Revista Cubana de Meteorología

Revista Cubana de Meteorología, Vol. 31, No. 1, enero-marzo 2025, ISSN: 2664-0880

Artículo Original

Control de calidad y reconstrucción de las series de precipitación diaria por estaciones para Cuba desde 1961 hasta 2022

https://cu-id.com/2377/v31n1e14

Quality control and reconstruction of daily precipitation gauge-based time series for Cuba from 1961 to 2022

Abel Centella-Artola[1*], Cecilia Fonseca-Rivera[1], Roberto Serrano-Notivoli[2], Ranses Vázquez-Montenegro[1], Arnoldo Bezanilla-Morlot[1], Maibys Sierra-Lorenzo[1], Dimitris A. Herrera[3,4]

[1]Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba.
[2]Departamento de Geografía y Ordenación del Territorio, Instituto Universitario de Ciencias Ambientales, Universidad de Zaragoza, Zaragoza, 50009, España.
[3]Department of Geography & Sustainability, University of Tennessee-Knoxville, TN, USA.
[4]Instituto Geográfico Universitario, Universidad Autónoma de Santo Domingo, Santo Domingo 10103, Dominican Republic.

RESUMEN: En este trabajo se describe el desarrollo de un conjunto de datos pluviómetros diarios para Cuba desde 1961 hasta 2022, utilizando 2871 estaciones proporcionadas por el Instituto Nacional de Recursos Hidráulicos. A todas las observaciones se les aplicó un exhaustivo control de calidad de coherencia espacial y se realizó un proceso de reconstrucción para completar los valores faltantes en las series temporales de las estaciones. El control de calidad permitió detectar un 0,13% de datos repetidos y alrededor de un promedio anual 14% de valores sospechosos. Las estimaciones utilizadas para la reconstrucción de las series temporales tuvieron una precisión de predicción de días secos y húmedos de 97,4% y 92,5%, respectivamente, mientras que en la estimación comparativa de la magnitud de la precipitación los valores del índice KGE fueron superiores a 0,7 en el 86% de las series analizadas. Para demostrar la utilidad de los datos reconstruidos, se realizó un análisis de tendencias del número de días secos consecutivos y de la proporción de lluvia anual debida a días con lluvias por encima de los indicadores extremos del percentil 95. Los resultados obtenidos sugieren que la tendencia al incremento de ambos índices en la región oriental de Cuba está produciendo un régimen de precipitaciones más extremo en esa zona. El producto de datos de estación generado se considera único y muy valioso para el desarrollo de diversas investigaciones y aplicaciones.

Palabras clave: Control de calidad, precipitación diaria, reconstrucción de series, análisis espacial, índices climáticos extremos.

ABSTRACT: This work describes the development of a daily rain-gauge data set for Cuba from 1961 to 2022, using 2871 stations provided by the National Institute of Hydraulic Resources. An exhaustive quality control of spatial coherence was applied to all observations and a reconstruction process was performed to fill the missing values in the stations time series. The quality control allowed to detect 0.13% of repeated data and about 14% of suspicious values. The estimates used for the time series reconstruction had a prediction accuracy of dry and wet days of 97.4% and 92.5%, respectively, while in the comparive estimation of the amount of precipitation, the values of the KGE index were higher than 0.7 in 86% of the series analyzed. To demonstrate the usefulness of the reconstructed data, a trend analysis of the number of consecutive dry days and the proportion of annual precipitation due to days with rainfall above the 95th percentile extreme indicators was performed. The results obtained suggest that the trend of increasing both indices in the eastern region of Cuba is producing a more extreme rainfall regime in that zone. The station data product generated is unique and critical for the development of various research and applications.

Keywords: Quality control, daily rainfall, time series reconstruction, spatial analysis, climate extreme indexes.

*Autor para correspondencia: *Abel Centella-Artola*. E-mail: abelcentella@gmail.com
Recibido: 12/09/2024
Aceptado: 02/12/2024
Conflicto de intereses: Los autores declaran que no existen conflictos de intereses
Contribución de los autores: Abel Centella Artola: **Conceptualización, Análisis formal, Investigación, Metodología, Software,Redacción- borrador inicial, Redacción - revisión y edición, Visualización, Supervisión.** Roberto Serrano Notivoli: **Metodología, Software, Redacción - revisión y edición.** Cecilia Fonseca Rivera: **Investigación, Redacción - revisión y edición, Validación.** Ransés Vázquez Montenegro: **Redacción - revisión y edición, Validación.** Arnoldo Bezanilla Morlot: **Redacción - revisión y edición.** Maibys Sierra Lorenzo: **Redacción - revisión y edición**

INTRODUCCIÓN

El análisis de los procesos relacionados con la precipitación necesita de conjuntos de datos de gran densidad espacial y alta resolución temporal, los cuales muchas veces no están totalmente disponibles para períodos largos de tiempo (Serrano-Notivoli y Tejedor, 2021). La elevada calidad de esa información es crucial para hacer más preciso el análisis hidrológico y para que sean mejores los resultados de investigaciones orientadas al estudio de los procesos vinculados con el ciclo hidrológico, el desarrollo del pronóstico hidrológico, así como la evaluación de modelos meteorológicos y climáticos (Yilmaz et al., 2005, Liu et al., 2017 y Barnett et al., 2005). Desde el punto de vista práctico, la determinación de los riesgos de inundaciones o sequías requiere de información sobre la lluvia con un nivel de calidad adecuado, con el fin de implementar servicios de información climática fiables para el manejo de los recursos hídricos, la agricultura y otras actividades relacionadas al uso del agua.

A pesar de que los estudios sobre las características espaciales y temporales de las precipitaciones en Cuba son abundantes (Davitaya y Trusov 1965, Trusov 1967, Gagua et al, 1976; Izquierdo, 1989; Trusov et al., 1983, Lecha et al., 1994), las descripciones sobre los métodos de control de calidad (QC) y rellenado de series a la que fueron sometidos los datos son relativamente escasas o no están disponibles. De acuerdo con Eischeid et al. (2000), la realización de controles de calidad y estimación de datos faltantes de forma indocumentada, no coordinada e independiente, puede resultar en procesos redundantes, costosos y hasta incorrectos. Así mismo, el uso de series de datos diarios sin un control de calidad apropiado afecta en mayor o menor medida los valores acumulados que se derivan a partir de ellos (p. ej., 10 días, mensuales, etc).

El volumen de datos diarios de lluvia que hay en Cuba es muy grande, tanto en el período de tiempo que abarcan (más de 60 años), como en su distribución espacial. Sin embargo, al menos en la escala diaria, esa extensa cantidad de información no ha sido sometida a un control de calidad riguroso, ni a un proceso de análisis que posibilite el rellenado de la información faltante. Más aun, en la literatura consultada se encontraron muy pocas referencias de trabajos enfocados directamente a esos temas (González et al., 2022 y Centella-Artola et al., 2023, por citar dos ejemplos), los que por demás sólo analizaron un grupo relativamente pequeño de estaciones pluviométricas.

Con el desarrollo actual de las técnicas de análisis de datos, la mayor capacidad computacional y el crecimiento en las capacidades de almacenamiento de información, las tareas de control de calidad y rellenado de series pueden ser ejecutadas con esfuerzos relativamente menores y empleando técnicas de análisis cada vez más robustas. Como es lógico, la alta variabilidad espacio-temporal de la precipitación hace que esos procesos sean complejos y que, independientemente de la metodología que se utilice para el QC, los conjuntos de datos no quedarán completamente libres de errores, mucho más si la información se relaciona con la escala de tiempo diaria.

González et al. (2022) aplicaron un método de control de calidad temporal y de consistencia interna a las series de datos diarios de 630 estaciones para el período 1961-2008 en todo el país. Ese estudio utilizó técnicas semiautomáticas para detectar los valores sospechosos en las series individuales, sin tomar en consideración las relaciones espaciales entre las mismas. Posteriormente, Centella-Artola et al. (2023) utilizaron los mismos datos y realizaron un QC de características espaciales empleando la metodología desarrollada por Serrano-Notivoli et al (2017a). Sin embargo, como el objetivo final de ese trabajo fue crear un conjunto de datos de lluvia diaria en forma de rejilla regular de alta resolución espacial (datos publicados en: https://doi.org/10.5281/zenodo.7847844), la base de datos resultante del QC no fue documentada detalladamente, ni se hizo disponible públicamente.

El método de Serrano-Notivoli et al (2017a) posibilita realizar el QC, pero también permite la reconstrucción de las series que poseen datos faltantes. La metodología se puede aplicar de forma completamente automatizada, dado que se encuentra implementada en el paquete de R *reddPrec* ver 2.0.3 (Serrano-Notivoli y Centella-Artola, 2024). Por ello es relativamente más fácil realizar el análisis de grandes volúmenes de datos diarios, luego de que los mismos sean explorados y organizados apropiadamente.

Tomando en consideración esos antecedentes, en este trabajo se aplica una metodología que combina el QC espacial y el rellenado de valores ausentes, con el objetivo de crear un conjunto de series de precipitación diaria para el período 1961-2022, con un control de calidad riguroso y sin información faltante. De acuerdo a la bibliografía consultada, es posible que esta sea la primera ocasión en que se obtenga un conjunto de datos diarios de precipitación para Cuba con estas características, y que al mismo tiempo se acompañe con la descripción detallada de la metodología utilizada, el proceso de análisis realizado y la discusión crítica de sus rasgos más relevantes. Se espera que este trabajo sea una contribución importante para el desarrollo del nuevo mapa isoyético nacional y que la disponibilidad de estos datos facilite las investigaciones sobre el cambio climático, la variabilidad del clima, y el estudio de los procesos hidrológicos en el país. Para demostrar esas posibles ventajas, en esta investigación también se incluye el análisis del comportamiento observado en eventos climáticos extremos de lluvia en Cuba.

A continuación se describen los métodos utilizados para organizar, controlar y reconstruir las series de datos; seguidamente hay una sección donde se presenta los resultados del QC, así como los relativos a la precisión con que se rellenaron las series. En la sección cuarta se discuten los resultados obtenidos, para finalizar con las conclusiones en la sección 5.

MATERIALES Y MÉTODOS

El proceso de creación del conjunto de datos de las series de precipitación diaria por estaciones se realizó en tres etapas principales (Figura 1). Inicialmente se compiló y limpió toda la información a la que se tuvo acceso, luego se realizó el QC de la misma y finalmente se procedió a realizar el rellenado de los datos faltantes de las series con un período de tiempo aceptable para realizar este tipo de estimación. A continuación, se detallan cada una de esas etapas.

Preparación de datos

Los datos diarios utilizados se obtuvieron a partir de la información almacenada por el INRH y en mucha menor medida por el INSMET en diferentes períodos y con formatos diversos. La primera fuente de estas informaciones (INRH-1995) se relaciona con los registros de 2194 estaciones que abarcan desde 1892 hasta 1995 (Planos, comunicación personal, 2023). La segunda fuente son datos de 630 estaciones que abarcan el período 1961-2008, y son los mismos que fueron utilizados en la creación del conjunto de datos CubaPrec1 (Centella-Artola et al., 2023). Por último, la tercera fuente (INRH-2023) corresponde con la información disponible en los archivos de cada una de las provincias, la cual estaba en una gran diversidad de formatos y contenía datos desde diferentes años de inicio hasta el 2022.

La identificación de las estaciones comunes entre las distintas fuentes de datos no resultó directa, pues los identificadores de estaciones utilizados en una eran diferentes a los de las otras. Los estándares utilizados por el INRH para identificar las estaciones combinan el código de la provincia (idprov) y el número de control o identificador de cada pluviómetro (ncontrol) dentro de cada provincia, que puede repetirse entre ellas. Esta práctica generó un problema, pues cuando la Distribución Política Administrativa (DPA) de Cuba cambió y se modificaron la cantidad de provincias y los límites fronterizos entre ellas, los identificadores en los grupos de datos INHR-1995 y Superclima eran distintos a los de INRH-2023. Por otro lado, también existían diferencias en el sistema de referencia utilizado en cada grupo para las coordenadas de las estaciones, pues en unos casos se encontraban en coordenadas planas y en otros en coordenadas geográficas.

Para solucionar estas limitaciones, se tomó el listado de estaciones pluviométricas utilizado en la confección del mapa isoyético nacional del período 1961-2000. En ese listado, los identificadores de las estaciones se forman con los 16 idprov la DPA actual, pero al proyectar sus coordenadas sobre los polígonos de la DPA anterior (fichero *shape* Old DPA), se pudo extraer el código idprov anterior y entonces se construyó otro indicador, asumiendo que los números de control no cambiaron. Finalmente, el nuevo listado con los dos indicadores se proyectó sobre los polígonos de la nueva DPA (fichero *shape* New DPA), con el fin de realizar un doble chequeo.

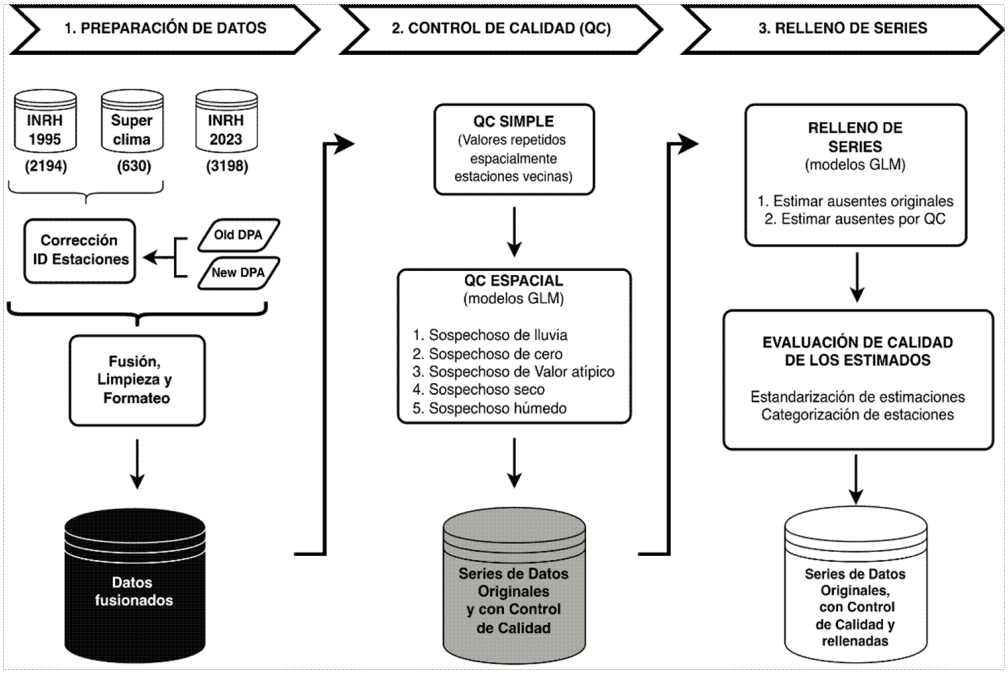

Figura 1. Diagrama de flujo de las principales etapas de trabajo. Los números entre paréntesis denotan la cantidad de estaciones presentes en los subconjuntos de datos originales, denominados INRH-1995, CubaPrec1 e INRH-2023.

Luego de esto, al disponer de un listado de estaciones con los dos identificadores de estaciones, se logró vincular las series de datos existentes en cada grupo y fusionar toda la información de forma coherente.

En la fusión y limpieza de los tres grupos de datos se siguieron las siguientes reglas:

1. Descartar las estaciones que sólo contenían información antes de 1961.
2. Eliminar los valores físicamente imposibles (fechas con formatos o valores imposibles, o identificadores de estaciones incorrectos debido a códigos de provincia inexistentes).
3. Quitar los registros de datos cuando la información de una estación existió en más de uno de los grupos, pero los valores de lluvia eran distintos. En estos casos no había criterios para identificar los valores correctos.
4. Cuando existió información de una estación en más de un grupo y los valores fueron similares, se tomaron los del grupo en que aparecían con cifras decimales.
5. Toda la información de lluvia que no fueron valores numéricos iguales o mayores que cero, fueron codificados como faltantes.

Control de Calidad (QC)

El QC se le aplicó a los datos de las estaciones que se obtuvieron a partir de la preparación de datos (etapa anterior). El proceso se inició con la identificación de datos repetidos (mayores que cero) en estaciones vecinas. Inicialmente, la regla que se siguió fue marcar y revisar los casos que, dentro de un grupo de 10 estaciones vecinas, 5 o más tuviesen exactamente igual valor de lluvia en el mismo día. Como resultado de este análisis se pudo determinar que la ocurrencia de valores duplicados se producía entre las estaciones de una misma provincia y no en las estaciones de provincias diferentes. Esto sugirió que, en el proceso de recopilación de los datos de cada provincia, existieron casos en que se asignó el mismo valor de lluvia a todas las estaciones. La decisión final que se adoptó fue eliminar todos los datos dentro de cada provincia cuya magnitud (superior a cero) fuese igual en 5 o más estaciones. El límite de 5 estaciones fue determinado por los autores a partir del análisis de la frecuencia en que esto sucedía anualmente. Debe notarse que en este chequeo no fue posible identificar los posibles casos donde el valor duplicado fue igual a cero.

Luego de esas verificaciones se procedió a realizar el QC utilizando la metodología propuesta por Serrano-Notivoli et al (2017a) y aplicada previamente en Cuba por Centella-Artola et al (2023). Ese método permite realizar el QC a todas las observaciones y también posibilita realizar el rellenado de las series, como se explicará posteriormente. Ambos procesos se basan en la estimación de valores de referencia (RV) utilizando los datos de 10 estaciones más cercanas (NNS). Los RV se determinan, para cada localidad y día, a partir de dos estimados que se obtienen mediante modelos lineales generalizados (GLM, por sus siglas en inglés). En un caso se hace una predicción binomial (BP) de la probabilidad de ocurrencia de un día húmedo, mientras que en el otro se realiza una estimación de la magnitud de la precipitación (MP).

En el caso de BP, se ajusta un GLM con la familia binomial, para lo cual los valores de las NNS se codifican como 1 o 0, en dependencia de si llovió o no, respectivamente. Para MP también se utilizan GLM pero con la famila cuasi-binomial. Por esta razón, los datos son transformados utilizando la ecuación (1).

$$\mathrm{Pcp_{i,l}} = \frac{\mathrm{NN}_{l,i} - (\min_i - (Q2_i - Q1_i))}{(\max_i + (Q3_i - Q1_i)) - (\min_i - (Q2_i - Q1_i))} \quad (1)$$

Donde: $Pcp_{i,l}$ son los valores transformados para el día i y la localidad l ; $NN_{i,l}$ son los valores de lluvia de las 10 NNS; *min, max, Q1,Q2 y Q3* son el mínimo, máximo y los cuartiles 1, 2 y 3, respectivamente, de los valores de las 10 NNS en el día i.

El valor final de RV es determinado por la combinación de MP y BP, considerando un umbral de BP mayor o igual que 0.5 para determinar un día húmedo. Así, si BP > 0.5, RV es igual a MP, de lo contrario RV es cero.

En el caso del QC, no se emplean directamente los valores de RV, pues resulta suficiente utilizar las magnitudes de BP y MP para aplicar secuencialmente los 5 criterios siguientes. Nótese que ellos se aplican para cada estación y día de manera independiente, permitiendo identificar y eliminar automáticamente los valores que resultan sospechosos.

1. **Sospechoso individual de lluvia (QC1)**: el valor observado es superior a cero y todos sus 10 NNS son ceros.
2. **Sospechoso individual de cero (QC2):** el valor observado es cero y todos sus 10 NNS están por encima de cero.
3. **Sospecha de valor atípico (QC3):** La magnitud del valor observado es 10 veces mayor o menor que la estimada (MP).
4. **Sospechoso seco (QC4):** el valor observado es cero, la probabilidad de día húmedo (BP) es superior a 0,99 y la magnitud estimada (MP) es superior a 5 mm.
5. **Sospechoso húmedo (QC5):** el valor observado es superior a 5 mm, la probabilidad (BP) de sequedad es superior a 0,99 y la magnitud estimada (MP) es inferior a 0,1 mm.

En este estudio, los GLM que se ajustaron, utilizaron como covariables las coordenadas y la distancia de costo[1]. Esta decisión se basó en los resultados publicados por Centella-Artola et al (2024) al evaluar el efecto de distintas variables topográficas en la estimación espacial de la lluvia, a la alta correlación de esa covariable con otras como la altura, y a la mayor contribución de esta en la precisión de los estimados. Los valores de las tres covariables (longitud, latitud y distancia de costo) fueron obtenidos a partir del modelo digital del terreno utilizado en el Atlas Nacional de Cuba (CITMA, 2019) y son las mismas que se utilizan en las estimaciones de los RV durante la etapa posterior de rellenado de las series.

Relleno de Series

Para el rellenado de series se empleó el mismo método de estimación con GLM, pero utilizando las series resultantes del QC, en las que los valores sospechosos fueron suprimidos (reemplazados por código faltante). Aquí sí se utilizan directamente los RV estimados para reemplazar los valores que originalmente estaban ausentes o aquellos que fueron eliminados durante el QC.

Los RV pueden ser ajustados o estandarizados, multiplicándolos por un factor de corrección que permite preservar las particularidades y la estructura de las series originales (Serrano-Notivoli et al., 2017a). El factor de corrección se calcula como la razón de las sumas respectivas de todas las observaciones y de todos los estimados, para los casos en que uno u otro valor sean distintos de cero. Las sumas de las observaciones y los estimados se calculan utilizando una ventana de *n* días, centrada en el día a estandarizar. Por ejemplo, para estandarizar el estimado de una estación para el día *x* con una ventana de tres días, se suman los valores de las respectivas series temporales de las observaciones y las estimaciones, correspondientes a los días *x*-1, *x,* y *x*+1. Con el cociente de estos totales se obtiene el factor de corrección que sirve para estandarizar el estimado de lluvia del día *x*.

Este procedimiento es diferente al utilizado anteriormente por Centella-Artola et al. (2023), quienes aplicaron la estandarización considerando los meses calendarios (cada día se estandarizaba con los datos del mes al que pertenecía). Ahora, al utilizar una ventana de *n* días anteriores y posteriores, se evita que se tomen en cuenta días que pueden tener poca o ninguna relación con el día a estandarizar. Es importante señalar que el proceso de estandarización puede producir valores irreales con magnitudes extremadamente elevadas, pues en ocasiones la suma de los valores observados es sustancialmente más alta que la de los estimados. Para evitar la presencia de esas magnitudes irreales, se verificó que los estimados estandarizados no fueran superiores al valor máximo absoluto de las observaciones, y se utilizó el estimado sin estandarizar.

Los valores que se predicen para cada estación se obtienen sin la participación del dato observado en la misma. De esa forma, siempre se tienen parejas de valores estimados y observados con los que se aplicó el método de validación cruzada dejando una estación fuera (LOSO-CV, por el inglés leave-one-station-out cross validation), que sirve para evaluar la habilidad del método en estimar los datos ausentes en cada una de las estaciones.

Las comparaciones se hicieron entre las observaciones y las estimaciones sin estandarizar, así como con estimaciones estandarizadas utilizando ventanas con diferentes intervalos de tiempo desde 1 hasta 15 días. En este proceso se utilizaron diagramas de dispersión para evaluar la habilidad general del método de estimación, así como tablas de contingencia para verificar la precisión de la predicción de los días con y sin lluvia (mayores e iguales que cero, respectivamente). También se calculó el índice de eficiencia modificado de Kling-Gupta (KGE) (Kling et al., 2012), que es un indicador adimensional que se emplea para realizar evaluaciones de modelos o hacer comparaciones de bases de datos de variables relacionadas con el ciclo hidrológico.

La formulación del KGE es como sigue:

$$\mathrm{KGE}=1-\sqrt{(r-1)^2+(b-1)^2+(v-1)^2} \quad (2)$$

$$\mathrm{r}=\frac{\sum_i^n (O_i-\bar{O})(S_i-\bar{S})}{\sqrt{\sum_i^n (O_i-\bar{O})^2}\sqrt{\sum_i^n (S_i-\bar{S})^2}} \quad (3)$$

$$\mathrm{b}=\frac{\bar{S}}{\bar{O}} \quad (4)$$

$$\mathrm{v}=\frac{\mathrm{CV}_S}{\mathrm{CV}_O} \quad (5)$$

Donde: n es el número de observaciones, O_i y S_i son los datos observados y estimados, respectivamente en la observación i; $\bar{O}$, $\bar{S}$, CV_o y CV_s son los valores medios y del coeficiente de variación de las observaciones y los estimados, respectivamente.

[1] Distancia de cada punto a la costa, pero asumiendo el costo de atravesar el relieve. En este caso, el costo está dado por una superficie de fricción (estimada a partir de la topografía), la cual se multiplica por la distancia para obtener el valor final. La definición es bastante común en los SIG, aunque el costo puede ser de otra naturaleza.

Por su propia formulación se evidencia que el valor óptimo del índice KGE y de sus distintas componentes (r, b y v) es la unidad. En esencia, r mide la correlación lineal entre las observaciones y los estimados, b refleja la tendencia de los valores estimados a ser mayores o menores que las observaciones, mientras que v indica si las observaciones tienen mayor o menor variabilidad que los estimados.

Indicadores extremos de lluvia

Para demostrar el uso potencial de la información resultante del proceso de reconstrucción (QC y rellenado de series), se procedió a realizar el análisis de las tendencias observadas en dos indicadores climáticos extremos, similares a los definidos por el Grupo de Expertos en Detección e Índices de Cambio Climático (ETCCDI, por sus siglas en inglés) (Klein Tank et al., 2009, Zang et al., 2011). Los indicadores son los siguientes:

- CDD: Cantidad de días consecutivos en los que la precipitación es menor que 1 mm.
- r95ptot: Porcentaje del acumulado anual de lluvia correspondiente a la suma de las precipitaciones diarias mayores que el percentíl 95 (95p). Los valores de 95p se obtuvieron a partir de la distribución percentílica del período de referencia 1971-2000.

Los índices fueron determinados anualmente y a diferencia del método de cálculo propuesto por el ETCCDI, la contabilidad de CDD se realizó dentro del mismo año natural sin permitir el solapamiento entre años colindantes. Los índices fueron calculados utilizando el software Climate Data Operators (CDO, por sus siglas en inglés) con los datos de estaciones seleccionadas. Posteriormente, se estimó la significación estadística de la tendencia de los indicadores utilizando la prueba no paramétrica de Mann-Kendall (Mann, 1945 y Kendall, 1970), mientras que la magnitud de la misma se estimó con el estimador de la pendiente de Sen, que es una prueba no paramétrica basada en el estadígrafo Tau de Kendall (Sen, 1968).

RESULTADOS

Limpieza y fusión de los datos originales

El proceso de limpieza y fusión de toda la información disponible generó un conjunto de datos con las series diarias de precipitación para las 2871 estaciones pluviométricas utilizadas. Como se puede observar en la Figura 2a, la densidad de estaciones es muy alta (1 estación por cada 36 km²) y, aunque no todas poseen series temporales con la misma duración, más del 60% de las estaciones tienen registros superiores a 30 años, mientras que cerca de un 10% posee muestras inferiores a 20 años. También se aprecia que el mayor número de estaciones con series temporales de mayor duración se concentra en algunas provincias (véase la Figura 2b como referencia de las provincias y el relieve).

En general, el número de estaciones aumentó abruptamente entre los años 1961 y 1966, manteniendo un ligero incremento hasta principios de la década de los 90s del siglo XX (Figura 2c). A partir de 1991 al presente se produce una reducción de la cantidad de estaciones, lo que parece ser un período de reorganización de las mediciones, donde finalizaron e iniciaron series distintas. No es posible indicar si el proceso estuvo relacionado con cambios en la ubicación de las estaciones, pues los metadatos necesarios no están disponibles para realizar ese análisis.

En correspondencia con las características del relieve cubano, la mayor cantidad de estaciones se localiza en alturas por debajo de los 100 m.s.n.m y entre 100 y 200 m.s.n.m (Figura 2d). Sobre los 700 m.s.n.m, afortunadamente existen 40 estaciones.

Como es lógico, las variaciones mostradas en la Figura 2c implican cambios en la densidad de las estaciones, y a pesar de que puede considerarse alta a lo largo del tiempo, existen áreas del país en las que varió mucho más que en otras o donde esta característica es menos favorable. Por ejemplo, en las provincias de Mayabeque y Holguín la densidad oscila en rangos de 1 estación por 25 - 30 km² y por 30-40 km², respectivamente (Figura 3). Sin embargo, en otros territorios como Matanzas, Camagüey o la Isla de la Juventud, la situación es menos favorable pues la densidad es menor (1 estación por 50 - 100 km²). La mayor densidad ocurre en La Habana, donde hay una gran concentración de 1 estación por ~14 km², triplicando la densidad de cualquier otra provincia del país.

Control de calidad espacial

La presencia de datos repetidos fue sólo el 0.13 % respecto al total de observaciones y su variación no fue uniforme a lo largo del tiempo (no se muestra). La mayor frecuencia de valores repetidos se concentra entre 1961-1963 y 1995-2008, a lo cual se adiciona el año 2021. Dado que los datos repetidos fueron eliminados, la cantidad de valores faltantes se incrementó ligeramente (originalmente los datos faltantes representaban el 35.4%). Entonces el control de calidad se aplicó al 64.5% de datos restante.

El porcentaje total de datos sospechosos detectados respecto a la cantidad de observaciones válidas por año (Figura 4), varió de 10.6% - 18.1%. El tipo más frecuente fue el QC1, seguido por el QC3, mientras que el menos frecuente fue el QC5. La mayor frecuencia de QC1 es comparable con los

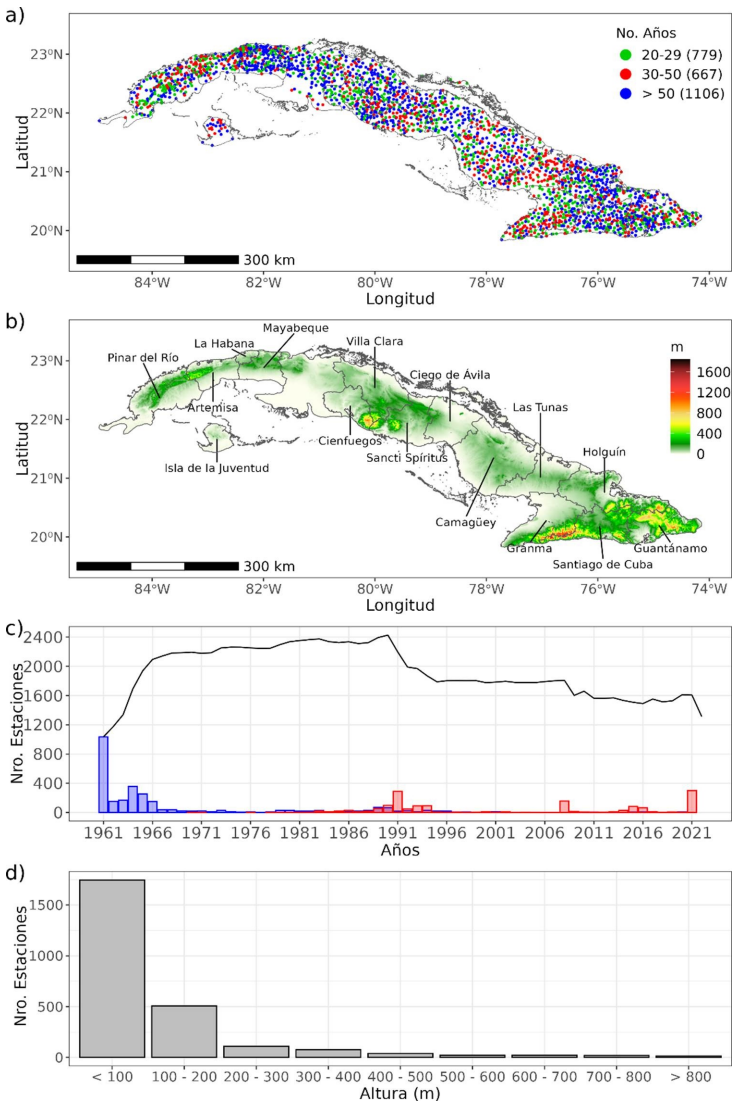

Figura 2. Distribución espacial de las estaciones y variación temporal de su cantidad. a) Localización de las estaciones por rangos de duración de las series. Los números entre paréntesis destacan la cantidad de estaciones. No se muestran las estaciones con menos de 20 años. b) Provincias y topografía de Cuba. c) Número de estaciones por año en el período 1961-2022 (línea negra). La cantidad de estaciones con registros que se inician o finalizan en años determinados se muestran en las barras azules y rojas, respectivamente. d) Histograma con la distribución de la cantidad de estaciones por rangos de altura.

resultados de Centella-Artola et al (2023), aunque el caso de QC3 refleja una posible contribución favorable del reescalado de los valores de lluvia para estimar la magnitud de la precipitación (Ecuación 1), en comparación con la formulación utilizada a tales efectos en trabajos anteriores. Ahora, prácticamente se anula la posibilidad de extrapolar estimaciones fuera de un rango considerablemente mayor que el de las observaciones. Esto hace que, si en un entorno de 10 NNS existe un valor que sobresale notablemente respecto a los demás, queda eliminado (marcándolo como QC3). No obstante, éste debe ser de al menos un orden de magnitud superior.

Al finalizar el QC, todos los datos sospechosos fueron removidos, lo que incrementó el volumen de datos faltantes a un 43.7 %. Estos datos faltantes fueron estimados con el proceso de rellenado de las series.

Rellenado de las series

La comparación de los valores observados (OBS) y estimados (PRED) brinda una idea objetiva sobre la habilidad del método utilizado para reconstruir las series de datos. Aquí, las observaciones fueron contrastadas con los estimados de lluvia sin estandarizar, así como con sus valores estandarizados, utilizando distintas ventanas de *n* días, con *n* variando desde 1 hasta 15. La evaluación se realizó en dos direcciones: i) estimación de la magnitud de la precipitación y ii) estimación de la ocurrencia de condiciones secas y húmedas representadas por valores iguales y mayores que cero, respectivamente.

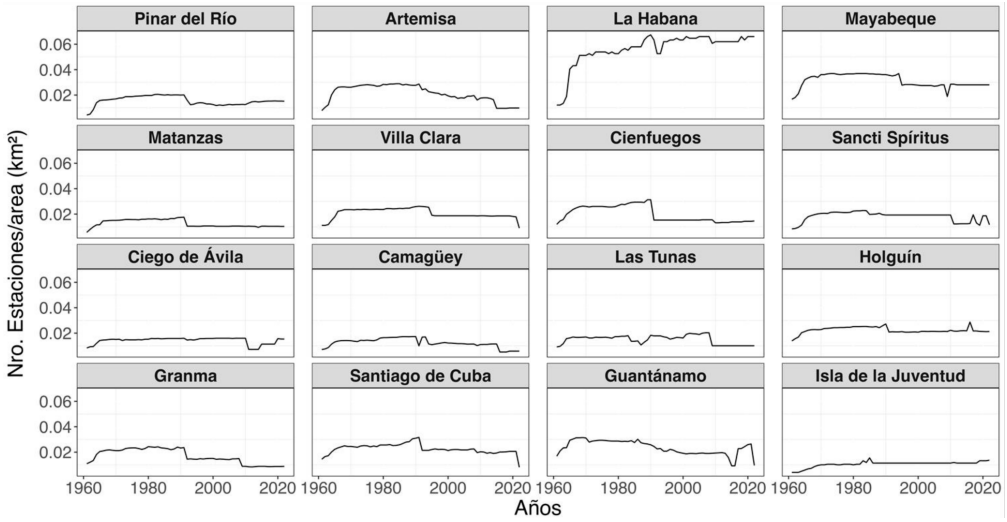

Figura 3. Variación de la densidad de estaciones con observaciones por provincias en el período 1961-2022. El área de las provincias se determinó sin tomar en cuenta los cayos asociados, por lo tanto la misma es menor que la real.

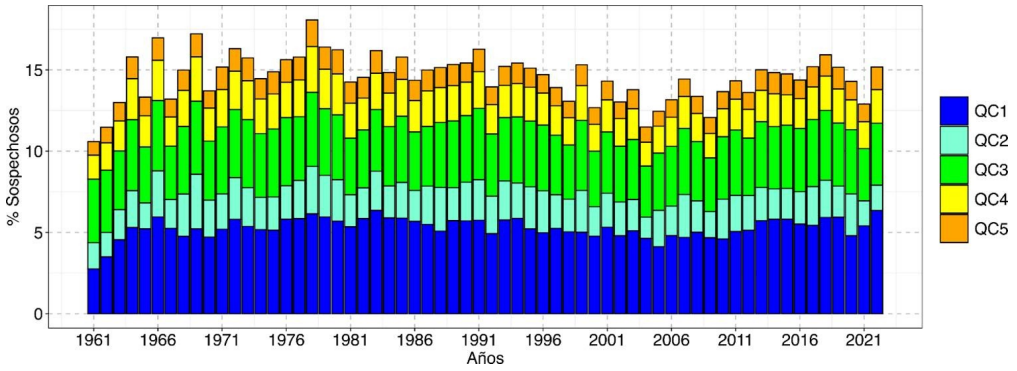

Figura 4. Porcentaje de valores sospechosos por tipo de QC respecto al total de valores válidos (no faltantes) por año.

Evaluación de la estimación de la magnitud de la precipitación

Al comparar las OBS con los PRED sin estandarizar y estandarizados, utilizando las diferentes ventanas de n días, se comprobó que los mejores resultados se produjeron para las magnitudes estandarizadas con una ventana de 1 día. En este caso, los valores del KGE y sus componentes r,b y v, tomando en cuenta todos los valores (0.79, 0.79, 0.99, 1.0, respectivamente), indican que la habilidad general del método de estimación es buena, mostrando que los estimados tienden a ser muy similares a los observados, sucediendo lo mismo con su variabilidad. Si los valores se agrupan por meses (Tabla 1) los resultados son similares. En este caso se destaca que el único componente del KGE que refleja estacionalidad es r, con una ligera reducción hacia los meses del período lluvioso. En general los valores del KGE aparecen dominados en mayor grado por las magnitudes del coeficiente de correlación de Pearson.

Alrededor del 35 % de las estaciones posee valores de KGE entre 0.8 y 0.89, mientras que un 51 % tiene magnitudes en el rango de 0.7 y 0.79. Aunque ambos grupos de estaciones están presentes en todo el país (Figura 5), las primeras se concentran más en la mitad occidental y las segundas en la oriental. Esta distribución parece estar relacionada con la estructura espacial de las estaciones mostrada en la Figura 2a, pues en las zonas donde hay mayor densidad de estaciones con series más largas, los resultados del KGE tienden a ser mejores.

Tabla 1. Valores de *r, b, v* y KGE como medida de la bondad de ajuste de las observaciones y los estimados de lluvia diaria agrupados por meses. *r, b, v* y KGE están definidos en las ecuaciones 2-5.

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
r	0.81	0.82	0.80	0.79	0.78	0.79	0.75	0.74	0.77	0.79	0.81	0.80
b	1.00	1.00	0.98	0.98	1.03	1.02	0.94	0.96	1.00	1.02	1.02	0.99
v	1.01	1.02	1.01	1.01	0.98	0.99	1.02	1.01	1.00	1.00	1.00	1.01
KGE	0.81	0.82	0.80	0.79	0.78	0.79	0.74	0.74	0.77	0.79	0.81	0.80

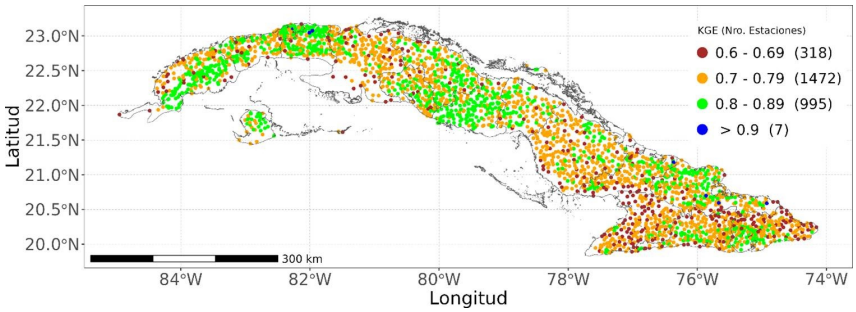

Figura 5. Distribución espacial de las magnitudes del KGE por estaciones y rangos de valores. Los números entre paréntesis reflejan la cantidad de estaciones en cada rango. Sólo se muestran las estaciones con valores de KGE mayores que 0.6.

En la comparación por días (todos los valores de las estaciones para cada día), los valores del coeficiente de Pearson (R) siempre fueron superiores a 0.90 (Figura 6a, b y c), con el menor valor (0.9005) para la mediana de los días húmedos (lluvia > 0 mm). Los resultados también fueron buenos en la comparación por estaciones (todos los valores para cada estación, Figura 7d, e y f), aunque en el caso de los valores extremos (p95) las magnitudes de R fueron menores.

Un análisis similar al anterior se realizó agrupando las estaciones en dos subconjuntos, uno con las estaciones con las series de 30 o más años (1590 estaciones) y otro con series igualmente de al menos 30 años, pero terminando en 2022 (1089 estaciones). Con este análisis se comprobó que con los estimados estandarizados de uno u otro subconjunto, los resultados que aparecen en la Figura 6d, e y f, son mejores y en ambos casos los valores de R son cercanos a 1, 0.91 y 0.88. Esto confirma que la estandarización es más robusta cuando la serie de datos observados es más prolongada (Serrano-Notivoli et al 2017a). En lo sucesivo al primer subconjunto se le denominará "*climático*" y al segundo "*operacional*".

La subdivisión anterior también permitió comprobar dos aspectos: i) que el subconjunto *climático* es el que engloba la mayor cantidad de estaciones con los mejores resultados de la estimación. Por esta razón puede ser tomado como referencia al momento de desarrollar futuras investigaciones relacionadas con la variabilidad, el cambio climático y otras y ii) que el subconjunto *operacional* es apropiado para el desarrollo de servicios climáticos operativos, toda vez que incluye las estaciones con series largas y que continúan reportando información diariamente, por lo que sus datos pueden ser actualizados. Ese subconjunto también puede ser el de mayor utilidad para apoyar la corrección de sesgos de los productos satelitales de precipitación. En ambos casos la distribución espacial es bastante buena, aunque que la variación temporal del número de registros por año es mucho menor en el conjunto *operacional* que en el *climático* (Figura 7).

A pesar de que durante el período 1961-1964 la cantidad de estaciones con datos válidos es mucho más baja y que la menor densidad de estaciones incrementó la distancia de las que participan en la estimación, los resultados del proceso de estimación fueron apropiados. En este caso, al realizar la

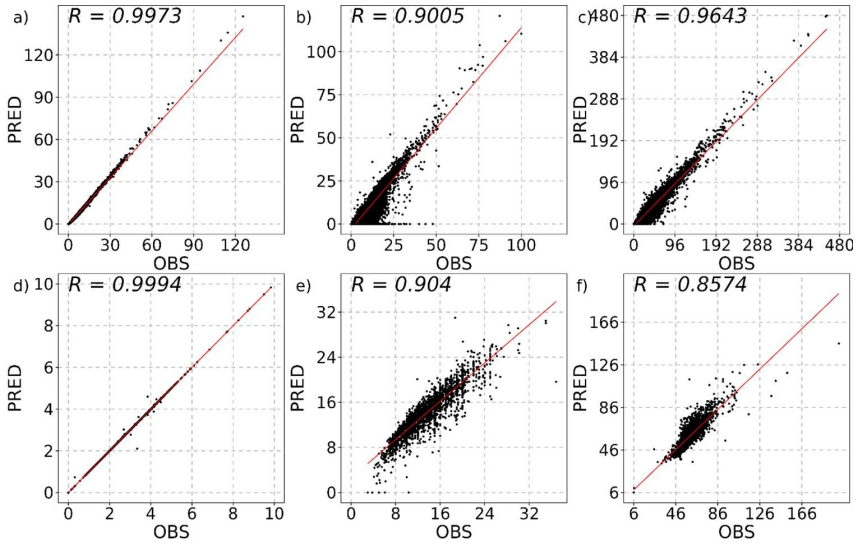

Figura 6. Comparación entre los estimados y las observaciones por días (a,b,c) y pluviómetros (d,e,f). (a y d) media de la precipitación; en (b y e) mediana de los días con lluvia y; en (c y f) 95 percentil de los días con lluvia.

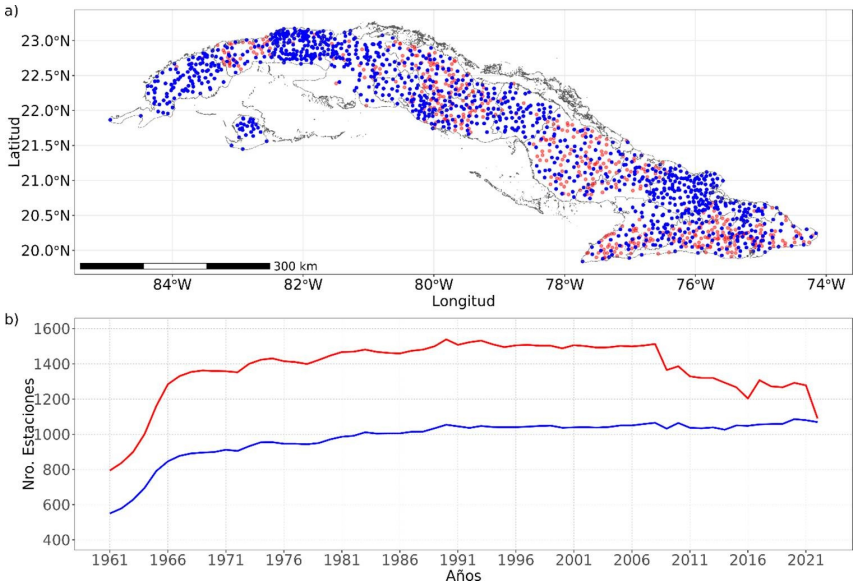

Figura 7. Distribución espacial y variación temporal de las estaciones en los subconjuntos *climático* y *operacional* (puntos y líneas roja y azul, respectivamente). Nótese que el conjunto *operacional* está contenido en el *climático*

comparación por estaciones, los valores de R para la media, la mediana y el percentil 95 fueron 0.8777, 0.7139 y 0.7270, respectivamente; mientras que en la comparación por días todos fueron superiores a 0.94.

Precisión en la estimación de condiciones secas y húmedas

Lo primero que destaca es que el método no está sesgado en la estimación de los días secos (valores iguales a cero) y húmedos (valores mayores que cero), si se toma en cuenta que la cantidad total de ceros en los datos originales es de 32 307 384 y que el número estimado fue de 32 546 432 (una relación de 0.99265). En general, la precisión de la estimación de verdaderos ceros (VC: OBS = 0 y PRED = 0) y verdaderos positivos (VP: OBS > 0 y PRED > 0) es muy elevada, mostrando un 97.4% y 92.5%, respectivamente (Tabla 2). Por otro lado, las razones de falsos ceros (FC: OBS > 0 y PRED = 0) y de falsos positivos (FP: OBS = 0 y PRED > 0) son bajas (2.6% y 7.5%, respectivamente). La precisión en la estimación de las condiciones secas es de 2 a 3% inferior en los meses más lluviosos con respecto a los que menos llueve, sucediendo lo contrario en la predicción de las condiciones húmedas.

Tendencias de los indicadores extremos

La Figura 8 muestra los resultados para los indicadores extremos CDD y r95ptot, utilizando las series de las 1590 estaciones del conjunto climático. Como se mencionó antes, en ese grupo de estaciones se incluyen todos pluviómetros con más de 30 años de datos luego del QC. Sin embargo, debe mencionarse que al emplear las 2871 estaciones reconstruidas, los resultados son muy similares.

En el caso de CDD las estaciones con tendencias (Z) positivas y negativas representan el 20% y 35% del total, respectivamente. Sin embargo, el aumento y la reducción de Z fue estadísticamente significativo (ES) sólo en el 5% y 11% de los

Tabla 2. Evaluación de la precisión en la estimación de días secos y húmedos (ceros y unos). Las etiquetas VC, FC, VP y FP, significan Verdadero Cero, Falso Cero, Verdadero Positivo y Falso Positivo.

MESES	VC	FC	VP	FP
TODOS	97.4	2.6	92.5	7.5
ENE	98.6	1.4	90.6	9.4
FEB	98.7	1.3	90.9	9.1
MAR	98.9	1.2	91.8	8.2
ABR	98.6	1.4	92.5	7.5
MAY	96.6	3.4	93.9	6.1
JUN	95.5	4.5	93.7	6.3
JUL	96.3	3.8	91.5	8.5
AGO	95.7	4.3	92.1	7.9
SEP	94.7	5.3	92.8	7.3
OCT	96.0	4.0	92.9	7.1
NOV	97.9	2.1	92.0	8.0
DIC	98.7	1.3	90.7	9.3

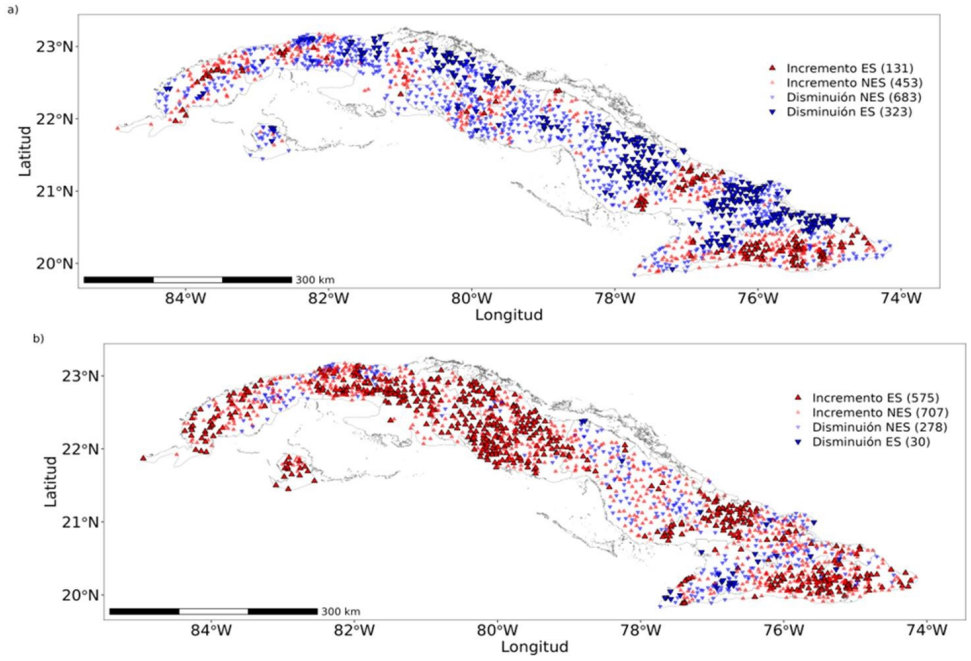

Figura 8. Tendencia de los indicadores CDD (a) y r95ptot (b) para el período 1961-2022. ES y NES se refiere a tendencias significativas y no significativas estadísticamente, respectivamente; mientras que los números entre paréntesis son la cantidad de estaciones en cada categoría. La significación estadística de la tendencia se determinó para $\alpha = 0.05$

casos, respectivamente. Espacialmente, se aprecia que la mayor cantidad de estaciones con incremento y disminución se localizan sobre la mitad oriental de Cuba (Figura 8a). Allí, se concentran los pluviómetros con los incrementos (Las Tunas y Guantánamo) y reducciones (Ciego de Ávila, Camagüey, Holguín y Granma) más pronunciadas, con pendientes de hasta 10 días/década. En la mitad occidental el número de puntos con tendencias ES es relativamente menor y aparecen distribuidas sobre La Habana, noroeste de Matanzas y Villa Clara (disminución), así como en Pinar del Río y Artemisa (aumento).

Por otro lado, la contribución de los días muy húmedos en los acumulados anuales presenta una tendencia mayoritaria hacia el incremento en casi toda Cuba (Figura 8b). En una quinta parte de las estaciones analizadas los valores de Z son ES. Existen algunas provincias, principalmente Camagüey y Granma, donde este indicador refleja Z negativas en un número de estaciones similar a las positivas, aunque por lo general no son ES.

Un análisis combinado de los resultados de CDD y r95ptot permite sugerir que existen zonas del país donde el régimen de precipitaciones tiende a ser cada vez más extremo, toda vez se incrementa el número de días secos consecutivos y aumentan los días muy húmedos (Figura 9). Resulta bastante claro que en casi todas las provincias del país este proceso está presente, aunque en Camagüey y Holguín casi no se observa. También sobresalen los casos de Las Tunas, Santiago de Cuba y Guantánamo, que es donde se concentra la mayor cantidad de pluviómetros donde las tendencias de ambos indicadores son ES.

DISCUSIÓN

El proceso de QC y rellenado de las series de lluvia diaria en el período 1961-2022 permitió crear un conjunto de datos con 2871 estaciones, que se puede considerar como un producto único en Cuba. Este producto puede constituir la base para nuevas investigaciones relacionadas con la actualización de las características del régimen pluviométrico en Cuba, la investigación de las variaciones espaciales y temporales de eventos extremos o la mejor caracterización de los procesos hidrológicos, entre otros.

Se logró identificar que los valores repetidos ocurrían dentro de las estaciones de la misma provincia, lo que sugiere que esas magnitudes fueron asignadas durante el proceso de concentración de datos del INRH, ya fuera en las unidades provinciales o a nivel nacional. Debe notarse que existieron períodos de años donde la detección de datos repetidos fue mucho mayor que en otros (Figura 4), aunque la frecuencia de ocurrencia resultó muy baja. Hay que tener en cuenta que el método para detectar los valores repetidos espacialmente no permitió identificar posibles casos donde el valor repetido fuese cero. La detección de estos casos requeriría de información externa (p. ej. satélite o radar) y de métodos diferentes para realizar ese control.

La aplicación del método de QC produjo resultados similares a los obtenidos por Centella-Artola et al (2023), aunque en este caso, el volumen de información analizada es poco más de 4 veces superior que la utilizada en ese estudio. En general,

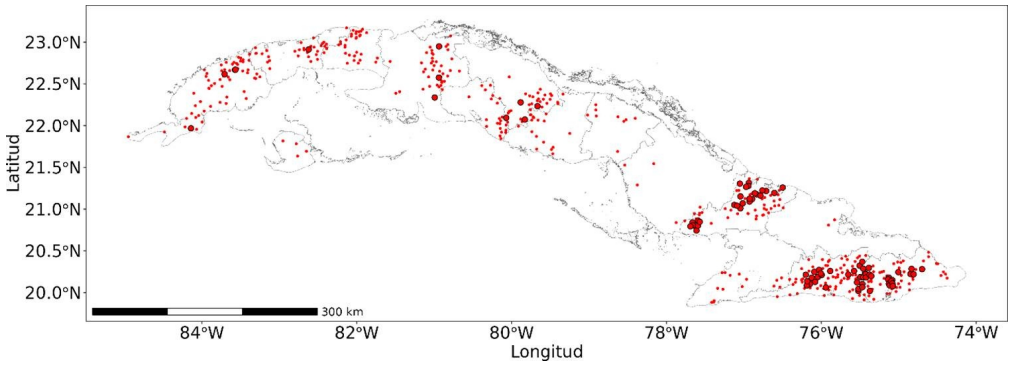

Figura 9. Distribución espacial de los pluviómetros donde la tendencia de los indicadores CDD y r95ptot fue positiva. Los círculos más grandes y con borde negro destacan los puntos en que Z fue ES para ambos indicadores.

el porcentaje de datos sospechosos fue cercano al 14% de la muestra de datos válidos y el tipo de sospechoso más frecuente resultó ser el relacionado con la presencia de "falsos" valores de lluvia en un entorno seco. De acuerdo con los resultados de Serrano-Notivoli et al, 2017b y Škrk N et al (2021), este tipo de dato sospechoso es el más frecuentemente detectado con el método de QC aplicado.

Aunque durante en el QC se detectaron y eliminaron los valores sospechosos, existe la posibilidad de aceptar o no esos resultados, ya sea parcial o totalmente. Hay que considerar, sin embargo, que realizar un proceso de análisis riguroso que permita rechazar los errores detectados tendría un costo elevado y sería muy difícil de implementar en la práctica. Por un lado, el volumen de datos es muy grande y hace muy difícil la revisión por medios manuales utilizando criterios de expertos. Por otro lado, la confirmación objetiva de esos "errores" requeriría de información adicional, la cual no está disponible en el primer tercio del período analizado (datos de satélites o radares).

Se reconoce que la definición de los criterios empleados en el QC es subjetiva, como siempre sucede en este tipo de análisis. Incluso, como señalan Serrano-Notivoli et al (2017a), los casos detectados como QC1 u otros, podrían ocurrir en realidad, pero son situaciones poco representativas de lo que sucede en el entorno definido por las estaciones cercanas, mucho más cuando la densidad espacial de las mediciones es elevada.

La reconstrucción de las series con datos faltantes, ya fuesen originales o resultantes del QC, mostró que ese proceso tuvo resultados satisfactorios en la precisión de la estimación de días secos y húmedos, así como en la magnitud de la lluvia. En general, la flexibilidad adicionada en la versión más reciente del paquete *ReddPrec* respecto al proceso de estandarización, así como la corrección aplicada a los estimados estandarizados "irreales" ofreció resultados más robustos en la reconstrucción de las series con datos faltantes, en comparación con lo obtenido por Centella-Artola et al (2023).

Es importante mencionar que el método aquí empleado ofrece mejores resultados cuando la densidad de estaciones es alta, como sucede en este trabajo. Con ello se asegura que la distancia entre las 10 estaciones vecinas sea relativamente pequeña y que, en los procesos de QC y rellenado, se logre incorporar mejor los efectos de la escala local. Está claro que mientras mayor sea la distancia, los valores de RV serán más regionales que locales. Por ejemplo, en este trabajo se encontró que en el período 1961-1964, donde la cantidad de estaciones era menor, las estimaciones fueron aceptables, pero con una calidad inferior a la lograda cuando el número de estaciones resultó mayor.

Serrano-Notivoli et al (2017a) discuten en detalle lo antes mencionado e indican que el empleo de un grupo menor de 10 estaciones utilizando tres covariables (tres fuentes de variación) necesita de un mínimo de observaciones (en este caso 10). Por un lado, una cantidad menor afecta la fortaleza de los GLM ajustados y la calidad de sus predicciones (se pierden los grados de libertad), mientras que por otro, el uso de más observaciones reduce el carácter local y afecta el contexto en el que se realizan las estimaciones. En el último caso las opciones o criterios del control de calidad pierden validez. Los mismos autores mencionan que el número de 10 observaciones se identificó luego de realizar pruebas utilizando tres covariables. Se debe mencionar también que Centella-Artola et al (2024) lograron buenos resultados empleando un número mayor de observaciones (datos de 15 estaciones vecinas), luego de realizar varias pruebas. Desde luego la cantidad de observaciones fue superior a 10, pues en ese caso se utilizaron modelos con 5 covariables.

Debe recordarse que la estimación de los RV empleados en el QC y el rellenado de las series se realiza ajustando modelos GLM adaptados a cada localidad y día de manera independiente. Esto asegura que se pueda utilizar la mayor cantidad de datos disponibles en cada momento, sin tener que prescindir de los pluviómetros que tengan series incompletas. Obviamente esto hizo que las observaciones que se utilizaron para estimar la lluvia de un punto dato

en un día, podían cambiar de un día a otro, pues se tomaron las 10 más próximas cuyos registros no estuviesen ausentes. Aunque esto pudiera ser considerado una limitación del método empleado, los resultados de la evaluación de la estimación de la magnitud y la ocurrencia de días secos y húmedos demuestran que el mismo funcionó con una precisión alta. Nuevamente, la elevada densidad de estaciones facilitó esos resultados.

Empleando las salidas de las diferentes etapas de trabajo, se conformó el conjunto de datos finales. En el Anexo A se describe la estructura general de los datos, explicando con el mayor grado de detalle su contenido y el significado de las variables que lo componen.

Este nuevo conjunto de datos ofrece nuevas oportunidades para realizar investigaciones relacionadas con la variabilidad y el cambio climático, la hidrología y otras. De hecho, los resultados del análisis de tendencia realizado para CDD y r95ptot, demuestran que en varias zonas del país existen tendencias ES de incremento de las magnitudes de ambos indicadores. Los resultados difieren a lo que se ha reportado anteriormente (Burgos y González 2012, González et al., 2017, CITMA, 2020), en cuanto a la ausencia de tendencias de interés en el caso de CDD, así como al carácter espacialmente limitado de las tendencias ES de aumento de los días muy húmedos. Está claro que en los trabajos anteriores se utilizó un número de estaciones bastante reducido en comparación con el que se ha empleado ahora y por ello las conclusiones a las que arribaron parecen sesgadas por el empleo de un número reducido de estaciones de medición. En estos casos, la extrapolación de los hallazgos a regiones mayores puede resultar inapropiado. Otra diferencia importante es que el período temporal utilizado por los otros autores fue inferior al empleado en este trabajo. Sin embargo, al replicar la evaluación realizada empleando períodos similares, los resultados (no se muestran) continuaron siendo distintos.

La distribución espacial de las tendencias de los indicadores extremos considerados presenta patrones que deben ser estudiados con mayor profundidad para comprender las causas y la dinámica a ellos asociados. Por ejemplo, resulta interesante poder comprender las causas que propician el incremento de los días secos consecutivos y de los días muy lluviosos en algunas zonas y en otras no. De hecho, este tipo de análisis es el que puede verse favorecido con el conjunto de datos desarrollado en este trabajo.

CONCLUSIONES

A partir de los datos diarios de lluvia existentes en Cuba durante el período 1961 - 2022 y utilizando diferentes técnicas estadísticas y de análisis espacial de datos, con el auxilio del lenguaje R, se generó un conjunto de series de datos diarios de lluvia para 2871 estaciones distribuidas en toda Cuba. Esos datos fueron sometidos a un extenso control de calidad espacial que identificó y eliminó automáticamente todas las observaciones sospechosas. Finalmente, las 2871 series fueron rellenadas, con las estimaciones proporcionadas por los modelos GLM ajustados para cada día y estación. Los resultados alcanzados permiten concluir que:

1. Se logró un conjunto de series de precipitación diaria para el período 1961-2022, con un control de calidad riguroso y con los datos faltantes rellenados. Hasta donde se conoce, este es un producto único en Cuba al incluir la documentación que describe su creación, así como una gran cantidad de metadatos.
2. El proceso de organización y limpieza de datos ejecutado utilizando técnicas de análisis espacial, junto con la metodología empleada para realizar el control de calidad y el completamiento de las series de datos produjo resultados favorables. Además se logró que fuera un proceso completamente automatizado que se puede ajustar y emplear para extender las series de datos y mantenerlas actualizadas operativamente.
3. La identificación de dos subconjuntos de datos, que incorporan las estaciones con series más largas, así como las que se actualizan operativamente en la actualidad, puede facilitar el empleo de este producto en estudios e investigaciones diferentes. Por ejemplo, se espera que esta sea una contribución importante para la ejecución del proyecto del nuevo mapa isoyético de Cuba.
4. Utilizando las series de datos reconstruidas en este trabajo se logró demostrar que la cantidad de días secos consecutivos mostró aumentos y reducciones estadísticamente significativas en distintas regiones del país, mientras que la proporción de la lluvia anual debida a días muy lluviosos mostró un incremento más generalizado y significativo. Así existen regiones, principalmente de la región oriental de Cuba, donde el incremento de ambos indicadores sugiere la tendencia a un régimen de lluvias más extremo.

AGRADECIMIENTOS

Los autores de este trabajo desean destacar la contribución del proyecto “Building resilience to drought in Cuba” bajo la subvención de International Development Research Center (IDRC), Ottawa, Canadá. Roberto Serrano disfruta de la ayuda RYC2021-034330-I financiada por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR. Se reconoce la contribución del Instituto Nacional de Recursos Hidráulicos por facilitar el acceso a los datos diarios de lluvia disponible en sus archivos. Se agradece especialmente al Dr. Eduardo Planos por las útiles observaciones realizadas al borrador del artículo.

REFERENCIAS

Barnett, T., Zwiers F., Hegerl, G., Allen, M., Crowly T., Gillett, N., Hasselmann, K., Jones, P., Santer, B., Schnur, R., Stott, P., Taylor, K., & Tett S. (2005). Detecting and attributing external influences on the climate system: A review of recent advances, *J. of Climate.*, 18(9), 1291 -1314. https://doi.org/10.1175/JCLI3329.1

Burgos, Y., y González, I. (2012). Análisis de indicadores de extremos climáticos en la isla de Cuba. *Revista de Climatología*, 12.

Centella-Artola, A., Bezanilla-Morlot, A., Serrano-Notivoli, R., Vázquez-Montenegro, R., Sierra-Lorenzo, M., and Chang-Dominguez, D. (2023). A new long term gridded daily precipitation dataset at high-resolution for Cuba (CubaPrec1). *Data in Brief*, 48, 109294. https://doi.org/10.1016/j.dib.2023.109294

Centella-Artola, A., Serrano-Notivoli, R., Fonseca-Rivera, C., Bezanilla-Morlot, Sierra-Lorenzo, M. (2024). Estimación espacial de la lluvia diaria en una región de Cuba usando modelos lineales generalizados y covariables topográficas. *Investigación Operacional*, FORTHCOMING 62J12-10-23-01, (disponible en https://rev-inv-ope. pantheonsorbonne.fr/sites/default/files/inline-files/PAPER-62J12-10-23-01_0.pdf)

CITMA (2019). Atlas Nacional de Cuba LX Aniversario. Versión 1.0. La Habana: Instituto de Geografía Tropical, GEOCUBA Investigación y Consultoría, CITMATEL.

CITMA (2020). Tercera Comunicación Nacional a la Convención Marco de las Naciones Unidas sobre Cambio Climático. Sello Editorial AMA, Ministerio de Ciencia, Tecnología y Medio Ambiente, ISBN: 978-959-300-170-0

Davitaya, F. F. y Trusov I. I. (1965). Los recursos climáticos de Cuba. Su utilización en la economía nacional. Ed. ACC-INRH, La Habana; 68 p.

Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S., and Lott, N. J. (2000). Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.*, 39 , 1580-1591. https://doi.org/10.1175/1520-0450(2000)039<1580:CASCND>2.0.CO;2

Gagua, G., Zarembo, S., y Izquierdo, A. (1976). Sobre el nuevo mapa isoyético de Cuba. Voluntad Hidráulica, La Habana, Cuba, 37: 35-41.

GarcíaI. T. G., SardiñasS. B., & GonzálezD. H. (2017). Comportamiento de Indicadores de extremos climáticos en la Isla de la Juventud. *Revista Cubana De Meteorología*, *23*(2), 217-225. http://rcm.insmet.cu/index.php/rcm/article/view/241

González GarcíaI. T., Martínez AlvarezM., Gil ReyesL., Alpizar TirzoM., Alonso DíazY., Bocalandro PalmeroM., & Hernández GonzálezD. (2022). Control de la calidad a series de datos diarios de lluvia en el periodo 1961-2008. *Revista Cubana De Meteorología*, *28*(2). http://rcm.insmet.cu/index.php/rcm/article/view/634.

Izquierdo, A. (1989). Precipitación media anual (1964-1983) [mapa 31]. En Nuevo Atlas Nacional de Cuba (p. VI.3.3). España: Gráficas ALBER.

Kendall, M.G. (1970). Rank Correlation Methods, fourth ed. Griffin, London.

Klein Tank, A. M. G., Zwiers, F. W. and Zhang, X. (2009). Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation, Climate data and monitoring WCDMP-No. 72, WMO-TD No. 1500, 56 pp.

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424-425*, 264-277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Lecha L. B., L. R. Paz y B. Lapinel. (1994). El clima de Cuba. Editorial ACC, 186 pp.

Liu, J., Yang, H., Gosling, S. N., Kummu, M., Flörke, M., Pfister, S., Hanasaki, N., Wada, Y., Zhang, X., Zheng, C., Alcamo, J., & Oki, T. (2017). Water scarcity assessments in the past, present, and future. *Earth's Future*, 5(6), 545-559. https://doi.org/10.1002/2016EF000518

Mann, H. B. (1945). Non-parametric tests against trend. Econometrica 13, 245-259.

R Core Team. (2017). R: a language and environment for statistical computing. https://www.r-project.org/

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat.Assoc*. 324, 1379-1389.

Serrano-Notivoli, R., Luis, M. de., Saz, M. A., y Beguería, S. (2017a). Spatially-based reconstruction of daily precipitation instrumental data series, *Clim. Res.*, 73(3), 167-186, https://doi.org/10.3354/cr01476.

Serrano-Notivoli, R., Beguería, S., Saz, M. Á., Longares, L. A., & de Luis, M. (2017). SPREAD: a high-resolution daily gridded precipitation dataset for Spain - an extreme events frequency and intensity overview. *Earth System Science Data*, 9(2), 721-738. https://doi.org/10.5194/essd-9-721-2017

Serrano-Notivoli, R. y Tejedor, E. (2021). From rain to data: A review of the creation of monthly and daily station-based gridded precipitation datasets. *WIREs Water*, 8 (6) e1555, https://doi.org/10.1002/wat2.1555.

Serrano-Notivoli, R. y Centella- Artola. (2024). Reconstruction of Daily Data - Precipitation. (ReddPrec) paquete R version 2.0.3. https://CRAN.R-project.org/package=reddPrec.

Servicio Hidrológico Nacional. (2006). Nuevos logros en el estudio de la pluviosidad en Cuba: Mapa Isoyético para el período 1961- 2000. *Revista Voluntad Hidráulica*, Año XLIV, No. 98, p. 2 - 14

Škrk, N., Serrano-Notivoli, R., Čufar, K., Merela, M., Črepinšek, Z., Kajfež Bogataj, L., de Luis, M. (2021). SLOCLIM: a high-resolution daily gridded precipitation and temperature dataset for Slovenia. *Earth System Science Data*, 13(7): 3577-3592. https://doi.org/10.5194/essd-13-3577-2021.

Trusov, I. I. (1967). Las precipitaciones en la Isla de Cuba. Ed. INRH, La Habana, 64p.

Trusov, I. I., Izquierdo, A. y Díaz, L.R. (1983). Características espaciales y temporales de las precipitaciones atmosféricas en Cuba. Ed. Academia, La Habana; 162 p.

Yilmaz, K. K., Hogue, T. S., Hsu, K., Sorooshian, S., Gupta, H. V., & Wagener, T. (2005). Intercomparison of Rain Gauge, Radar, and Satellite-Based Precipitation Estimates with Emphasis on Hydrologic Forecasting. *Journal of Hydrometeorology*, 6(4), 497-517. https://doi.org/10.1175/JHM431.1

Zhang, X., Alexander, L., Hegerl, G. C., Jones, P., Tank, A. K., Peterson, T. C., Trewin, B., & Zwiers, F. W. (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *WIREs Climate Change*, 2(6), 851-870. https://doi.org/10.1002/wcc.147

Debe notarse que los datos resultantes de todo el proceso no son públicos, pues forman parte de los archivos de datos del Instituto Nacional de Recursos Hidraúlicos de Cuba (https://www.hidro.gob.cu/). A pesar de ello, se consideró oportuno asegurar el máximo de transparencia y facilitar que se pueda reproducir el trabajo realizado. Por esta razón, se generó un archivo datos finales con la mayor cantidad de información posible.El archivo tiene formato CSV (comma separated values), con 64974076 filas y 12 columnas. Utilizando la Tabla A.1 como ejemplo, el archivo se puede describir como sigue:

- **Fecha:** Son los días desde el 1 de enero de 1961 hasta el 31 de diciembre de 2022 en formato yyyy-mm-dd
- **ID**: Es el código utilizado para identificar las estaciones. Utiliza la filosofía empleada por el INRH tomando en consideración el código de la provincia (21-35) y del municipio especial Isla de la Juventud (40), más el número de control de cada equipo dentro de cada provincia.
- **ori**: son los datos originales de lluvia
- **after_dup:** valores de **ori** luego de la revisión de los datos duplicados. Si el dato original estaba duplicado se asignó el código NA (línea 3 de la Tabla A.1)
- **dup:** código que indica si el dato fue removido (1) o no (0) en la revisión de los duplicados.
- **after_qc:** Son los valores luego del control de calidad (QC). Si determinado valor resultó sospechoso, el valor de **after_dup** fue removido (línea 3 de la tabla)
- **qc_codes**: código del tipo de sospechoso que fue detectado y removido en **after_qc.** En la línea 3 se observa que el valor de 31,8 milímetros fue removido al ser detectado como un sospechoso de valor atípico (código 3)
- **after_fill**: Valores estimados durante el proceso de rellenado de las series. Como se explica en el texto esas magnitudes son los estimados estandarizados. Por lo general siempre se pudo tener un estimado y sólo en algunos casos de años iniciales de las estaciones localizadas en la Isla de la Juventud, la estimación no se pudo realizar, pues no existieron 10 estaciones con datos válidos debido a la ausencia de registros, a valores duplicados o a magnitudes que fueron removidas al detectarse como sospechosos.
- **data_end**: Contiene las series de datos reconstruidas, las cuales tienen los datos originales que pasaron el control de calidad (línea 1 de la tabla) o por los datos estimados, cuando originalmente faltaban (fila 4), cuando se removieron como duplicados (fila 3) o uando se eliminaron en el QC (fila 2)
- **error**: Es el error estándar del modelo que estimó el valor **after_fill**. Las magnitudes son en milímetros y sólo se muestran si el valor estimado es utilizado para rellenar la serie (filas 2-4)

Tabla A.1. Ejemplo del fichero con el conjunto de datos finales. El significado de las columnas se explica en el texto.

Fecha	ID	ori	after_dup	dup	after_qc	qc_codes	after_fill	data_end	error
1961-01-01	p21_11	2.5	2.5	0	2.5	NA	11.69	2.5	NA
1961-01-01	p21_7	31.8	31.8	0	NA	3	5.6	5.6	3.2
1961-01-01	p22_53	9.0	NA	1	NA	NA	0	0	0
1961-01-01	p26_812	NA	NA	NA	NA	NA	0	0	0

En el conjunto de datos también se incorpora el fichero con las coordenadas de las estaciones y los valores de altura. Estos últimos se corresponden con las alturas extraídas del Modelo Digital del Terreno que fue utilizado en la confección del último Atlas Nacional de Cuba. A este fichero también se le adicionan dos columnas (st_fill y st_updated) para identificar las estaciones de los subconjuntos climático y operacional, respectivamente (ver Figura 6).

Abel Centella-Artola. Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba, Apdo. 17032, C.P. 11700, Habana 17.
Cecilia Fonseca-Rivera. Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba, Apdo. 17032, C.P. 11700, Habana 17. E-mail: ceciliafonseca91@gmail.com
Roberto Serrano-Notivoli. Departamento de Geografía y Ordenación del Territorio, Instituto Universitario de Ciencias Ambientales, Universidad de Zaragoza, Zaragoza, 50009, España. E-mail: roberto.serrano@unizar.es
Ranses Vázquez-Montenegro. Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba, Apdo. 17032, C.P. 11700, Habana 17. E-mail: ranses.vazquez@insmet.cu
Arnoldo Bezanilla-Morlot. Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba, Apdo. 17032, C.P. 11700, Habana 17. E-mail: arnoldo.bezanilla@gmail.com
Maibys Sierra-Lorenzo. Instituto de Meteorología, Loma de Casablanca, Regla, La Habana, Cuba, Apdo. 17032, C.P. 11700, Habana 17. E-mail: maibyssl@gmail.com
Dimitris A. Herrera. Department of Geography & Sustainability, University of Tennessee-Knoxville, TN, USA. Instituto Geográfico Universitario, Universidad Autónoma de Santo Domingo, Santo Domingo 10103, Dominican Republic. E-mail: dherrer3@utk.edu