	Satisfactorio Malo	(Santhi et al., 2001)
$RMSE = \left[\frac{(\sum_{t=1}^{n} Q_t^{sim} - Q_t^{obs})^2}{n}\right]^{0.5}$	Valor inferior a la mitad de la desviación estándar	Satisfactorio	(Moriasi et al., 2007)
$NSE = 1 - \frac{\sum_{t=1}^{n}(Q_t^{obs} - Q_t^{sim})^2}{\sum_{t=1}^{n}(Q_t^{obs} - \bar{Q}^{sim})^2}$	0.75 < NSE ≤ 1.00 0.65 < NSE ≤ 0.75 0.50 < NSE ≤ 0.65 0.40 < NSE ≤ 0.50 NSE ≤ 0.40	Excelente Bueno Satisfactorio Aceptable Malo	(Moriasi et al., 2007)
$PBIAS = \frac{\sum_{t=1}^{n}(Q_t^{obs} - Q_t^{sim})}{\sum_{t=1}^{n}(Q_t^{obs})} * 100$	PBIAS < ±10 ±10 ≤ PBIAS <±15 ±15 ≤ PBIAS <±25 PBIAS ≥ ±25	Excelente Bueno Satisfactorio Malo	(Moriasi et al., 2007)
$PEV = \frac{Vol_t^{obs} - Vol_t^{sim}}{Vol_t^{obs}} * 100$	PEV > ±25 ±15 < PEV ≤ ±25 ±10 < PEV ≤ ±15 0 < PEV ≤ ±10 PEV = 0	Malo Satisfactorio Bueno Muy bueno Excelente	(Fanta and Tadesse, 2022, Natarajan and Radhakrishnan, 2021)
$PEP = \frac{Q_t^{p-obs} - Q_t^{p-sim}}{Q_t^{p-obs}} * 100$	PEP > ±25 ±15 < PEP ≤ ±25 ±10 < PEP ≤ ±15 0 < PEP ≤ ±10 PEP = 0	Malo Satisfactorio Bueno Muy bueno Excelente	(Belayneh et al., 2020)

Donde, *"Q^{sim} y Q^{obs}"* son los caudales simulados y observados, respectivamente; *"$\bar{Q}^{sim}$ y $\bar{Q}^{obs}$"* representa la media de cada uno, en el tiempo "t", "n" es el número de observaciones del periodo considerado, "Vol^{sim} y Vol^{obs}" son volumen simulados y observados (m^3), "Q^{p-sim} y Q^{p-obs}" son caudales picos simulados y observados.

3. Resultados y discusión

3.1. Interpolación Espacial

La calidad de los datos fue evaluada utilizando diferentes técnicas como graficas o métodos empíricos. El diagrama de caja (Figura 6), nos muestra gráficamente una serie de datos numéricos a través de sus cuartiles, y también se observan los probables valores atípicos presente en cada estación.

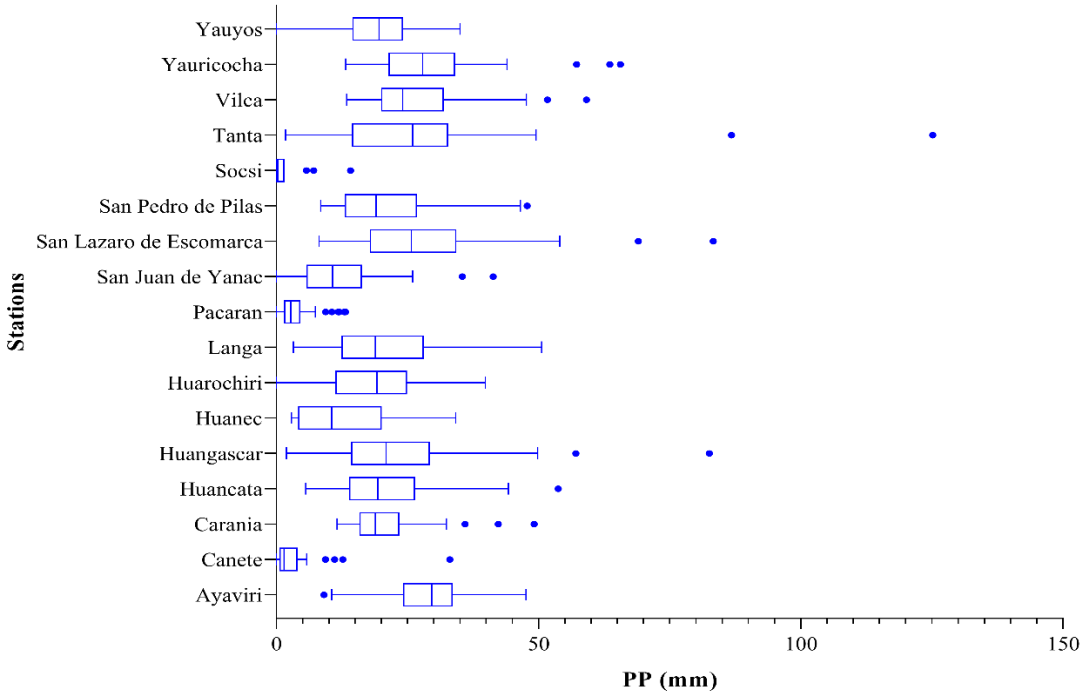


Figura 6: Diagrama de caja por estaciones pluviométricas de precipitaciones máximas en 24 horas

3.1.1 Selección de estaciones en base a la calidad de información

Desde las estaciones pluviométricas se obtuvieron los valores de precipitaciones máximas diarias anuales y estas fueron transformadas a precipitaciones máximas en 24 horas en base a las ecuaciones (2) y (3). Siguiendo con el proceso de análisis de los datos, se realizó la prueba de normalidad utilizando métodos gráficos (gráficos de caja y gráficos cuantiles normales) y pruebas formales de normalidad (AD, SW y KM, ecuaciones (5), (6) y (8) respectivamente). Además, para identificar y eliminar valores atípicos se aplicaron los métodos de Tukey y Z-score (ecuaciones (9) y (12) respectivamente). Los resultados de la prueba de normalidad al 95% de nivel de confianza ($\alpha = 0.05$), se muestran en la Tabla 9. Además, el análisis de la prueba

de normalidad fue implementado mediante el lenguaje de programación Python (Van Rossum and Drake, 2009)).

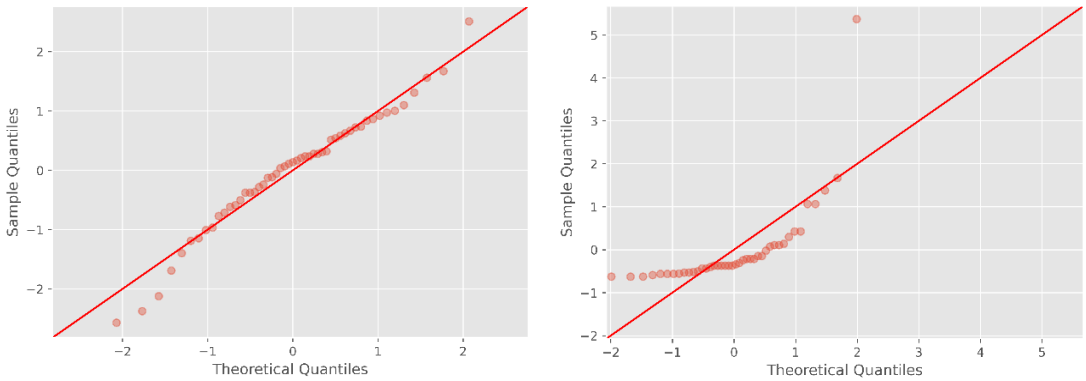
Tabla 9: Análisis de valores atípicos y prueba de normalidad al 95% nivel de confianza

ID	Stations	Iteration	No of Data	Outliers method	Anderson-Darling α = 0.05		Shapiro-Wilk α = 0.05		Kolmogorov-Smirnov α = 0.05		Normality test
					AD* statistic	AD* calculated	SW statistic	SW calculated	D* statistic	D* calculated	
1	Ayaviri	1	51	-	0.736	0.444	0.977	0.428	0.085	0.475	Yes
2	Cañete	1	37	Tukey/ Z-score	0.722	1.362	0.888	0.001	0.182	0.004	Not
3	Carania	2	52	Tukey	0.752	0.766	0.948	0.024	0.095	0.293	Yes
4	Huancata	2	39	Tukey	0.725	0.674	0.942	0.044	0.133	0.082	Yes
5	Huangascar	2	53	Tukey	0.738	0.549	0.965	0.116	0.094	0.305	Yes
6	Huañec	2	54	Z-score	0.739	1.376	0.920	0.001	0.135	0.018	Not
7	Huarochiri	1	55	-	0.739	0.307	0.981	0.522	0.072	0.697	Yes
8	Langa	1	40	-	0.726	0.481	0.955	0.113	0.105	0.323	Yes
9	Pacaran	2	42	Tukey	0.728	0.443	0.953	0.085	0.086	0.595	Yes
10	San Juan de Yanac	2	51	Tukey	0.736	0.403	0.965	0.141	0.077	0.639	Yes
11	San Lazaro de Escomarca	2	53	Tukey	0.738	0.344	0.972	0.247	0.093	0.325	Yes
12	San Pedro de Pilas	2	34	Z-score	0.718	0.666	0.934	0.041	0.148	0.061	Yes
13	Socsi	1	11	Tukey/ Z-score	0.680	3.042	0.541	0.000	0.461	0.001	Not
14	Tanta	2	53	Tukey	0.752	0.746	0.962	0.089	0.134	0.021	Yes
15	Vilca	2	53	Tukey	0.738	0.593	0.955	0.046	0.111	0.114	Yes
16	Yauricocha	2	31	Tukey	0.713	0.247	0.970	0.517	0.080	0.882	Yes
17	Yauyos	1	48	-	0.734	0.362	0.977	0.464	0.067	0.850	Yes

Las iteraciones realizadas en la prueba de normalidad (Tabla 9), la máxima iteración fue 2. Las estaciones que se presentaron 1 iteración son aquellas que fueron procesadas con el total de sus valores, por lo tanto, no se aplicaron método de análisis de valores atípicos. Sin embargo, las estaciones que presentan 2 iteraciones, son aquellas que sufrieron eliminación de valores en comparación a sus valores iniciales, por lo tanto, se aplicaron uno u otro método de análisis de valores atípicos. La prueba de normalidad es básicamente el proceso de realizar la comparación de los valores estadísticos (teóricos) versus los valores calculados. Los métodos utilizados en prueba de normalidad formales como Anderson-Darling, Shapiro-Wilk y Kolmogorov-Smirnov muestran valores estadísticos $AD^*_{statistic}$, $SW_{statistic}$, y $D^*_{statistic}$ respectivamente, representan

valores teóricos y son comparados con sus respectivos valores calculados $AD^*_{calculated}$, $SW_{calculated}$, $D^*_{calculated}$. Las estaciones que pasan la prueba de normalidad mediante el método de AD, sus valores $AD^*_{calculated}$ deben ser menor que $AD^*_{statistic}$ (Bayoud, 2021). Para SW los valores $SW_{calculated}$, deben estar entre cero y uno teóricamente. Sin embargo, los valores máximos para cada estación en particular están representadas por $SW_{statistic}$, se debe cumplir dos criterios: primero, que $SW_{calculated}$, debe ser menor que $SW_{statistic}$ y segundo, que $SW_{calculated}$ no tener valores pequeños, no cumplir este criterio, se rechaza la normalidad bajo el método SW (Razali et al., 2011). Mediante el método KS al 95% de nivel de confianza, para aceptar la prueba de normalidad los valores de $D^*_{calculated}$ debe estar por debajo de 0.895, además tener valores mayores a $D^*_{statistic}$ (Bayoud, 2021, Stephens, 1974)

La gráfica cuantiles normales, visualmente nos permite observar los valores de precipitación ploteado alrededor o cercano a una línea recta, al presentarse de esta forma es posible afirmar que se tiene un ajuste una distribución normal teórica. En la Figura 7a se observa a la estación Ayaviri, la misma que muestra un buen ajuste a la prueba de normalidad. La figura 7b a la estación Cañete, la cual muestra gráficamente que no pasa la prueba de normalidad. La Figura 7c muestra los valores de la estación Huancata, en un primer análisis no pasa la prueba de normalidad, sin embargo, al realizar el análisis para detectar y eliminar valores atípicos y realizado nuevamente la prueba de normalidad la gráfica se ajusta a la línea recta (Figura 7d). Las demás graficas cuantiles normales, correspondientes a las otras estaciones, se adjunta en anexos. Finalmente, los resultados en esta etapa revelaron que los métodos de prueba de normalidad formales suelen ser más objetivas ante las discriminaciones visuales basadas en gráficos (Bayoud, 2021). Por lo tanto, el resumen de los resultados se muestra en la Tabla 9, Cañete, Huañec y Socsi fueron las estaciones que no se ajustaron a la prueba de normalidad.



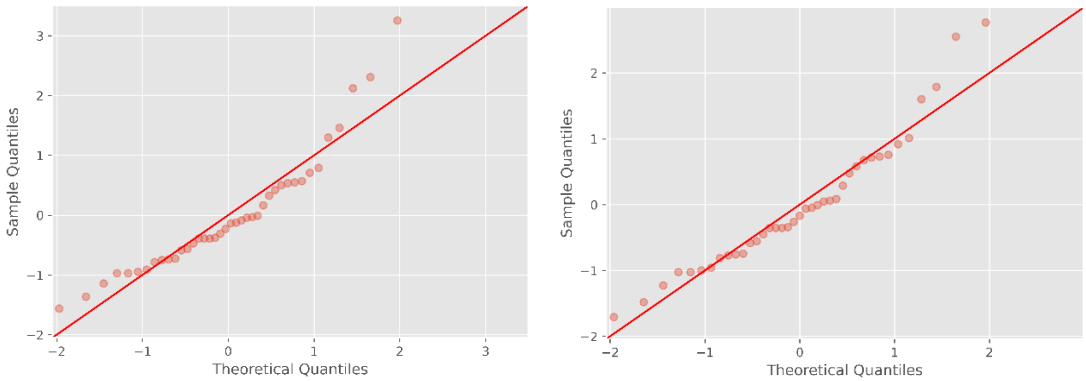


Figura 7: Graficas de cuartiles normales de las estaciones a) Ayaviri b) Cañete c) Hancata (primer análisis) d) Huancata (segundo análisis)

3.1.2 Análisis de función de distribución de probabilidades

Después de realizado el control de calidad de las precipitaciones máximas en 24 horas (prueba de normalidad y eliminación de valores atípicos), como resultado se obtuvieron que 14 de 17 estaciones pasaron las pruebas de control de calidad. El análisis de función de distribución se realizó a las 14 estaciones que pasaron la prueba de normalidad para periodos de retorno de 5, 10, 20, 50 y 100 años. Así mismo, para validar los resultados obtenidos al aplicar las funciones de distribuciones, se realizaron prueba de bondad de ajuste mediante el método KS, y se seleccionaron las funciones de distribución que mejores se ajustaron en cada estación. La Tabla 10 presenta las funciones de distribución teórica de mejor se ajustó de cada estación, para precipitaciones con Tr 5, 10, 20, 50 y 100 años, (valores finales después del análisis de prueba de bondad de ajuste). El análisis de distribución de frecuencias y prueba de bondad de ajuste fueron realizado con el programa hidrológico Hydrognomon (http://hydrognomon.org/).

Tabla 10: Función de distribución teórica y valores de precipitaciones para periodos de retornos de 5, 10, 20, 50 y100 años

ID	Stations	Distribution Function	PP (mm)				
			Tr 5 year	Tr 10 year	Tr 20 year	Tr 50 year	Tr 100 year
1	Ayaviri	GEV-Max	35.43	38.22	40.19	42.00	42.96
3	Carania	EV1-Max	22.93	25.94	28.82	32.55	35.34
4	Huancata	GEV-Max	26.77	32.04	37.12	43.75	48.74
5	Huanguascar	GEV-Max	30.02	35.75	41.02	47.53	52.18
7	Huarochiri	Normal	26.58	30.61	33.94	37.68	40.18
8	Langa	GEV-Max	28.84	35.06	40.85	48.08	53.31
9	Pacaran	Pearson III	4.41	5.64	6.78	8.17	9.18
10	San Juan de Yanac	GEV-Max	16.41	19.94	22.96	26.39	28.66
11	San Lazaro de Escomarca	GEV-Max	34.48	40.26	45.30	51.15	55.10
12	San Pedro de Pilas	Pearson III	26.04	30.61	34.72	39.71	43.26
14	Tanta	GEV-Max	34.17	40.30	45.51	51.40	55.26
15	Vilca	GEV-Max	31.69	36.17	40.32	45.49	49.21
16	Yauricocha	Normal	33.59	37.19	40.17	43.52	45.75
17	Yauyos	GEV-Max	25.52	29.07	32.01	35.25	37.32

Un aspecto muy importante en el análisis de distribución de frecuencias en hidrología es identificar una distribución estadística que mejor se ajuste a un conjunto de datos observados, entre varios modelos probabilísticos. Para ajustar estas variables, se realizó las pruebas de bondad de ajuste entre los diferentes modelos, se selecciona varios modelos candidatos al conjunto de datos, a menudo tenemos que decidir o discriminar entre los diferentes modelos competitivos, eligiendo uno por cada estación. La Tabla 10 muestra los modelos de mejor ajuste para cada estación. De mismo modo, el diagrama de caja (Figura 8), muestra valores fuera del rango del diagrama, que podría interpretarse como valores atípicos. Sin embargo, corresponde a Pacaran, estación que se ubica en la parte suroeste de la cuenca cañete y corresponde a parte seca de cuenca, región 3 de (Portuguez-Maurtua et al., 2022).

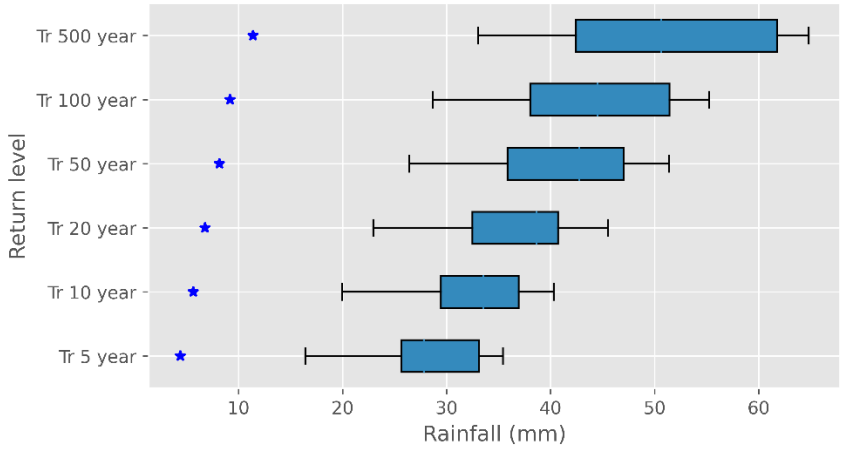


Figura 8: Características de los seis niveles de retorno de precipitaciones

3.1.3 Interpolación espacial de las precipitaciones

La técnica KO requieren funciones de variograma teóricas ajustadas (Bárdossy et al., 2021). El variograma desempeña un papel fundamental en la presentación de la variación espacial de una variable de interpolación (Das, 2021). En este estudio se utilizó un modelo de variograma experimental isótropo a partir del conjunto de datos de precipitación para cada Tr, asumiendo una correlación espacial idéntica en todas las direcciones y despreciando la influencia de la anisotropía en los parámetros del variograma (Adhikary et al., 2017). El variograma experimental y los parámetros óptimos que permita que el modelo prediga con errores mínimo se muestran en la Tabla 11 para cada conjunto de datos. El semivariograma teórico que mejor se ajustó fue el modelo lineal, los valores del parámetro nugget es 0, sill toma valores entre 98.81 y 254.26 (mínimo y máximo respectivamente) y el valor del rango va entre 36.91 a 60.22 km. Los cálculos y optimización del variograma, y el proceso de interpolación KO fueron realizados mediante el programa QGIS 3.18 bajo el complemento Smart-Map Plugin (SMP), plugin gratuito de código abierto (Pereira et al., 2022).

El variograma experimental explica como la distancia máxima y la ubicación espacial de una estación influye sobre otra estación a diferentes distancias (autocorrelacion espacial). Por lo tanto, las observaciones cercanas tienen mayores probabilidades de ser similares (Javari, 2016). La selección de los modelos de variograma se realizaron mediante un proceso de ensayo y error con SMP, eligiéndose los que presentaron mejor ajuste a través de la cuantificación de los valores de R^2 (Tabla 11). Los parámetros del variograma fueron modificados de forma iterativa para obtener el modelo mejor ajustado, que produce el mejor valor de R^2, cercano a 1. Los

valores de R^2 oscila entre 0.80 a 0.84, para Tr más corto (5 años) presenta mayor R^2, caso contrario sucede con Tr 50 y 100 años, presentaron menor R^2.

Tabla 11: Resultados de los parámetros de los modelos de variogramas ajustado para diferente periodo de retorno (Tr)

Return period	Model name	Variogram parameters			R^2
		Nugget, C_0 (mm^2)	Sill, $C_0 + C_1$ (mm^2)	Range, a (km)	
5 year	Linear	0.00	98.81	51.14	0.84
10 year	Linear	0.00	114.75	47.15	0.82
20 year	Linear	0.00	162.23	55.17	0.81
50 year	Linear	0.00	134.74	36.91	0.80
100 year	Linear	0.00	254.26	60.22	0.80

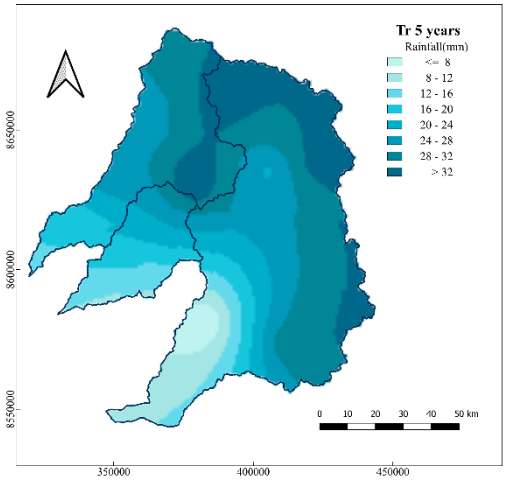
Explorar la autocorrelación espacial de una variable de interpolación mediante el análisis de los variograma es crucial al aplicar KO, obteniendo estimaciones de puntos desconocidos (Zou et al., 2021). Como se describe líneas arriba el proceso de interpolación KO se realizó mediante el complemento SMP en QGIS. Las principales estadísticas de cada cuencas y total de la zona de estudio se muestran en la Tabla 12. La intensidad media de la precipitación para Tr 5 años osciló entre 27.09, 19.09 y 25.15 mm (Mala, Omas y Cañete respectivamente), con una precipitación media de 24.92 mm para MOC, disminuyendo desde noreste a suroeste. La precipitación máxima osciló entre 35.38, 32.19 y 34.13 mm, (Mala, Omas y Cañete respectivamente), con precipitación media de valor máximo de 35.38 mm para MOC. La precipitación mínima osciló entre 14.01, 11.28 y 4.64 mm, (Mala, Omas y Cañete respectivamente), con una precipitación media de valor mínimo de 4.64 mm para MOC. La Tabla 12 presenta los valores mínimos, máximos, medias y desviación estándar para cada cuenca y total (MOC) para Tr 10, 20, 50 y 100 años. En general, para Tr 5 años la cuenca Mala presenta el mayor valor de precipitación máxima y la cuenca Cañete presento los mayores valores de precipitación máxima para Tr 10, 20, 50, 100 y 500 años. Además, Cañete también muestra los valores más bajo de precipitación mínima en todos los Tr. En particular, las estaciones que se ubican por encima de los 2000 msnm. mostraron valores altos de precipitaciones, caso contrario sucede a las estaciones por debajo que tiende a valores bajos o escaso.

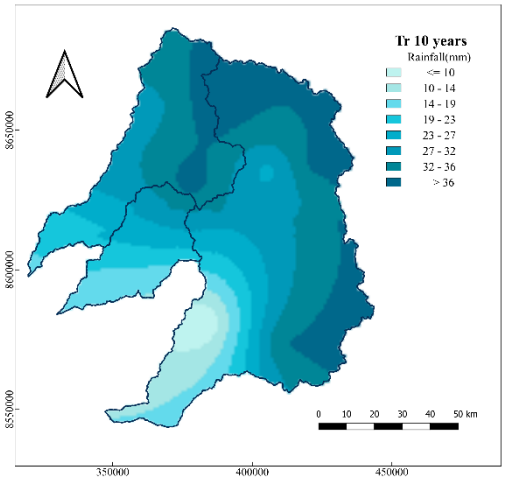
Tabla 12: Estadística de precipitaciones para las cuencas Mala, Omas, Cañete y área total (MOC)

Return period	Mala Watershed				Omas Watershed				Cañete Watershed				MOC Watersheds			
	Min	Max	Mean	Std dev	Min	Max	Mean	Std dev	Min	Max	Mean	Std dev	Min	Max	Mean	Std dev
5 year	14.01	35.38	27.09	4.85	11.28	32.19	19.09	5.24	4.64	34.13	25.15	7.74	4.64	35.38	24.92	7.23
10 year	17.26	38.89	31.48	5.07	13.77	35.39	22.64	5.65	5.92	40.25	29.11	8.58	5.92	40.25	28.94	7.95
20 year	20.34	43.80	35.34	5.19	16.05	37.88	25.83	5.97	7.10	45.45	32.63	9.34	7.10	45.45	32.51	8.57
50 year	24.27	49.38	39.94	5.28	18.84	41.36	29.72	6.32	8.55	51.33	36.84	10.25	8.55	51.33	36.78	9.31
100 year	27.22	53.06	43.17	5.34	20.85	44.23	32.49	6.55	9.60	55.19	39.77	10.90	9.60	55.19	39.76	9.83

La distribución espacial de la precipitación para diferente Tr (5, 10, 20, 50 y 100 años) son presentados en las Figura 9 (a, b, c, d y e, respectivamente). Del análisis de la distribución espacial de la precipitación, las mayores concentraciones de precipitaciones se producen en la parte alta (nor-este y sureste) y va descendiendo en dirección a la parte baja (sur-oeste) para todo los Tr. Además, también se aprecia que las precipitaciones mínimas o escasas se ubican en la parte baja de la cuenca Cañete, esto debido a estar situado en la región más larga situada entre la llanura costera y las estribaciones de los Andes occidentales, considerado como más seca del país (Rau et al., 2017).

Los regímenes de precipitaciones varían espacialmente en todos los lugares. La situación geográfica del lugar y el entorno que lo rodea son factores importantes en la variación del régimen pluviométrico. En la parte nororiental se registra la mayor cantidad de precipitaciones, mientras que en la parte sudoccidental se registra la menor cantidad, estos patrones se presentan para los 5 Tr analizadas.





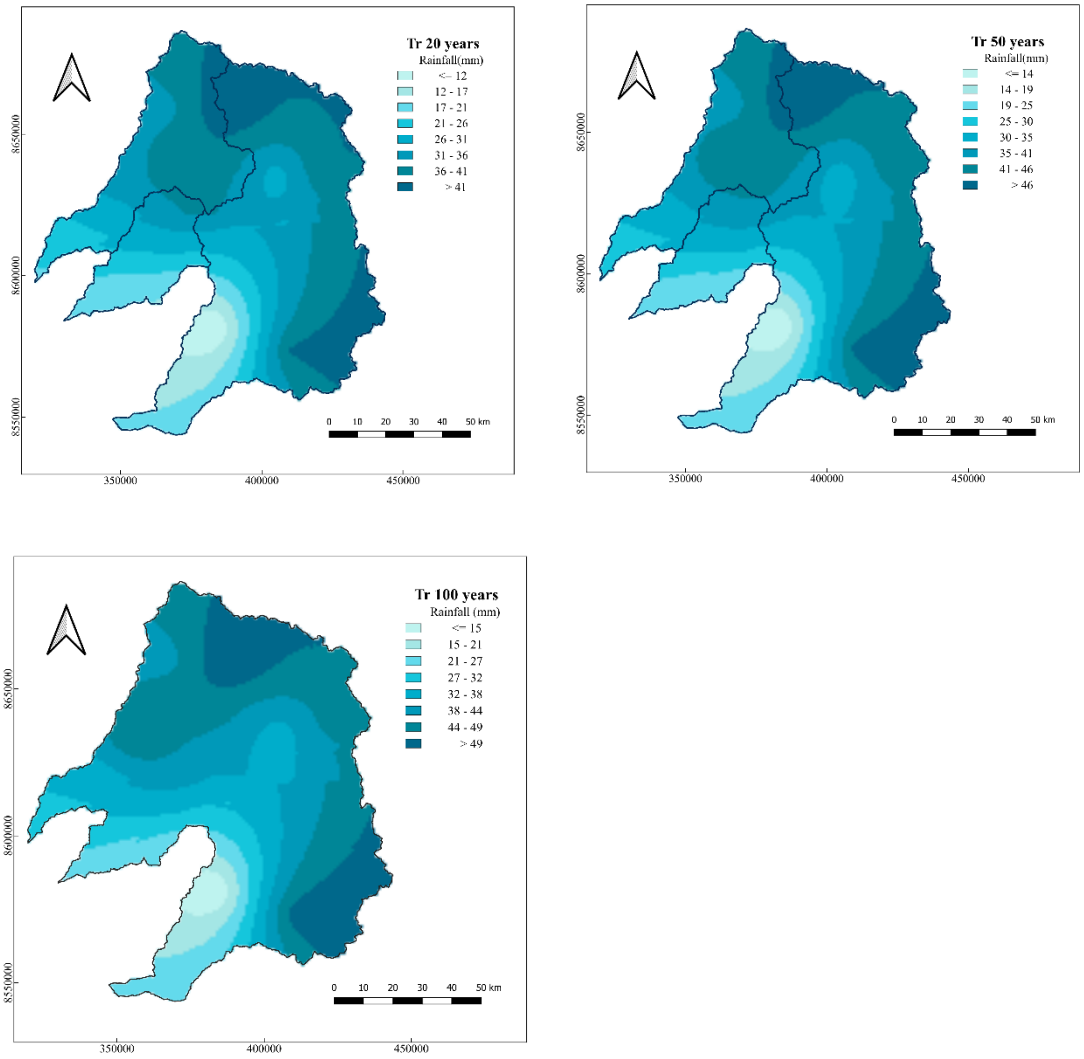


Figura 9: Mapas de isoyetas a Tr a) 5 años b) 10 años c) 20 años d) 50 años y e) 100 años.

3.1.4 Precisión de la interpolación

La interpolación de las precipitaciones fue comprobada mediante la evaluación de métricas estadísticas basadas en R^2, RMSE, NSE y Pbias (Adhikary et al., 2017). La Tabla 13 presenta las diferentes medidas de rendimiento del método de interpolación para estimar la precipitación para diferentes Tr. en las cuencas de estudio. Los valores más pequeños de RMSE, aceptable valor inferior a la mitad de la desviación estándar (Moriasi et al., 2007), y con un valor más alto de R^2 recomendado mayor a 0.5 es considerado satisfactorio (Santhi et al., 2001). Las métricas y su clasificación son presentadas en la Tabla 2, todos ellos son indicadores de una buena o mala predicción por el método KO.

Las estadísticas indican un ajuste excelente, con valores de NSE desde 0.85 a 0.91 (Tr 100 años y Tr 5 años, respectivamente), clasificándose como excelente. Los valores de PBIAS se encuentran en 2.78 a 3.97 (Tr 5 años y Tr 100 años respectivamente), los valores predichos con respecto al original son subestimadas en menos de 10%. RMSE los valores varían entre 2.32 a 4.50 (Tr 5 años y Tr 100 años, respectivamente) estos valores se ubican por debajo a la mitad de la desviación estándar y finalmente los valores de R2 son valores altos en todos los Tr. Los valores de las métricas estadísticas que permite evaluar la eficiencia de la interpolación mediante kriging ordinario son mostrado en la Tabla 13. Por lo tanto, de acuerdo a los resultados obtenidos de las métricas estadísticas, KO es una alternativa de interpolación eficiente para la generación de isoyetas.

Tabla 13: Valores de métricas estadísticas de la interpolación espacial mediante kriging ordinario

Metrics	Tr 5 year	Tr 10 year	Tr 20 year	Tr 50 year	Tr 100 year
R^2	0.92	0.91	0.89	0.88	0.88
RMSE	2.32	2.85	3.37	4.02	4.50
NSE	0.91	0.89	0.88	0.86	0.85
PBIAS (%)	2.78	3.21	3.48	3.78	3.97

3.2 Modelación hidrológica

Para el desarrollo del modelo hidrológico se consideraron tres modelos para las subcuencas y cauce, que incluyen el Modelo de Pérdidas, Modelo de Transformación y Modelo de transito de cauce.

3.2.1 Resultado inicial de la simulación con HEC-HMS

La simulación inicial (sin calibración) se realizó utilizando los parámetros iniciales de la cuenca (Tabla 5 y 6), Curva número grillada (Figura 5) y precipitaciones extremas grilladas (Figura 9). Las Figuras 10a y 10b muestran la implementación del modelo hidrológico utilizando precipitación de estaciones pluviométricas y precipitación grillada respectivamente. Los valores de caudales observados y simulados en la simulación inicial fueron evaluados mediante métricas estadísticos, la Tabla 14 muestra las métricas de la modelación hidrológica usando precipitación de estaciones pluviométricas y precipitaciones grilladas.

Las métricas bajo la modelación hidrológica usando precipitación de estaciones pluviométricas (Tabla 14), los valores de R^2 oscilan entre 0.77 a 0.97, el menor y mayor valor son presentado

por Tr 5 años y Tr 100 años respectivamente, teniendo rendimiento satisfactorio, de acuerdo a la clasificación presentada por (Santhi et al., 2001, Moriasi et al., 2007). RMSE oscila entre 49.35 m^3/s y 94.05 m^3/s, el menor y mayor valor son presentado por Tr 100 años y Tr 5 años respectivamente; para ser considerado de rendimiento satisfactorio los valores de RMSE deben presentar valores menores a la mitad de la desviación estándar de los datos simulados, solo los Tr 50 y 100 años fueron satisfactorio (Moriasi et al., 2007). Para NSE los valores oscilan entre 0.39 y 0.95, los Tr 20, 50 y 100 presentan rendimiento excelente por tener valores por encima de 0.75, Tr 10 años de rendimiento bueno y Tr 5 años de rendimiento satisfactorio (Moriasi et al., 2007). Los valores de PBIAS oscilan entre 15.95 % a 63.55 %, clasificado como malo Tr 5, 10 y 20 años y satisfactorio Tr 50 y 100 años, según (Moriasi et al., 2007), en general los resultados de PBIAS muestran la subestimación o sobrestimación de los caudales, por lo tanto, el modelo subestima los caudales en comparación a los caudales observados. PEV varía entre 15.85 % a 63.58 %, el menor y mayor valor es para Tr 100 años y Tr 5 años, respectivamente. Los resultados de Tr 50 y 100 años son clasificado como satisfactorio y el resto de Tr son clasificados como malo, esta métrica cuantifica la subestimación o sobreestimación del volumen de caudales estimados, en general el modelo tiende a subestimar el volumen de caudal. Finamente, PEP oscilan entre 8.74 % a 53.93 %, Tr 100 años es de rendimiento muy bueno, el modelo explica que los valores de caudales picos observado y simulado están por debajo de ±10. Tr 50 años su rendimiento es satisfactorio y para Tr 20, 10 y 5 años los resultados revelan rendimiento malo.

Las métricas bajo la modelación hidrológica usando precipitación grillada (Tabla 14), los valores de R^2 oscilan entre 0.72 a 0.82, Tr 5 años muestra el menor valor y Tr 100 años el mayor valor, todos los valores de R^2, fueron clasificados como satisfactorio, de acuerdo a la clasificación presentada por (Santhi et al., 2001, Moriasi et al., 2007). RMSE oscila entre 69.26 m^3/s y 81.57 m^3/s, el menor y mayor valor son Tr 100 años y Tr 5 años, respectivamente. Los valores de RMSE de Tr 50 y 100 años son menores a la mitad de la desviación estándar, clasificándose como satisfactorio y el resto como no satisfactorio (Moriasi et al., 2007). Para NSE los valores oscilan entre 0.62 y 0.91, los Tr 20, 50 y 100 se clasifican con rendimiento excelente por tener valores por encima de 0.75, Tr 10 años de rendimiento bueno y Tr 5 años de rendimiento satisfactorio (Moriasi et al., 2007). Los valores de PBIAS oscilan entre 11.24 % a 49.56 %, clasificado como malo Tr 5, 10 y 20 años, como satisfactorio Tr 50 y bueno Tr 100, según (Moriasi et al., 2007). PEV varía entre 11.15 % a 49.45 % el menor y mayor valor es para

Tr 100 años y Tr 5 años, respectivamente. Los resultados de Tr 100 y 50 años son clasificados como bueno y satisfactorio respectivamente, esta métrica nos cuantifica la subestimación del volumen de caudales estimados, en concordancia a los porcentajes dado. Finamente PEP oscilan entre 4.52 % a 40.49 %, Tr 50 y 100 años, mostraron rendimiento muy bueno, el modelo explica que los valores de caudales picos observado y simulado están por debajo de ±10. Tr 20 años su rendimiento es satisfactorio y para Tr 10 y 5 años los resultados revelan rendimiento malo.

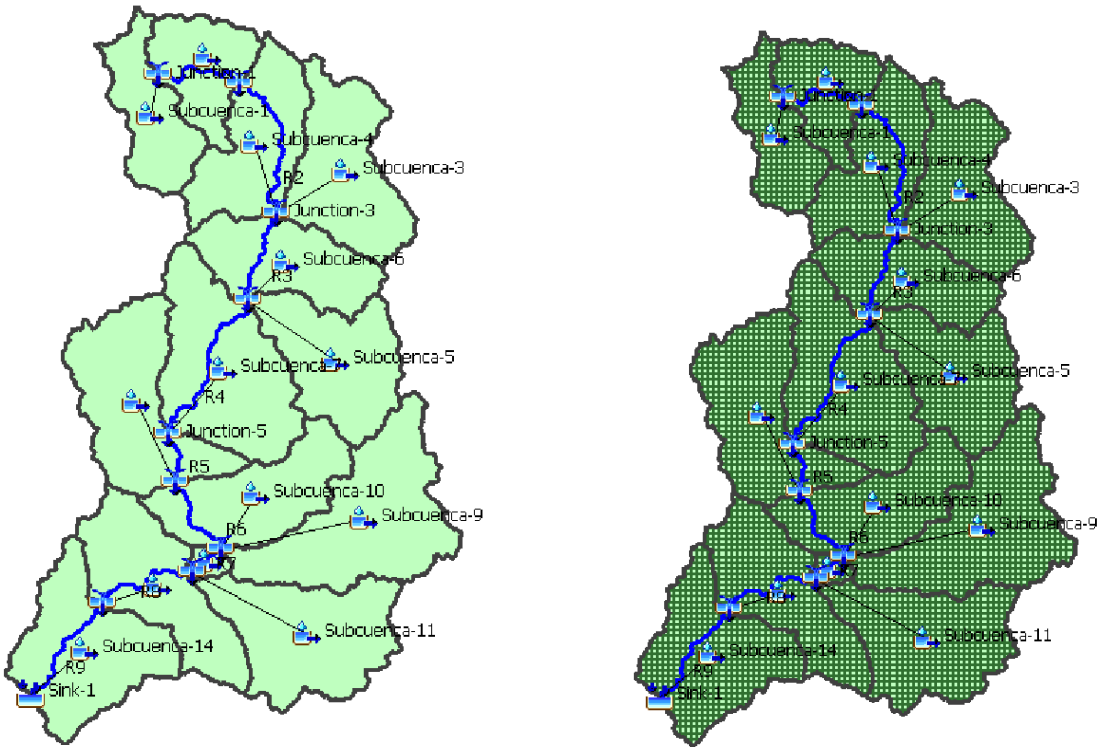


Figura 10: Esquema del modelo hidrológico implementado en HEC-HMS a) con subunidades hidrográficas utilizando precipitación desde estaciones pluviométricas y b) utilizando precipitación grillada

Analizando la comparación de los resultados en base a la evaluación de las métricas de la modelación hidrológica usando precipitación de estaciones pluviométricas y precipitaciones grilladas en la simulación inicial en el punto de control, estación hidrométrica Socsi, los resultados obtenidos mediante el uso de precipitación grilladas en la modelación hidrológica, muestran mejores resultados en casi todas las métricas excepto en R^2. Sin embargo, los resultados de caudales picos simulados y PEP muestran mejores performances, es un indicador muy importante por tratarse de modelación hidrológica de evento.

Tabla 14: Métricas estadísticas de simulación inicial en la estación hidrométrica Socsi

Modelación Hidrológica con:	Métricas	Tr 5 años	Tr 10 años	Tr 20 años	Tr 50 años	Tr 100 años
Precipitaciones Estaciones Pluviométricas	$Q_{pico\ observado}$ (m3/s)	387.02	466.71	544.74	648.29	728.03
	$Q_{pico\ simulado}$ (m3/s)	178.30	288.10	400.60	550.10	664.40
	R^2	0.77	0.87	0.92	0.95	0.97
	RMSE (m3/s)	94.05	85.44	73.90	58.31	49.35
	NSE	0.39	0.66	0.81	0.92	0.95
	PBIAS (%)	63.55	48.39	36.10	23.39	15.95
	PEV (%)	63.58	48.35	36.05	23.32	15.85
	PEP (%)	53.93	38.27	26.46	15.15	8.74
Precipitaciones Grilladas	$Q_{pico\ observado}$ (m3/s)	387.02	466.71	544.74	648.29	728.03
	$Q_{pico\ simulado}$ (m3/s)	230.30	335.50	439.90	586.40	695.10
	R^2	0.72	0.80	0.85	0.89	0.92
	RMSE (m3/s)	81.57	78.24	74.62	70.24	69.26
	NSE	0.62	0.74	0.82	0.88	0.91
	PBIAS (%)	49.56	36.79	27.26	16.90	11.24
	PEV (%)	49.45	36.63	27.27	16.86	11.15
	PEP (%)	40.49	28.11	19.25	9.55	4.52

3.2.2 Análisis de sensibilidad y calibración

La calibración del modelo consiste en ajustar determinados parámetros del modelo hasta que los resultados coincidieran aceptablemente con los datos observados (Bastia et al., 2021). El modelo se calibró comparando el hidrograma de los datos simulados (precipitación desde estaciones y grilladas) y observadas. Además, la evaluación del rendimiento del modelo HEC-HMS incluye el análisis de sensibilidad y la calibración, ambos procesos se realizaron simultáneamente. Mediante el análisis de sensibilidad se identificaron los parámetros causantes de la mayor escorrentía. Se realizaron varias corridas de ensayos de optimización a través del proceso de calibración (Fanta and Tadesse, 2022, Tassew et al., 2019, Kousari et al., 2010). Estudios de investigaciones anteriores para el análisis de sensibilidad han utilizado diferentes rangos de parámetros (Fanta and Tadesse, 2022, Tassew et al., 2019, Zelelew and Melesse, 2018, Majidi and Shahedi, 2012). En este estudio, el análisis de sensibilidad se realizó cambiando los valores de los parámetros en un rango del 25% con intervalos del 5%, recomendado por (Fanta and Tadesse, 2022, Tassew et al., 2019). El grado de sensibilidad de cada parámetro se basó en la magnitud del cambio en el volumen total de escorrentía a la salida

de la cuenca (estación Socsi). La calibración se realizó mediante la auto-calibracion disponible en HEC-HMS (optimization trials). La selección de los parámetros se realizó en función de su efecto sobre la descarga máxima y el volumen total (Cahyono and Adidarma, 2019). En el proceso de optimización en HEC-HMS se eligió el método Simplex y la función objetivo Peak-Weighted RMSE, esto debido a su simplicidad y rendimiento. Los parámetros característicos seleccionados fueron CN, Tc, R, K y X, principalmente.

3.2.3 Resultado de la calibración de HEC-HMS

En un modelo hidrológico aplicado a inundaciones, el aspecto más importante del hidrograma es el caudal pico, ya que éste corresponde a la inundación máxima aguas abajo. Los resultados de este estudio mostraron un ajuste razonable entre el caudal modelado y observado después de la optimización (calibración); la forma del hidrograma y el momento de los picos para los eventos de los cinco periodos de retorno (Figura 11). Además, se observa que en la mayoría de Tr, la forma del hidrograma se reprodujo con exactitud en la salida de la cuenca. La Tabla 15 presenta los resultados de la evaluación de las métricas estadísticas durante la calibración.

Tabla 15: Métricas estadísticas de simulación calibrada en la estación hidrométrica Socsi

Modelación hidrológica con:	**Métricas**	**Tr 5 años**	**Tr 10 años**	**Tr 20 años**	**Tr 50 años**	**Tr 100 años**
Precipitaciones Estaciones Pluviométricas	$Q_{pico\ observado}$ (m3/s)	387.02	466.71	544.74	648.29	728.03
	$Q_{pico\ simulado}$ (m3/s)	238.30	351.20	413.60	549.30	642.10
	R^2	0.93	0.92	0.90	0.87	0.85
	RMSE (m3/s)	75.29	69.87	81.27	86.41	96.88
	NSE	0.74	0.84	0.84	0.88	0.88
	PBIAS (%)	58.12	43.78	41.12	31.16	25.89
	PEV (%)	64.46	43.77	41.07	31.09	25.82
	PEP (%)	38.43	24.75	24.07	15.27	11.80
Precipitaciones Grilladas	$Q_{pico\ observado}$ (m3/s)	387.02	466.71	544.74	648.29	728.03
	$Q_{pico\ simulado}$ (m3/s)	371.90	455.10	523.10	646.10	725.70
	R^2	0.90	0.88	0.88	0.87	0.87
	RMSE (m3/s)	41.09	54.27	62.54	79.77	91.01
	NSE	0.88	0.86	0.86	0.84	0.84
	PBIAS (%)	12.61	10.04	11.87	7.88	7.71
	PEV (%)	12.58	9.89	11.76	7.77	7.63
	PEP (%)	3.91	2.49	3.97	0.34	0.32

Las métricas de la modelación hidrológica calibrada usando precipitación de estaciones pluviométricas (Tabla 15), los valores de R^2 oscilan entre 0.85 a 0.93, clasificándose como satisfactorios. RMSE oscila entre 75.29 m^3/s y 96.88 m^3/s. Los valores de NSE oscilan entre 0.74 y 0.88, muestran rendimiento excelente por tener valores por encima de 0.75, a excepción de Tr 5 años. Los valores de PBIAS oscilan entre 25.89% a 58.12%, clasificándose como Malo (Tr 5, 10, 20, 50 y 100 años), en general estos resultados mostraron una subestimación de los resultados de los caudales simulados. PEV varía entre 25.82% a 64.46%, los resultados se clasifican desde Satisfactorio a Malo, en general el modelo tiende a subestimar el volumen de caudal. Finamente PEP oscilan entre 11.80% a 38.43%, mostraron rendimiento bueno (Tr 100 años), satisfactorio (Tr 10, 20 y 50 años) y malo para Tr 5 años. Los caudales picos simulado están por debajo de caudal pico observado (Figura 11 y Figura 12).

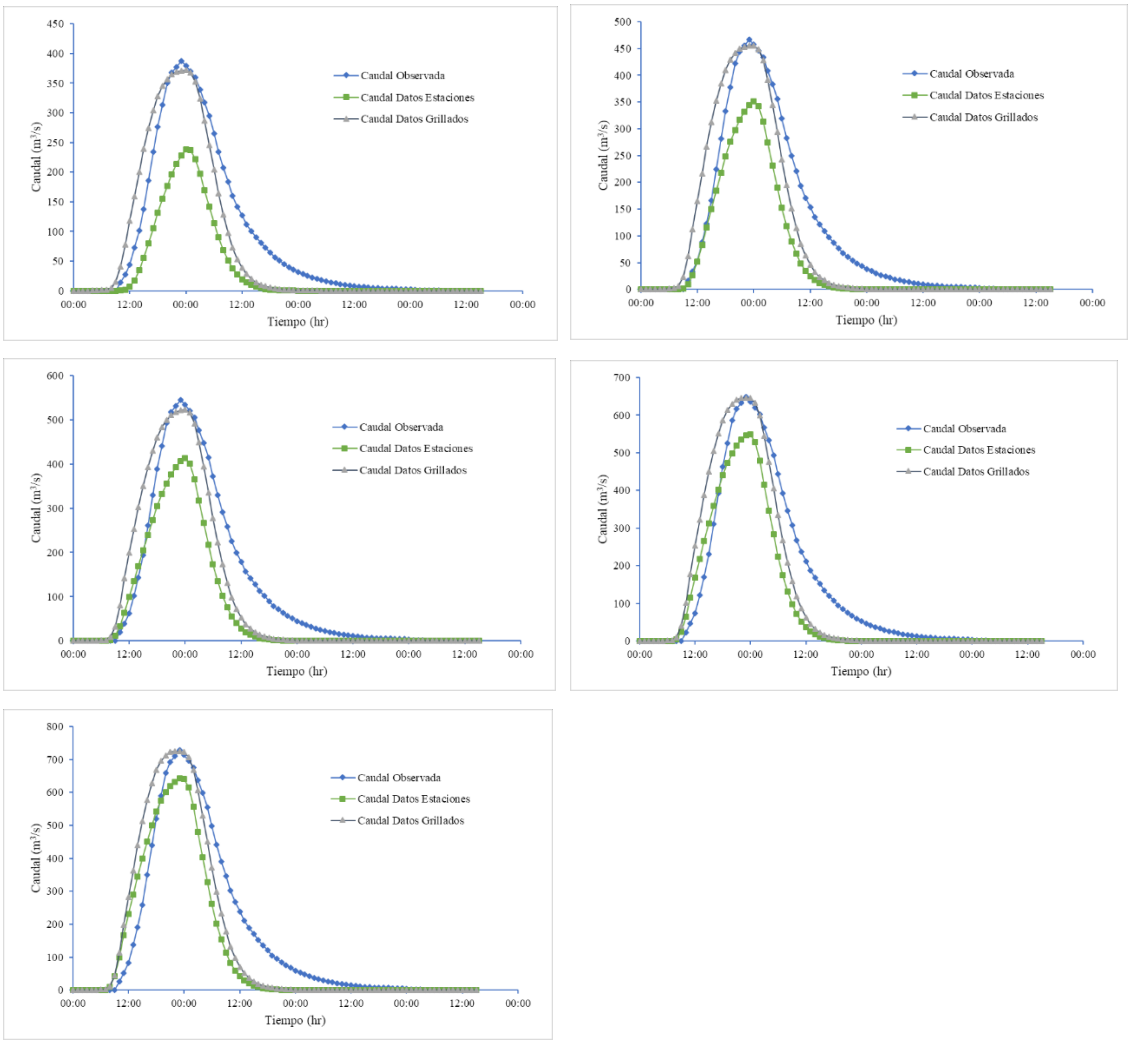


Figura 11: Comparación de hidrograma de caudal observado y simulado durante la calibración a) Tr 5 años, b) Tr 10 años, c) Tr 20 años d) Tr 50 años y e) Tr 100 años

Las métricas de la modelación hidrológica calibrada usando precipitación grilladas (Tabla 15), los valores de R^2 oscilan entre 0.87 a 0.90, clasificándose como satisfactorios. RMSE oscila entre 41.09 m^3/s y 91.01 m^3/s, todos los valores de Tr están por debajo de la mitad de la desviación estándar de los datos simulados. Los valores de NSE oscilan entre 0.84 y 0.88, presentan rendimiento excelente por tener valores por encima de 0.75 para todos los Tr. Los valores de PBIAS oscilan entre 7.71% a 12.61%, de rendimiento excelente (Tr 10, 50 y 100 años) y bueno (Tr 5 y 20 años), estos resultados muestran que el modelo ha subestimado los caudales simulados. PEV varía entre 7.63% a 12.58%, los resultados de Tr 10, 50 y 100 años son clasificados como muy bueno, y el resto como bueno, en general el modelo tiende a subestimar el volumen de caudal. Finamente PEP oscilan entre 0.32% a 3.91%, mostraron rendimiento muy bueno para los cinco periodos de retorno. Los caudales picos simulado están por debajo de caudal pico observado (Figura 11 y 12).

En el desarrollo de este estudio se empleó el hidrograma unitario Clark y Clark modificado, seleccionados principalmente porque poseen características de ser modelo basado en eventos, semidistribuidos (precipitaciones desde estaciones) y distribuido (precipitaciones grilladas) y además poseer parámetros ajustables. Así mismo, el modelo HMS para la calibración empleo el método de búsqueda y aproximación a los valores óptimos: el método Simplex y la función objetivo Peak-Weighted RMSE, obteniendo buenos resultados principalmente para la modelación hidrológica con precipitación grillada.

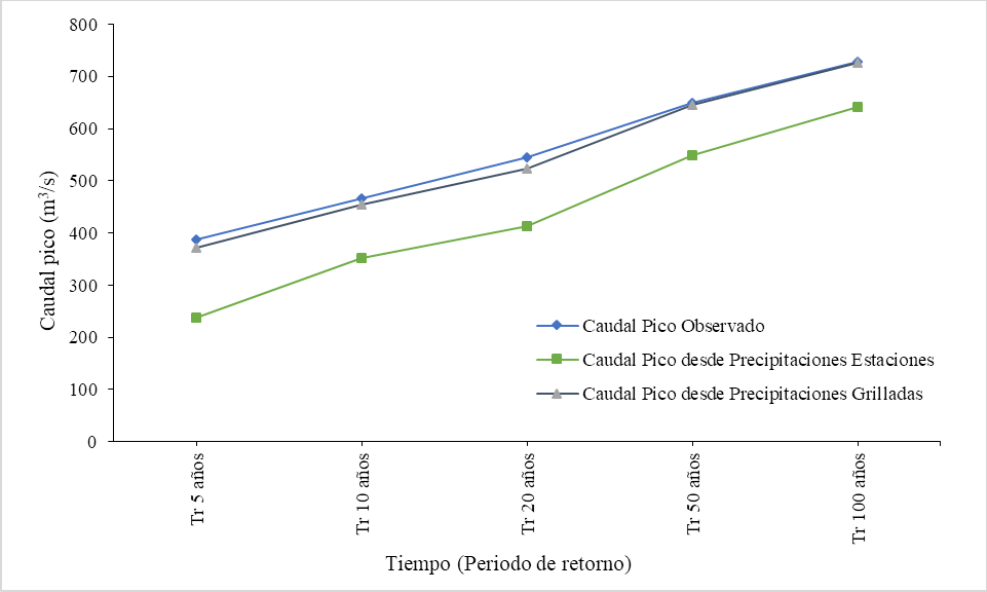


Figura 12: Comparación de los caudales pico observado y simulado durante la calibración

Además, para evaluar la confiabilidad de los resultados en el proceso de calibración, fue muy importante tener en cuenta la cantidad y calidad de los datos, estos afectan directamente en los resultados, dependiendo de las condiciones del modelo se analizó la sensibilidad de los parámetros, que consistió en observar si un cambio en el valor del parámetro afecta con gran importancia el resultado, para este caso se dice que existe sensibilidad en los parámetros, pero si el valor del parámetro no afecta el resultado se dice que no existe sensibilidad. Finalmente, el proceso de calibración del modelo mejoró los resultados con mayor eficiencia en la modelación hidrológica con precipitación grillada en comparación a precipitaciones desde estaciones, como lo muestra los resultados de las métricas estadísticas presentada en las Tabla 15 y las Figuras 11 y 12.

Referencias

Adhikary, S. K., Muttil, N., and Yilmaz, A. G. (2017). Cokriging for enhanced spatial interpolation of rainfall in two australian catchments. Hydrological processes, 31(12):2143–2161.

Alam, M. A., Emura, K., Farnham, C., and Yuan, J. (2018). Best-fit probability distributions and return periods for maximum monthly rainfall in bangladesh. Climate, 6(1):9.

Alamri, N., Afolabi, K., Ewea, H., and Elfeki, A. (2023). Evaluation of the time of concentration models for enhanced peak flood estimation in arid regions. Sustainability, 15(3).

Ali, G., Sajjad, M., Kanwal, S., Xiao, T., Khalid, S., Shoaib, F., and Gul, H. N. (2021). Spatial–temporal characterization of rainfall in pakistan during the past half-century (1961–2020). Scientific reports, 11(1):1–15.

Amini, M. A., Torkan, G., Eslamian, S., Zareian, M. J., and Adamowski, J. F. (2019). Analysis of deterministic and geostatistical interpolation techniques for mapping meteorological variables at large watershed scales. Acta Geophysica, 67(1):191–203.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. Journal of the American Statistical Association, 49(268):765–769.

Ashkar, F. and Ba, I. (2017). Selection between the generalized pareto and kappa distributions in peaks-over-threshold hydrological frequency modelling. Hydrological Sciences Journal, 62(7):1167–1180.

Atwan, T. A. (2022). Time Series Analysis with Python Cookbook: Practical recipes for exploratory data analysis, data preparation, forecasting, and model evaluation. Packt Publishing.

Bárdossy, A., Modiri, E., Anwar, F., and Pegram, G. (2021). Gridded daily precipitation data for iran: A comparison of different methods. Journal of Hydrology: Regional Studies, 38:100958.

Bastia, J., Mishra, B. K., and Kumar, P. (2021). Integrative assessment of stormwater infiltration practices in rapidly urbanizing cities: A case of Lucknow City, India. Hydrology, 8(2).

Bayoud, H. A. (2021). Tests of normality: new test and comparative study. Communications in Statistics - Simulation and Computation, 50(12):4442–4463.

Belayneh, A., Sintayehu, G., Gedam, K., and Muluken, T. (2020). Evaluation of satellite precipitation products using HEC-HMS model. Modeling Earth Systems and Environment, 6(4):2015–2032.

Bhusal, A., Parajuli, U., Regmi, S., and Kalra, A. (2022). Application of machine learning and process-based models for rainfall-runoff simulation in DuPage River Basin, Illinois. Hydrology, 9(7).

Cahyono, C. and Adidarma, W. K. (2019). Influence analysis of peak rate factor in the flood eventsâ€™ calibration process using HEC-HMS. Modeling Earth Systems and Environment, 5(4):1705–1722.

Cheah, R., Billa, L., Chan, A., Teo, F. Y., Pradhan, B., and Alamri, A. M. (2019). Geospatial modelling of watershed peak flood discharge in Selangor, Malaysia. Water, 11(12).

Chen, F., Gao, Y., Wang, Y., Qin, F., and Li, X. (2019). Downscaling satellite-derived daily precipitation products with an integrated framework. International Journal of Climatology, 39(3):1287–1304.

Cho, Y. (2020). Application of NEXRAD radar-based quantitative precipitation estimations for hydrologic simulation using ArcPy and HEC Software. Water, 12(1).

Chow, V. T., Maidment, D. R., and Mays, L. W. Hidrología aplicada. Editorial Mc Graw-Hill.

Chu, H.-J. and Chang, L.-C. (2009). Applying particle swarm optimization to parameter estimation of the Nonlinear Muskingum Model. Journal of Hydrologic Engineering, 14(9):1024–1027.

D'Agostino, R. and Stephens, M. (1986). Goodness-of-Fit Techniques. Marcel Dekker.

Das, S. (2021). Extreme rainfall estimation at ungauged locations: Information that needs to be included in low-lying monsoon climate regions like bangladesh. Journal of Hydrology, 601:126616.

Eryani, G. A. P., Amerta, I. M. S., and Jayantari, M. W. (2021). Model calibration parameter using optimization trial in HEC-HMS for Unda Watershed. IOP Conference Series: Earth

and Environmental Science, 930(1):012040.

Fanta, S. S. and Tadesse, S. T. (2022). Application of HEC-HMS for runoff simulation of Gojeb Watershed, Southwest Ethiopia. Modeling Earth Systems and Environment, 8(4):4687–4705.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The Shuttle Radar Topography Mission. Reviews of Geophysics, 45(2).

Flores Tapia, C. E. and Flores Cevallos, K. L. (2021). Pruebas para comprobar la normalidad de datos en procesos productivos: Anderson-darling, ryan-joiner, shapiro-wilk y kolmogórov-smirnov. Societas, 23(2):83–106.

Foehn, A., García Hernández, J., Schaefli, B., and De Cesare, G. (2018). Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in alpine catchments. Journal of Hydrology, 563:1092–1110.

Hershfield, D. M. (1961). Rainfall frequency atlas of the united states. Technical paper, 40:1–61.

Hershfield, D. M. (1965). Method for estimating probable maximum rainfall. Journal-American Water Works Association, 57(8):965–972.

Hromadka, T. V, I., HROMADKA, T. V, I., and PHILLIPS, M. (2010). Use of Rainfall Statistical Return Periods to Determine Threshold for Mass Wasting Events. Environmental & Engineering Geoscience, 16(4):343–356.

Huang, J., Jing, C., Fu, J., and Huang, Z. (2018). Uncertainty analysis of rainfall spatial interpolation in urban small area. In International Conference on Testbeds and Research Infrastructures, pages 79–95. Springer.

Jabbar, L. A., Khalil, I. A., and Sidek, L. M. (2021). HEC-HMS hydrological modelling for runoff estimation in Cameron Highlands, Malaysia. International Journal of Civil Engineering and Technology, 12(9):40–51.

Javari, M. (2016). Geostatistical and spatial statistical modelling of precipitation variations in

iran. J. Civ. Environ. Eng, 6:1–30.

Kousari, M. R., Malekinezhad, H., Ahani, H., and Asadi Zarch, M. A. (2010). Sensitivity analysis and impact quantification of the main factors affecting peak discharge in the scs curve number method: An analysis of iranian watersheds. Quaternary International, 226(1):66–74. Larger Asian Rivers: Climate Change, River Flow, and Watershed Management.

Kropp, S. (2015). Climate change and risk of flooding in germany. Research Collection, page 155.

Kull, D. W. and Feldman, A. D. (1998). Evolution of clark's unit graph method to spatially distributed runoff. Journal of Hydrologic Engineering, 3(1):9–19.

Kumarasamy, K. and Belmont, P. (2018). Calibration parameter selection and watershed hydrology model evaluation in time and frequency domains. Water, 10(6).

Li, D., Christakos, G., Ding, X., and Wu, J. (2018). Adequacy of TRMM satellite rainfall data in driving the SWAT modeling of tiaoxi catchment (taihu lake basin, china). Journal of Hydrology, 556:1139 – 1152.

Majidi, A. and Shahedi, K. (2012). Simulation of rainfall-runoff process using green-ampt method and HEC-HMS model (case study: Abnama Watershed, Iran). International Journal of Hydraulic Engineering, 1(1):5–9.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3):885–900.

Naghettini, M. (2017). Fundamentals of Statistical Hydrology. Springer International Publishing, 1 edition.

Natarajan, S. and Radhakrishnan, N. (2021). Simulation of rainfall-runoff process for an ungauged catchment using an event-based hydrologic model: A case study of koraiyar basin in Tiruchirappalli city, India. Journal of Earth System Science, 130(1):30.

Nathan, G. (2021). Practical Data Science with Python: Learn tools and techniques from hands-on examples to extract insights from data. Packt Publishing.

Navlani, A., Fandango, A., and Idris, I. (2021). Python Data Analysis, Third Edition. Packt Publishing, 3 edition.

Nharo, T., Makurira, H., and Gumindoga, W. (2019). Mapping floods in the middle zambezi basin using earth observation and hydrological modeling techniques. Physics and Chemistry of the Earth, Parts A/B/C, 114:102787.

Paudel, M., Nelson, E. J., and Scharffenberg, W. (2009). Comparison of Lumped and Quasi-Distributed Clark Runoff Models Using the SCS Curve Number Equation. Journal of Hydrologic Engineering, 14(10):1098–1106.

Pereira, G. W., Valente, D. S. M., Queiroz, D. M. d., Coelho, A. L. d. F., Costa, M. M., and Grift, T. (2022). Smart-map: An open-source qgis plugin for digital mapping using machine learning techniques and ordinary kriging. Agronomy, 12(6).

Pérez, R. E., Cortés-Molina, M., and Navarro-González, F. J. (2021). Analysis of rainfall time series with application to calculation of return periods. Sustainability, 13(14):8051.

Portuguez-Maurtua, M., Arumi, J. L., Lagos, O., Stehr, A., and Montalvo Arquiñigo, N. (2022). Filling gaps in daily precipitation series using regression and machine learning in inter-andean watersheds. Water, 14(11).

Portuguez-Maurtua, M. and Verano Zelada, C. (2016). Generación de mapa de número de curva con Sistema de Información Geográfica. Apuntes de Ciencia & Sociedad, 6(1).

Rau, P., Bourrel, L., Labat, D., Melo, P., Dewitte, B., Frappart, F., Lavado, W., and Felipe, O. (2017). Regionalization of rainfall over the Peruvian Pacific slope and coast. International Journal of Climatology, 37(1):143–158.

Razali, N. M., Wah, Y. B., et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, 2(1):21–33.

Sabol, G. V. (1988). Clark unit hydrograph and R - parameter estimation. Journal of Hydraulic Engineering, 114(1):103–111.

Salhi, H. (2022). Evaluation of the spatial distribution of the annual extreme precipitation using kriging and co-kriging methods in algeria country. In Tiefenbacher, J. P., editor, Climate

Change in Asia and Africa, chapter 4. IntechOpen, Rijeka.

Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., Srinivasan, R., and Hauck, L. M. (2001). Validation of the Swat model on a large rwer basin with point and nonpoint sources 1. JAWRA Journal of the American Water Resources Association, 37(5):1169–1188.

Sen, Z. (2018). Flood modeling, prediction and mitigation. Springer.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4):591–611.

Stephens, M. A. (1974). Edf statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347):730–737.

Tassew, B. G., Belete, M. A., and Miegel, K. (2019). Application of HEC-HMS model for flow simulation in the lake Tana Basin: The Case of Gilgel Abay Catchment, Upper Blue Nile Basin, Ethiopia. Hydrology, 6(1).

Teng, F., Huang, W., and Ginis, I. (2018). Hydrological modeling of storm runoff and snowmelt in Taunton River Basin by applications of HEC-HMS and PRMS models. Natural Hazards, 91(1):179–199.

Van Rossum, G. and Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

WMO (2009). Manual on estimation of probable maximum precipitation (PMP).

Zelelew, D. G. and Melesse, A. M. (2018). Applicability of a spatially semi-distributed hydrological model for watershed scale runoff estimation in Northwest Ethiopia. Water, 10(7).

Zou, W.-Y., Yin, S.-Q., and Wang, W.-T. (2021). Spatial interpolation of the extreme hourly precipitation at different return levels in the haihe river basin. Journal of Hydrology, 598:126273.

CAPITULO V: Conclusiones generales

5.1 Machine learning en el relleno de serie de precipitación diaria faltante

Este estudio ha demostrado las ventajas de rendimiento del uso de los algoritmos de machine learning en relleno de datos de precipitaciones diarias faltantes. Los resultados obtenidos mostraron que los modelos de machine learning presentaron menor variabilidad en los errores de estimación y mejor aproximación a los datos reales, interpretando eficientemente la variabilidad espacio-temporal de la precipitación, tal y como demostraron las métricas estadísticas analizadas. Sin embargo, es importante precisar que es necesario realizar el proceso de regionalización, antes de la aplicación de los algoritmos de machine learning.

5.2 Caracterización de riesgo de inundación mediante parámetros morfométricos

Mediante el análisis de los parámetros morfométricos se evaluó la influencia a la respuesta hidrológica de eventos extremos, zonificando zonas de riesgo de inundaciones repentinas en la cuenca del Rio Cañete. El mapeo de zonas vulnerables a inundaciones fue realizado a nivel de subunidades hidrográficas. Los resultados mostraron que las subunidades hidrográficas SC-3, SC-9 y SC-11, son propensas a inundaciones y a una pérdida de suelo muy altas. En general, la cuenca del rio Cañete presentan riesgo desde moderada a muy altas. Las subunidades localizadas aguas abajo de la cuenca Cañete se caracterizan principalmente por presentar inundaciones fluviales. Estos resultados son de gran utilidad a los responsables de toma (gobierno regional y local) de decisiones en aplicación de técnicas adecuadas de gestión de cuencas hidrográficas en cuanto a medidas de conservación del suelo y del agua, permitiendo salvaguardar y mitigar la degradación de la misma en la zona estudiada.

5.3 Modelación hidrológica en base a precipitación grillada

Mediante la aplicación de método geoestadístico (Kriging Ordinario- KO) se generó el grillado de la precipitación extrema, de acuerdo con la evaluación de las métricas estadísticas; este método mostro eficiente en la interpolación espacial. Se comparó la modelación hidrológica utilizando precipitación de estaciones pluviométricas y precipitación grillada, los resultados iniciales mostraron que existe una cercana diferencia entre los caudales máximos observados y simulados usando precipitación grillada, en comparación a lo simulado desde precipitaciones de estaciones.

En el proceso de calibración del modelo, permitió optimizar los parámetros, encontrándose que los parámetros más sensibles en la modelación son: curva número, tiempo de concentración y el coeficiente de almacenamiento. El proceso de calibración permitió mejorar los resultados con mayor eficiencia en la modelación hidrológica con precipitación grillada y en menor eficiencia en la modelación con precipitaciones desde estaciones.

5.4 Hipótesis de la investigación

Finalmente, la investigación evidenció que con el uso de información grillada (precipitación, topografía y numero de curva) se obtuvieron resultados eficientes en predecir la magnitud de las inundaciones mediante el uso de modelo hidrológico.

5.5 Futuras investigaciones

La metodología y análisis realizada en esta investigación, y conjuntamente con las nuevas tecnologías de detección remota para la estimación de precipitaciones grilladas como la Tropical Rainfall Measuring Mission – TRMM, Precipitation Estimation from Remotely Sensed Information using Artificial Neural Net- works – PERSIANN, Global Precipitation Measurement – GPM, etc, se representan de gran utilidad como datos de entradas a los modelos hidrológicos distribuidos basado físicamente, permitiendo evaluar con precisión el desempeño de respuesta de los procesos hidrológicos a nivel de cuenca.