Descripción de la base de datos climáticos diarios y los controles de calidad implementados en el Centro Regional del Clima para el Sur de Sudamérica

VERSIÓN 2023

REPORTE TÉCNICO CRC-SAS-2023-001

WWW.CRC-SAS.ORG

Registro de actualizaciones de la publicación y sus antecedentes previos

Fecha	Código del Reporte Técnico	Antecedente o actualización
15/11/2013	CRC-SAS-2013-001	Diseño del proceso del control de calidad de datos climáticos diarios en el Centro Regional del Clima para el Sur de Sudamérica
15/12/2015	CRC-SAS-2014-001	Descripción de controles de calidad de datos climáticos diarios implementados por el Centro regional del Clima para el Sur de Sudamérica
20/04/2023	CRC-SAS-2023-001	Descripción de las estaciones automáticas y su control de calidad. Restructuración del reporte completo de calidad de datos climáticos implementados por el Centro Regional del Clima para el Sur de Sudamérica

Índice

1. Introducción

Para desarrollar nuevos productos y servicios climáticos es indispensable contar con una base de datos robusta, en la que todos sus registros sean de buena calidad. La consistencia de los datos es fundamental para que los productos o servicios derivados a partir de ellos sean confiables y puedan proveer soluciones fundamentadas a los usuarios.

Uno de los principales componentes de una base de datos climáticos regionales es el esquema de control de calidad de los datos que contiene. A través de este esquema de control se identifican los datos erróneos o dudosos posiblemente relacionados con (a) problemas en los registros tomados por los observadores meteorológicos, (b) errores de digitalización o transcripción de la información y (c) fallas electrónicas o mecánicas. El propósito de este documento es describir la compilación y control de calidad de la base de datos de variables climáticas compilados por las instituciones participantes en el Centro Regional del Clima para el Sur de América del Sur (en adelante, CRC-SAS).

El CRC-SAS es un esfuerzo liderado por los servicios meteorológicos e hidrológicos de Argentina, Brasil, Bolivia, Chile, Paraguay y Uruguay, como parte del Marco Mundial para los Servicios Climáticos de la Organización Meteorológica Mundial (OMM) de la Organización de las Naciones Unidas (Freires Lúcio y Grasso, 2016). El objetivo principal del CRC-SAS es la producción y diseminación de datos, información y conocimiento climático que sea útil para apoyar la toma de decisiones en sectores de la sociedad sensibles a la variabilidad y el cambio climático.

2. Datos climáticos

Una de las principales actividades del CRC-SAS es la recopilación de datos diarios de un conjunto determinado de variables meteorológicas (tabla 1) para el período que se extiende desde el 1 de enero de 1961 hasta el presente. Como paso inicial, se compilan los datos provenientes de las estaciones meteorológicas convencionales operadas por los seis países miembros del CRC-SAS. Entre otros requisitos, estas estaciones convencionales deben estar equipadas con instrumentos manuales operados por observadores meteorológicos, quienes deben encargarse de efectuar las mediciones meteorológicas, y pueden también ser responsables del funcionamiento y mantenimiento del instrumental si se les proporciona la formación adecuada (OMM, 2017). La base también incluye datos de estaciones meteorológicas automáticas, en las cuales los datos se registran electrónicamente sin un operador.

La cobertura geográfica de la base de datos del CRC-SAS incluye las estaciones convencionales de Argentina, Bolivia, Brasil (las estaciones ubicadas al sur del paralelo 10 °S), Chile, Paraguay y Uruguay; así como también las estaciones automáticas de Argentina, Brasil (estaciones al sur del paralelo 10 °S), Chile y Uruguay. De las variables meteorológicas que se incluyen en la base de datos, hay tres (temperatura máxima y mínima diaria, precipitación acumulada) que son provistas

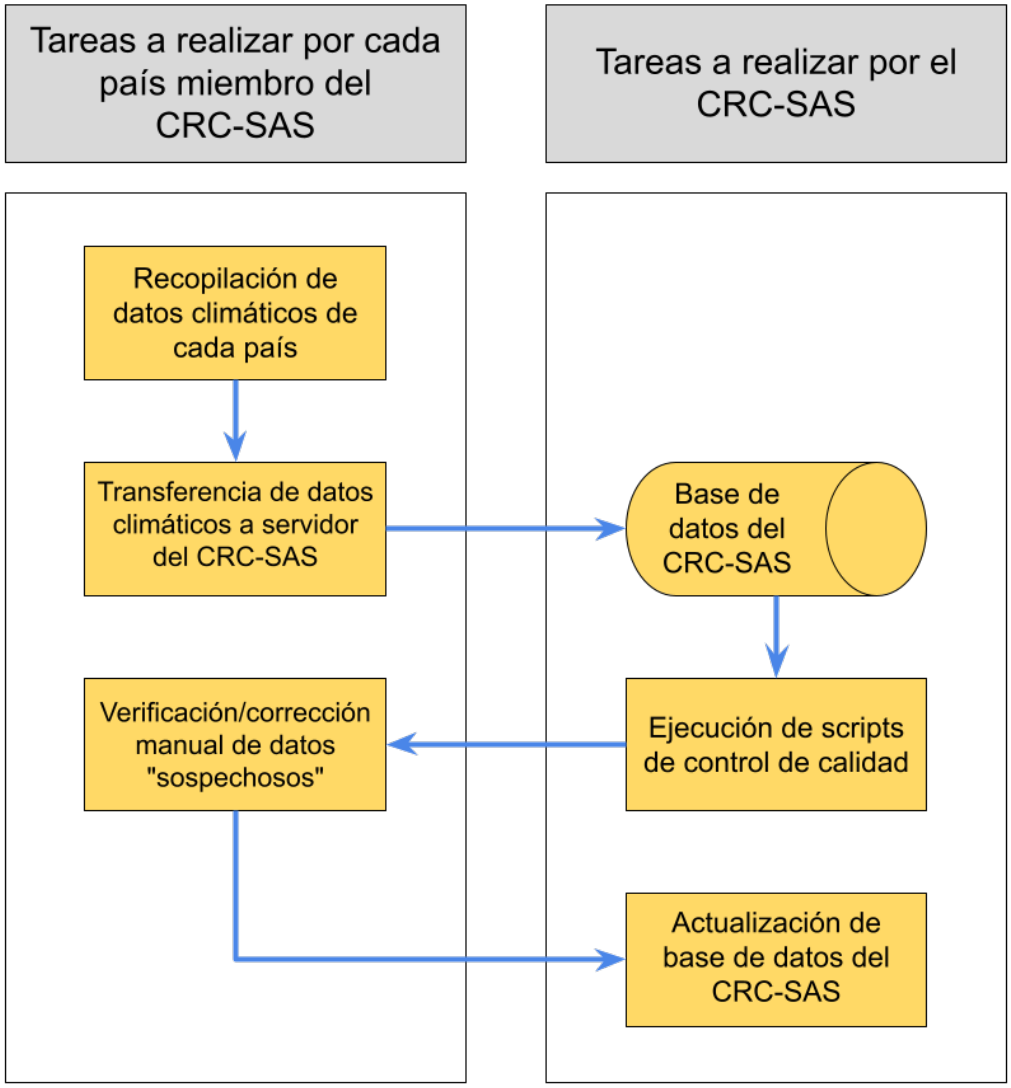
por *todos* los miembros del CRC-SAS que operan tanto estaciones meteorológicas convencionales como automáticas. Las restantes variables representan cantidades útiles para el cálculo de variables climáticas derivadas (por ejemplo, evapotranspiración potencial), necesarias para apoyar la toma de decisiones en sectores sensibles al clima. Se espera que eventualmente todas estas variables (y otras que sean identificadas como necesarias) sean añadidas a la base de datos del CRC-SAS por todos sus miembros.

Además de los datos climáticos, la base de datos del CRC-SAS también incluye metadatos, o sea datos sobre las estaciones meteorológicas. La OMM desarrolla estándares para la recopilación de metadatos. Por este motivo, los metadatos incluidos en la base de datos del CRC-SAS, por el momento, se restringen a lo listado en el Apéndice B.

Tabla 1. Variables meteorológicas incluidas en la base de datos del Centro Regional del Clima para el Sur de América del Sur (CRC-SAS). Las variables en las líneas coloreadas serán contribuidas por todos los miembros del CRC-SAS; las restantes son, por el momento, opcionales y servirían para calcular una serie de productos derivados. El número de observaciones diarias indica cuántas observaciones se usan para calcular variables agregadas para un día (por ejemplo, la temperatura media diaria).

VARIABLE	NOMBRE ABREVIADO	UNIDADES
Temperatura máxima diaria	tmax	grados Celsius (°C)
Temperatura mínima diaria	tmin	grados Celsius (°C)
Temperatura media diaria	tmed	grados Celsius (°C)
Temperatura de rocío	td	grados Celsius (°C)
Presión atmosférica al nivel de la estación	pres_est	hectopascales (hPa)
Presión atmosférica reducida al nivel medio del mar	pres_nm	hectopascales (hPa)
Precipitación acumulada	prcp	milímetros (mm)
Humedad relativa	hr	porcentaje (%)
Horas diarias de sol (heliofanía)	helio	horas
Cobertura nubosa	nub	octas (octavos)
Dirección del viento máximo diario	vmax_d	decenas de grado
Velocidad del viento máximo diario	vmax_f	metros por segundo (m s^{-1})
Velocidad media del viento	vmed	metros por segundo (m s^{-1})
Número de observaciones diarias	num_observaciones	sin unidades

Figura 1. Organización general del proceso de compilación y control de calidad de datos meteorológicos diarios en la base de datos del CRC-SAS.



3. Flujo del procesamiento de datos meteorológicos en el CRC-SAS

El uso de los datos climáticos en la base de datos del CRC-SAS para la generación de información climática útil y relevante requiere que los datos hayan sido sometidos a un proceso de control de calidad. Este proceso debe identificar valores sospechosos que podrían ser incorrectos y, en consecuencia, afectar indebidamente los productos o estadísticas derivados a partir de los datos originales. Actualmente, el control de calidad de los datos es realizado en forma centralizada por el CRC-SAS, siguiendo el flujo que se ilustra en la figura 1. En las siguientes secciones se describe muy sucintamente cada una de las etapas principales en el proceso.

3.1 Recopilación de datos climáticos de cada país miembro del CRC-SAS

Cada país miembro definirá las estaciones meteorológicas convencionales y automáticas cuyos datos se contribuirán a la base de datos del CRC-SAS. Los criterios que se han consensuado inicialmente sugieren que se contribuyan datos para las estaciones convencionales:

- que estén actualmente activas; y
- que hayan operado por al menos 10 años consecutivos en el período desde el 1 de enero de 1961 al presente.

Debe enfatizarse que la acción de contribuir datos nacionales al CRC-SAS no implica que esos datos sean distribuidos a cualquier entidad o persona que los solicite. La diseminación de datos meteorológicos diarios será definida por cada país miembro. Un miembro puede estipular que sus datos puedan ser usados para la generación de productos por parte del CRC-SAS, pero que no sean publicados o transferidos a terceros.

Cada país se ha comprometido a proveer los datos en el formato consensuado por los miembros. El formato de los archivos de texto que se definió para la transferencia de los datos se ilustra en la figura 2.

Esto es válido para las estaciones convencionales. El flujo para las automáticas es diferente y no responde a la tabla 2.

Figura 2. Ejemplo de formato de archivo de texto con datos meteorológicos diarios a ser incluidos en la base de datos del CRC-SAS. Las variables cuyo nombre se abrevia en el encabezamiento del archivo se describen en el Apéndice A; el texto se escribe en forma vertical por razones de espacio.

omm_id	fecha	tmax	tmin	tmed	td	pres_est	pres_nm	prcp	hr	helio	nub	vmax_d	vmax_f	vmed	num_obs
87444	2002-07-31	11.4	2.2	6.1	-5.1	953.1	1027.8	0.001	48	8.7	1	\N	\N	7.4	4
87444	2002-08-01	12.6	-3.2	3.9	-8.1	952.8	1027.9	0	48	7.8	3	36	5.1	0.9	4
87444	2002-08-02	15.5	-8.8	4.3	-5.7	949.8	1025.2	23.2	56	8.9	1	36	4.6	0.9	4

3.2 Transferencia de datos climáticos a servidor FTP en CRC-SAS

Los archivos de transferencia de datos climáticos preparados por cada país deberán ser copiados a un sitio FTP del CRC-SAS. Para realizar la transferencia, cada miembro del Centro utilizará un nombre de usuario y contraseña a ser provisto por el CRC-SAS.

Un *script* de importación se activará periódicamente y listará archivos de transferencia en el sitio FTP que no se hayan procesado anteriormente. Este *script* realizará algunos controles de integridad de los datos –como verificar que no haya fechas duplicadas o que todos los datos importados para una estación tengan el mismo código de identificación– e importará los datos a la base de datos relacional del CRC-SAS. En caso de que haya problemas en el formato o contenido de los archivos de transferencia, se enviará en forma automática un mensaje a la casilla de correo electrónico del contacto designado por cada país para asuntos relacionados con la transferencia de datos.

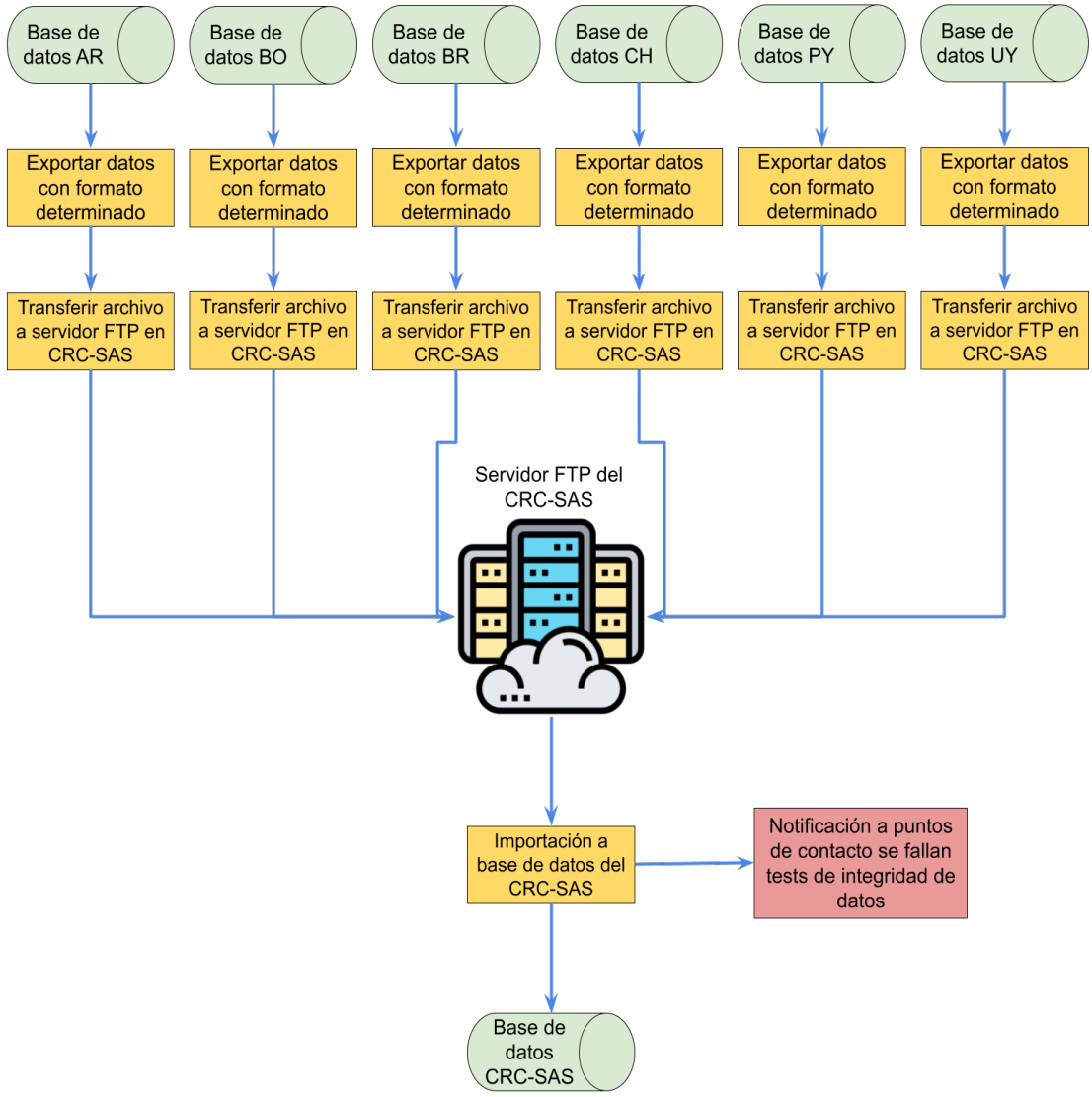
Cada registro en la base de datos corresponde a una combinación única de una fecha y un ID de estación meteorológica. En algunos casos, los datos meteorológicos de una estación solamente

incluyen aquellas fechas para las cuales haya datos. Por ejemplo, si la estación no ha observado datos entre el 25 y el 31 de agosto de 1962, los datos incluirán una fila para el 24 de agosto y, a continuación, una fila para el 1 de septiembre. Durante el proceso de importación, se completarán las series de modo de que haya una secuencia temporal completa. Para los registros que se agreguen, los valores de todas las variables (excepto fecha e ID de estación) se llenarán con el código establecido para datos faltantes. Siguiendo con el ejemplo anterior, en la base de datos se incluirán registros para las fechas del 25 al 31 de agosto de 1962. Todos los datos contendrán el valor “\N”.

Como parte del proceso de importación, se asigna una etiqueta “pendiente” a cada registro recientemente agregado a la base. De esta manera, la base de datos identifica los registros que deben pasar por los controles de calidad, que se realizan sólo una vez para cada registro. El proceso de transferencia de los datos se ilustra en la figura 3. Este proceso puede sufrir modificaciones para diferentes países en distintos momentos.

Cabe aclarar que para la transferencia de datos meteorológicos provenientes de estaciones automáticas se utilizan apis o servicios web que proveen las instituciones.

Figura 3. Transferencia de datos meteorológicos diarios de los países miembros a la base de datos del CRC-SAS. El formato del archivo (o archivos) de transferencia se describe en el texto. Cada miembro del CRC transferirá esos archivos a un sitio FTP, de donde serán importados a la base de datos del Centro.



4. Organización general del esquema de controles de calidad

El uso de variables climáticas para la generación de información climática útil y relevante requiere que los datos hayan sido sometidos a un proceso de control de calidad. Este proceso debe identificar valores sospechosos que podrían ser incorrectos y, en consecuencia, afectar indebidamente los productos o estadísticas derivados a partir de ellos. Por esta razón, el CRC-SAS –en colaboración con proyectos de investigación financiados por el Instituto Interamericano para la Investigación del Cambio Global (IAI), el Banco Interamericano de Desarrollo (BID) y el programa Euroclima+ de la Unión Europea– ha implementado una serie de procedimientos, publicados en la literatura científica, para el control de calidad de los datos meteorológicos. El flujo de información asociado con la recopilación, actualización y control de calidad de la base de datos del CRC-SAS se describe en la sección anterior. En esta, en cambio, se describen los detalles de los controles de calidad implementados.

Los controles a la base de datos se realizan en dos etapas. En la primera etapa, los datos climáticos se someten a una serie de controles estadísticos que identifican registros que contienen variables meteorológicas con valores sospechosos. La segunda etapa del control de calidad involucra la verificación manual de los valores sospechosos identificados en la etapa anterior mediante la verificación de los registros originales (en papel o formato digital), que tienen una mayor resolución temporal. Esto último facilita controlar las observaciones horarias de ese día (sea en papel o en formato digital). Por ejemplo, si es necesario corroborar el valor de la temperatura máxima de un día determinado, se puede revisar la libreta meteorológica (que contiene el registro diario de la estación, generalmente en papel). Dado que la verificación de datos sospechosos es realizada previamente por cada servicio meteorológico miembro del CRC-SAS (es decir, cada servicio verifica los datos de su país) la proporción de datos que pasan por el control de calidad puede variar entre países. En el caso de las estaciones automáticas no se realiza un control manual de datos sospechosos.

4.1 Familias de controles de calidad

Los controles de calidad se organizan en seis "familias" con características similares (figura 4). En esta sección se presenta una breve descripción de cada familia, mientras que en las secciones subsiguientes se detallan los controles incluidos en cada uno de ellos.

1. **Controles generales.** Estos controles verifican la integridad general de los datos. Por ejemplo, se determina que no haya fechas duplicadas o fuera de secuencia en las observaciones diarias. También se verifica la frecuencia con la cual se registran los valores decimales para cada variable: desvíos muy marcados con respecto a una distribución aproximadamente uniforme de valores decimales de 0 a 9 (en el caso de las temperaturas, que se registran con un solo decimal) pueden alertar sobre la existencia de problemas potenciales en los datos. Algunos de estos controles generales se implementan en el proceso de actualización periódica de datos (por ejemplo, se identifica la repetición de fechas antes de actualizar la información para una estación dada en la base de datos del CRC-SAS).

2. **Controles de rango fijo.** Estos controles aseguran que no existan valores físicamente imposibles o nunca observados en el registro histórico. Los límites propuestos son fijos para cada variable durante todo el periodo de datos y para todas las estaciones meteorológicas. Por ejemplo, una temperatura máxima diaria de 99 °C está por encima del récord mundial. Es posible que un valor así corresponda a un código de valor faltante que no se ha definido apropiadamente como tal.

3. **Controles de rango variable.** En esta familia, los rangos o umbrales usados para "marcar" valores sospechosos varían con el tiempo, tomando valores específicos para cada día o mes del año, por lo que los controles son más finos o sensibles que los controles de rango fijo. Por ejemplo, se puede ajustar un ciclo estacional a los valores de temperatura mínima diaria, y los valores extremos se evalúan con respecto al valor esperado del ciclo anual para una fecha determinada.

4. **Controles de continuidad temporal.** Estos controles estudian las secuencias de valores de cada variable en días consecutivos. Algunos de los controles en esta familia detectan la presencia de saltos o picos inusuales en las series de datos. Por ejemplo, un valor de temperatura media muy bajo en relación con los valores de los días adyacentes (el anterior y el siguiente) puede ser marcado como *sospechoso*. Otros controles en esta familia identifican secuencias largas con valores idénticos.

5. **Controles de consistencia entre variables.** Los controles en esta familia evalúan la consistencia entre valores de pares o grupos de variables que deben guardar cierta consistencia entre sí. Un ejemplo obvio es la verificación de que la temperatura mínima diaria sea menor o igual que la temperatura máxima diaria.

6. **Controles de consistencia espacial.** Todos los controles descriptos anteriormente se realizan sobre los datos de una única estación meteorológica (aunque en algunos controles se use más de una variable). En esta familia de controles se incorporan, además, controles

entre estaciones, en los cuales los valores de una variable para una estación determinada (que generalmente se denomina *estación central*) se comparan con valores de esa variable registrados en estaciones geográficamente cercanas (o *estaciones vecinas*).

Cada familia puede incluir varios tipos de controles basados en distintos cálculos. Por otra parte, no todos los controles se aplican a todas las variables: por ejemplo, hay controles que se utilizan solamente para datos de precipitación. La tabla 2 lista los controles incluidos hasta el momento en cada familia y las variables a las cuales se aplica cada uno de ellos.

Figura 4. Familias de controles de calidad. Dentro de cada familia existen controles que se aplican a diferentes variables.

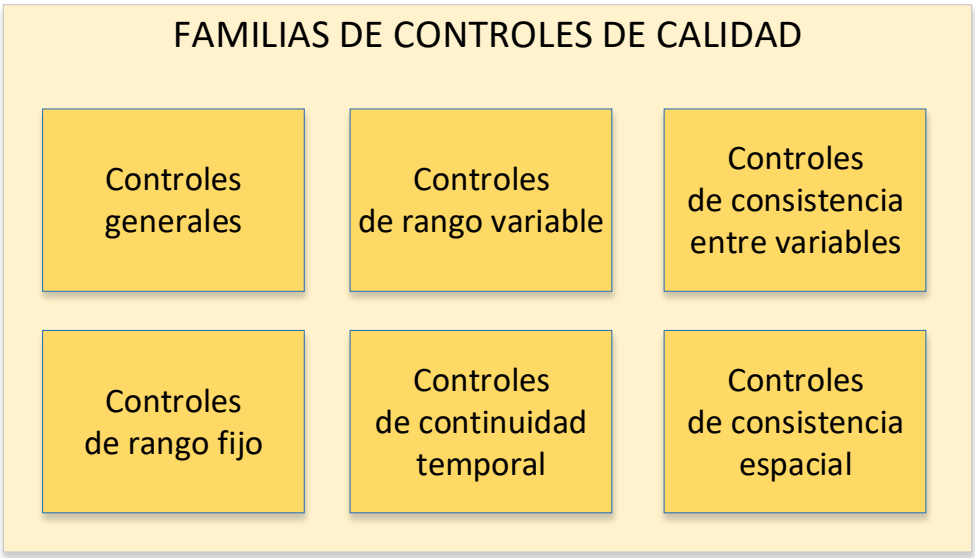


Tabla 2. Controles incluidos en cada familia de controles de calidad y variables meteorológicas a las cuales se aplica cada control.

1. CONTROLES DE RANGO FIJO (SECCIÓN 5)	tmax	tmin	tmed	td	hr	prcp	pres_est	pres_nm	vmax_d	vmax_f	vmed	helio	nub
1.1 Límites inferiores y superiores constantes (valores en tabla 3) (RF01)	*	*	*	*	*	*	*	*	*	*	*	*	*

2. CONTROLES DE RANGO VARIABLE (SECCIÓN 6)	tmax	tmin	tmed	td	hr	prcp	pres_est	pres_nm	vmax_d	vmax_f	vmed	helio	nub
2.1 Apartamientos respecto al ciclo estacional (RV01)	*	*	*	*	*		*	*				*	
2.2 Apartamientos respecto a múltiplos del rango intercuartil para ventanas de 3 o 5 días (RV02)	*	*	*	*	*		*	*					
2.3 Apartamientos respecto a estadísticos resistentes (*biweight*) para ventanas de 3 o 5 días (RV03)	*	*	*	*	*		*	*					
2.4 Heliofanía medida y heliofanía astronómica teórica (RV04)												*	
2.5 Apartamientos respecto al rango intercuartil de la precipitación para ventanas mensuales (RV05)						*							
2.6 Valores extremos de precipitación (cuantiles de la distribución gamma) (RV06)						*							
2.7 Apartamientos de la amplitud térmica, respecto a estadísticos resistentes (*biweight*) para ventanas mensuales (RV07)	*	*											

3. CONTROLES DE CONTINUIDAD TEMPORAL (SECCIÓN 7)	tmax	tmin	tmed	td	hr	prcp	pres_est	pres_nm	vmax_d	vmax_f	vmed	helio	nub
3.1 Persistencia de valores constantes por más de 3 días consecutivos (CT01)	*	*	*	*	*	*	*	*	*	*	*	*	*
3.2 Persistencia extrema (definida climatológicamente) de días sin precipitación (CT02)						*							
3.3 Saltos excesivos entre días consecutivos (CT03)	*	*	*	*	*		*	*					
3.4 Picos de corta duración (1 día) (CT04)	*	*	*	*	*		*	*					
3.5 Picos muy pronunciados entre días previos y posteriores (CT05)	*	*	*	*	*		*	*					
3.6 Picos de corta duración (1 día) con el paquete "*anomalize*" en CRAN (CT06)	*	*	*	*	*		*	*		*	*		

4. CONTROLES DE CONSISTENCIA ENTRE VARIABLES (SECCIÓN 8)		tmax	tmin	tmed	td	hr	prcp	pres_est	pres_nm	vmax_d	vmax_f	vmed	helio	nub
4.1	Consistencia entre temperaturas mínima, media y máxima diarias (CEV01)	*	*	*										
4.2	Consistencia entre temperaturas medias diarias calculadas de diferentes formas (CEV02)	*	*	*										
4.3	Consistencia entre la temperatura máxima en un día y las temperaturas mínimas en días adyacentes (CEV03)	*	*											
4.4	Consistencia entre la temperatura mínima en un día y las temperaturas máximas en días adyacentes (CEV04)	*	*											
4.5	Consistencia entre temperatura media diaria y temperatura de rocío (CEV05)			*	*									
4.6	Consistencia entre presión atmosférica en una estación y la presión equivalente a nivel del mar (CEV06)							*	*					
4.7	Consistencia entre velocidad y dirección medias del viento (CEV07)										*	*		
4.8	Consistencia entre velocidad y dirección máximas del viento (CEV08)									*	*			
4.9	Consistencia entre nubosidad y precipitación (CEV09)						*							*
4.10	Consistencia entre nubosidad y heliofanía (CEV10)												*	*
4.11	Consistencia en la amplitud térmica diaria (CEV11)	*	*											

5. CONTROLES DE CONSISTENCIA ESPACIAL ENTRE ESTACIONES (SECCIÓN 9)		tmax	tmin	tmed	td	hr	prcp	pres_est	pres_nm	vmax_d	vmax_f	vmed	helio	nub
5.1	Regresión espacial ponderada (CES01)	*	*	*	*									
5.2	Regresión espacial basada en el índice de concordancia (*index of agreement* de Legates y McCabe) (CES02)	*	*	*	*									
5.3	Corroboración espacial de la temperatura (CES03)	*	*	*	*									
5.4	Corroboración espacial de la precipitación (CES04)						*							
5.5	Diferencias con valores interpolados a partir de datos vecinos (CES05)	*	*	*	*	*			*		*	*		

4.2 Implementación de los controles de calidad

Todos los controles de calidad están implementados en el lenguaje R, un entorno de programación diseñado para realizar análisis estadísticos y visualizar datos (R Core Team, 2013). R es un software abierto y sin costo y está disponible para varias plataformas (Windows, Mac OS, Linux), bajo los términos de la Licencia Pública General GNU (GNU-GPL, por sus siglas en inglés; ver http://www.r-project.org/Licenses/GPL-3).

Una ventaja del lenguaje R es la existencia de una gran variedad de paquetes o librerías contribuidos por la comunidad de usuarios a nivel mundial que expanden la funcionalidad del lenguaje (ver, por ejemplo http://crantastic.org o http://dirk.eddelbuettel.com/cranberries). Estos paquetes contienen cálculos ya programados, lo que permite reducir tanto el tiempo de implementación como la probabilidad de cometer errores de programación. Por ejemplo, existen al menos dos paquetes para calcular el largo teórico del día (o sea, el número máximo de horas de sol) en función de la latitud y día del año.

Para facilitar la organización y mantenimiento del código, cada familia de controles se implementó en un *script* separado. Los controles incluidos en el *script* correspondiente a cada familia se listan en la sección 4.1. Los *scripts* que ejecutan cada familia de controles son llamados desde un *script* maestro que además realiza tareas generales, como conectarse con la base de datos, etc. Otro *script* contiene funciones programadas en R que realizan tareas que se utilizan más de una vez. El encapsular tareas repetidas en funciones –por ejemplo, el cálculo de una media y un desvío estándar resistentes de una variable usando la función de ponderación *biweight* (Lanzante, 1996)– permite evitar la duplicación de código y, por lo tanto, disminuir la chance de errores y facilitar el mantenimiento de los *scripts*.

4.3 Resultados de los controles de calidad

Cada control de calidad produce como resultado un valor lógico (TRUE o FALSE) para cada variable, cada fecha y cada estación meteorológica. Un valor que supera exitosamente o "pasa" un control determinado toma el valor TRUE (verdadero); si, en cambio, el control falla, el resultado es FALSE y el valor de la variable se identifica como sospechoso. En algunos casos, los resultados de los controles se asignan a más de una variable a la vez: por ejemplo, si un registro falla un control de consistencia entre dos variables, los valores de esas dos variables pueden ser sospechosos y, por lo tanto, ambas variables reciben el valor FALSE como resultado del control en cuestión.

Los umbrales o valores límite usados para cada uno de los controles propuestos fueron ajustados para que la tasa de falsas alarmas –o sea, datos identificados como potenciales errores, pero realmente correctos– fuera lo más baja posible. A la vez, se procuró una detección eficiente de errores, buscando balancear la relación entre el tiempo necesario para la verificación manual subsiguiente de los datos dudosos y la tasa de detección de registros realmente erróneos. Estos aspectos se discuten en detalle más adelante en este documento.

El resultado de esta etapa es una serie de datos marcados como “sospechosos”, que a continuación deben verificarse en forma manual usando como fuente los registros oficiales. Es decir, los controles de calidad no intentan “corregir” los datos, sino que se limitan a identificar valores que pueden ser errores verdaderos o no.

Etiquetas de registro. Antes del control de calidad, los registros recientemente importados tienen una etiqueta que indica su estado “pendiente” (o sea, que debe pasar por los controles). Una vez que un registro pasa por el control de calidad, se le asigna una nueva etiqueta cuyo valor depende del resultado de los controles de calidad. Las etiquetas asignadas a cada registro después de los controles de calidad pueden tomar los valores:

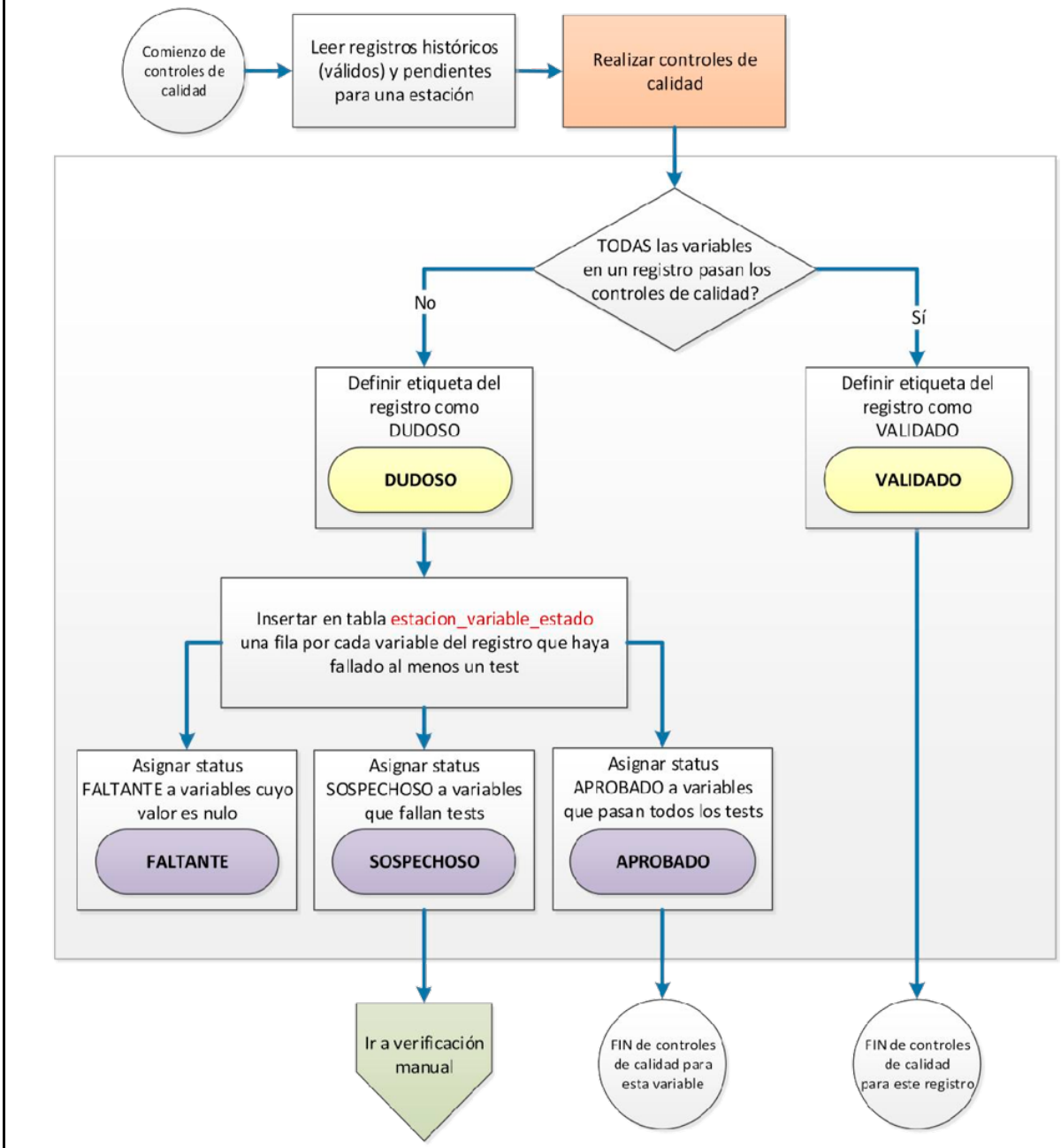
- “Dudoso” en el caso de que al menos una variable en el registro falle algún control de calidad; o
- “Validado” si todas las variables en un registro pasan todos los controles de calidad.

Etiquetas de variables. Además de las etiquetas que indican el estado de todo un registro completo, se asignan etiquetas a cada variable en un registro dentro de la base de datos. Después del control de calidad (pero antes de la verificación manual de los datos dudosos, descrita en la sección siguiente), una variable en un registro puede tomar los valores:

- “Faltante” si el valor de la variable no está disponible;
- “Sospechoso” si la variable falla al menos uno de los controles de calidad; o
- “Aprobado” si la variable pasa exitosamente todos los controles.

El flujo de asignación de etiquetas tanto a registros como a variables individuales se ilustra en la figura 5.

Figura 5. Flujo de asignación de etiquetas de estado a registros y variables individuales durante la etapa de controles de calidad.

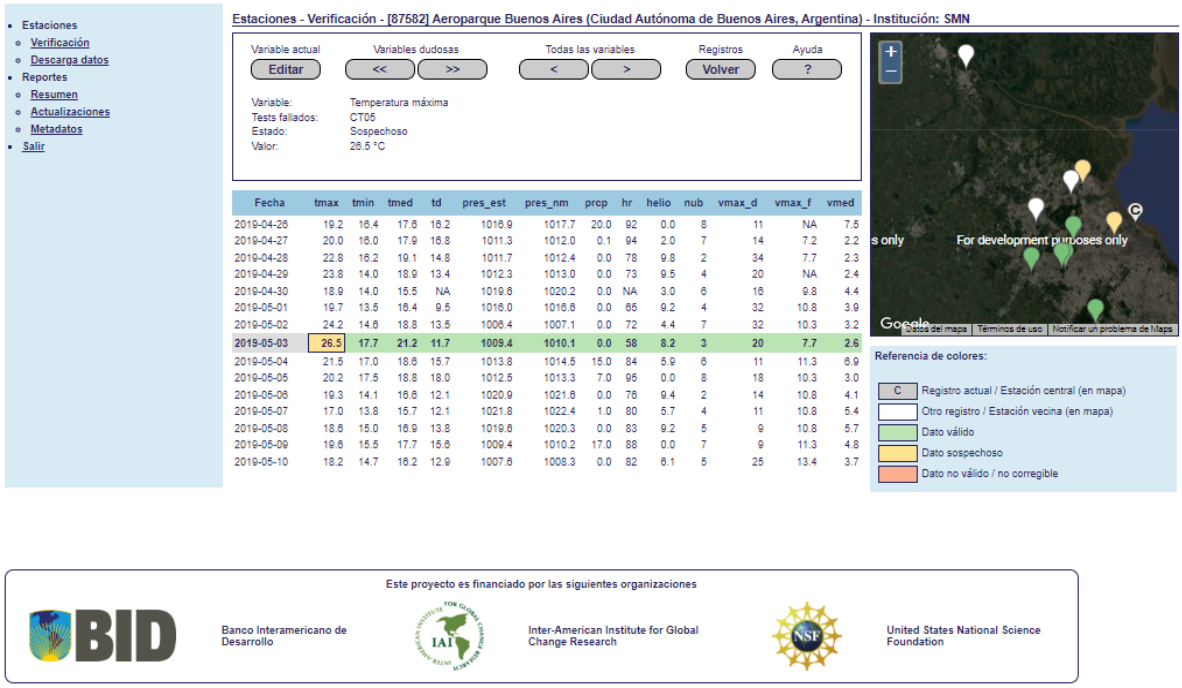


4.4 Verificación manual de datos sospechosos

Los controles de calidad descriptos en la sección anterior no intentan corregir automáticamente datos marcados como sospechosos. Los controles simplemente generan listas de posibles errores que en esta etapa deben ser verificadas manualmente. La verificación involucra la comparación de los valores sospechosos con la versión “oficial” de estos valores, por ejemplo, las libretas donde los observadores de cada estación meteorológica registran los datos. Esta verificación será realizada por personal de los servicios meteorológicos e hidrológicos (SMHs) de cada país miembro del CRC-SAS, ya que las fuentes oficiales de datos se encuentran en cada país.

Para la verificación de datos sospechosos, el CRC-SAS (en colaboración con un proyecto financiado por el Banco Interamericano de Desarrollo) ha desarrollado un procedimiento para que el personal de los SMHs miembros pueda acceder remotamente (vía un *browser*) a la base de datos del CRC-SAS. Se muestra una pantalla con cada registro “dudoso” (o sea, aquel para el cual por lo menos una variable haya fallado en algún test) y los registros para 5-6 días antes y después del registro examinado (para tener un contexto temporal). Para cada variable con valores sospechosos en un registro, el procedimiento también muestra los valores de esa variable en las estaciones geográficamente adyacentes (lo que proporciona un contexto espacial para la interpretación de valores sospechosos). Se espera también que, durante este procedimiento, el operador de un SMH tenga acceso a los registros meteorológicos oficiales. Un ejemplo de una de las pantallas del sistema de verificación se muestra en la figura 6.

Figura 6. Ejemplo de una de las pantallas del sistema de verificación manual de datos sospechosos. Personal de cada país miembro accederá a la base de datos del CRC y controlará los valores considerados sospechosos utilizando registros. El registro siendo analizado se resalta, y las variables se colorean de acuerdo a los resultados de los controles de calidad (por ejemplo, en amarillo se indican los valores sospechosos). El mapa a la derecha de la pantalla proporciona contexto espacial, mostrando valores en estaciones vecinas (capacidad a implementar).



Durante el proceso de verificación manual, un operador (a) selecciona una estación meteorológica, (b) recibe una pantalla con registros sospechosos que todavía no han sido verificados, (c) controla cada valor sospechoso contra los registros oficiales) y (d) selecciona un estado final para cada variable sospechosa en un registro. Una vez que el operador selecciona un estado final para todas las variables sospechosas en un registro, la base de datos actualiza las etiquetas para ese registro, que pasa de "pendiente" a "revisado". También se actualizan las etiquetas correspondientes a cada variable en el registro. De acuerdo a la decisión del operador, la etiqueta de una variable puede tomar diferentes valores:

- "Ratificado" en el caso de que el valor sospechoso de la variable se haya confirmado usando los registros oficiales;
- "Corregido" si el valor sospechoso de la variable se corrigió en base a los registros oficiales (por ejemplo, una temperatura máxima de "41 °C" en julio se cambia a "14 °C", sugiriendo una trasposición de dígitos);
- "Eliminado" en el caso de que un valor sospechoso no pueda corregirse usando los

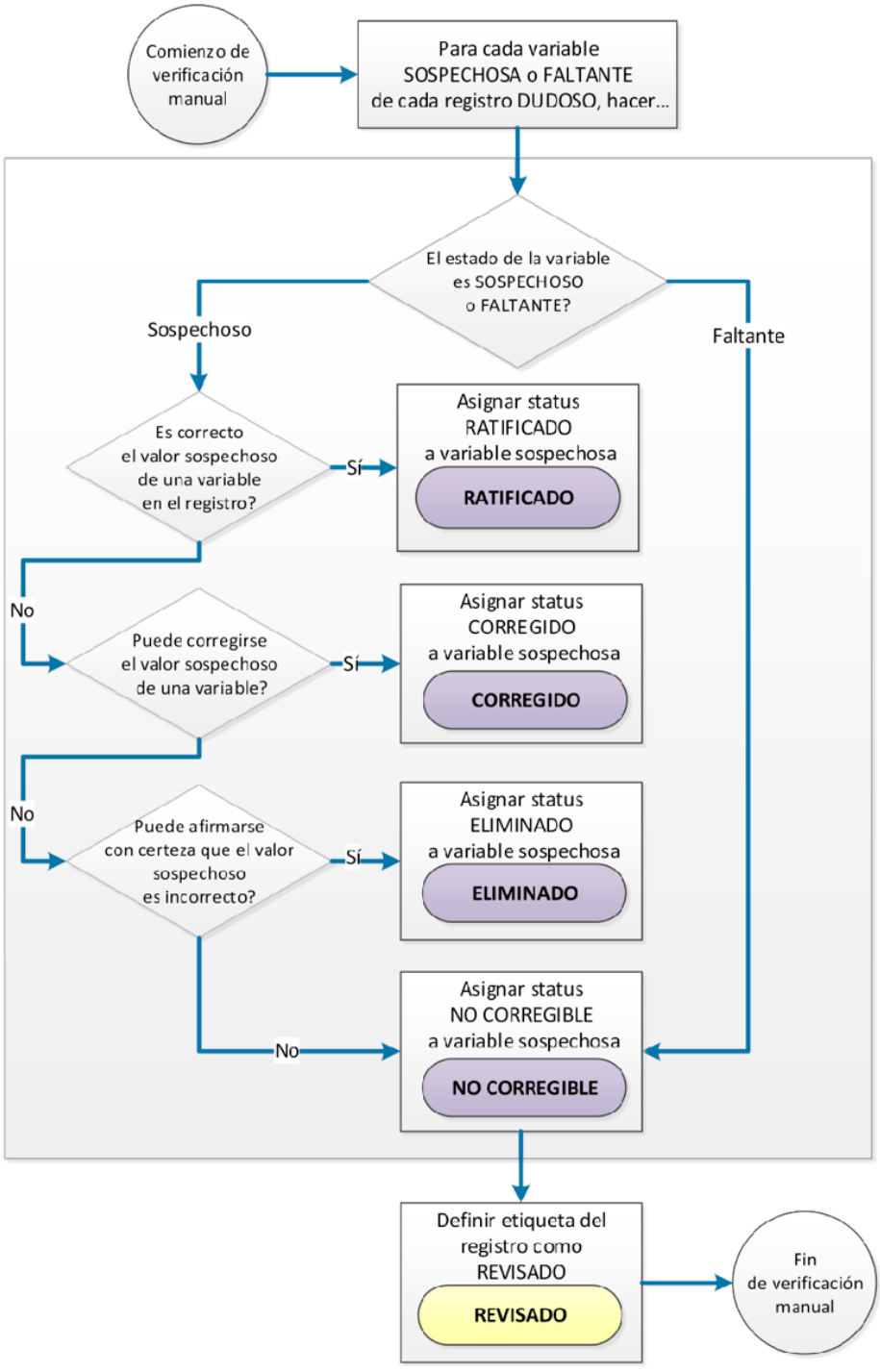
registros oficiales, pero haya elementos para concluir que el valor es claramente erróneo;
- "No corregible" cuando (a) un valor sospechoso no puede corregirse usando los registros oficiales pero (b) NO hay elementos suficientes para concluir que el valor es erróneo (o sea, el valor es plausible).

4.5 Actualización de la base de datos del CRC-SAS

El proceso de verificación manual de datos sospechosos actualiza automáticamente la base de datos del CRC-SAS cada vez que el operador de cada SMH pasa de un registro sospechoso al siguiente. Al terminar de verificar cada registro, las etiquetas de ese registro y de cada variable en el registro se actualizan automáticamente de acuerdo a las opciones elegidas por el operador. Una vez completada la verificación manual de un registro, la base de datos le asigna la etiqueta "Revisado." El flujo de asignación de etiquetas de estado a registros y variables individuales durante la etapa de verificación manual de datos se ilustra en la figura 7. Los valores posibles de las etiquetas para las variables en cada registro se discutieron en la sección anterior.

Un aspecto importante del proceso de control y actualización de datos es que los valores sospechosos que se confirmen como erróneos se guardan en la base de datos en tablas separadas. Es decir, si los registros oficiales indican que el valor registrado para una variable es incorrecto y debe reemplazarse, el valor originalmente almacenado en los datos se preserva, de modo de poder reconstruir los datos originales si fuera necesario.

Figura 7. Flujo de asignación de etiquetas de estado a registros y variables individuales durante la etapa de verificación manual de datos.



5. Familia de controles de rango fijo

La primera familia de controles –controles de rango fijo– compara el valor de una variable meteorológica con valores extremos preestablecidos. Un valor se identifica como sospechoso si queda fuera del intervalo válido definido para cada variable. Los extremos del intervalo se consideran incluidos dentro del rango correcto de datos: por ejemplo, si el límite inferior del rango aceptado para temperatura mínima diaria es de -39.0 °C, un valor de -39.1 °C fallará el control, mientras que un registro de -39.0 °C será considerado válido por este control. El intervalo o rango aceptado es fijo para el análisis de todos los datos de cada variable y para todas las estaciones meteorológicas almacenadas.

Los límites propuestos para cada variable se seleccionaron en base a los datos históricos y/o valores físicamente plausibles (tabla 3). Por ejemplo, para las temperaturas (máxima, mínima, media) por el momento se utilizan como umbrales valores extremos históricos para Argentina (-39.0 °C, 49.0 °C), mientras que para los valores de humedad relativa se considera el intervalo teórico (0 %, 100 %). El intervalo aceptado para cada variable es muy amplio con el objetivo de encontrar valores claramente mal transcriptos y/o errores en el traspaso de la información de distintas fuentes a la base de datos.

Los controles de rango fijo fueron utilizados por Feng et al. (2004), Estévez et al. (2011) y Meek & Hatfield (1994), entre otros, y constituyen el primer paso en los controles de calidad. En las siguientes familias se aumenta la exigencia para que la identificación de datos sospechosos sea más rigurosa.

Tabla 3. Límites de los rangos de valores aceptables utilizados en los controles de rango fijo (RF1).

VARIABLE	INTERVALO VÁLIDO
Temperaturas	-39.0 °C ≤ temperaturas ≤ 49.0 °C
Presión en estación	530 hPa ≤ presión en estación ≤ 1060 hPa
Presión a nivel del mar (pnm)	Si elevación ≤ 800 m: 930 hPa ≤ pnm ≤ 1060 hPa; Si 800 m < elevación ≤ 2300 m: 1000 hPa ≤ pnm ≤ 1650 hPa; Si 2300 m < elevación ≤ 3700 m: 1600 hPa ≤ pnm ≤ 3200 hPa; Si elevación > 3700 m: 3200 hPa ≤ pnm ≤ 6300 hPa.
Precipitación	0 ≤ precipitación ≤ 300 mm
Humedad relativa	0 ≤ humedad relativa ≤ 100%
Heliofanía	0 ≤ heliofanía ≤ 18 horas
Nubosidad	0 ≤ nubosidad ≤ 9
Dirección de viento máximo diario	0 ≤ vmax.d ≤ 36 (direcciones en decenas de grado, redondeadas a valores enteros)
Velocidad de viento máximo diario	0 ≤ vmax.f ≤ 62 m s^{-1} (120 nudos)
Velocidad media del viento	0 ≤ vmed ≤ 26 m s^{-1} (50 nudos)

6. Familia de controles de rango variable

En esta familia de controles, los intervalos usados para detectar datos posiblemente erróneos son de rango variable. Es decir, para aceptar o rechazar datos, los umbrales varían dinámicamente (de allí el nombre de esta familia de controles), tomando valores específicos para cada día o mes del año. Al igual que en los controles de rango fijo, un valor se identifica como sospechoso si queda fuera del intervalo válido o aceptable definido para cada variable. En las secciones siguientes se discute en detalle los diferentes controles de esta familia.

6.1 Apartamientos respecto al ciclo estacional

Dado que muchas variables meteorológicas presentan un comportamiento regular o repetido estacionalmente, se realiza un control para cuantificar las diferencias entre los valores observados de una variable y un ciclo estacional ajustado a la serie de datos. El ajuste del ciclo estacional se realiza mediante un Modelo Aditivo Generalizado o GAM (Hastie y Tibshirani, 1990), utilizando una curva diferenciable definida por polinomios, que permite una representación flexible del ciclo estacional y no asume una forma funcional predeterminada (por ejemplo, una sinusoide).

Este control (RV01) identifica como sospechosos a aquellos valores que muestran desvíos sospechosamente diferentes respecto del ciclo estacional ajustado de una variable. Los desvíos o residuos extremos se definen en términos de percentiles estimados a partir de estos desvíos. Por ejemplo, si la variable muestra valores excepcionalmente bajos o altos se pueden utilizar los percentiles 1 y 99, respectivamente ($perc_{01}$ y $perc_{99}$) como límites. Los percentiles pueden estimarse (a) para cada día del año o (b) para cada mes. En la implementación actual, los percentiles de los desvíos se calculan para cada mes.

En la figura 8 se presenta un ejemplo específico para el año 1962 en Pehuajó. Los valores de temperatura máxima diaria (Tmax) de ese año en particular se representan con una línea roja y el ciclo estacional ajustado, con una curva negra. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un punto rojo, indicando que el control lo identifica como sospechoso. La figura 9 muestra la serie de diferencias entre valores observados y el ciclo estacional para la serie presentada en la figura 8. En los tonos grises se indican los distintos límites del rango inferior aceptable para cada mes según el percentil escogido (percentiles 10.1, y 0.1, indicados en la figura como 0.1, 0.01 y 0.001 respectivamente). El valor sospechoso (punto azul) cae claramente fuera (debajo) inclusive del rango inferior más estricto (percentil 0.001).

Figura 8. Temperatura máxima diaria observada en Pehuajó (línea roja) y ciclo estacional estimado (línea negra) para el año 1962. El valor de Tmax para el 22 de agosto de 1962 (2.9 °C) está resaltado con un punto rojo, indicando que el control lo identifica como sospechoso.

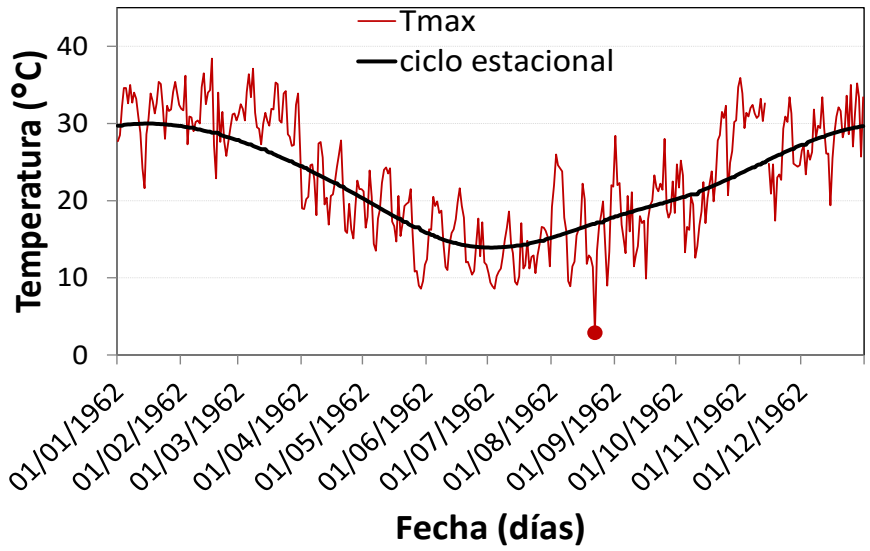
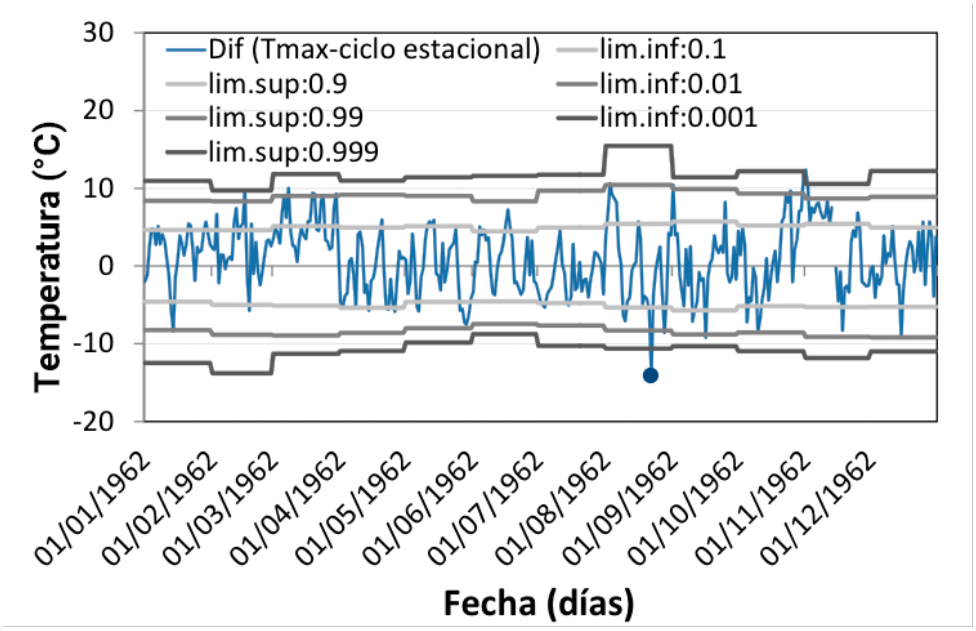
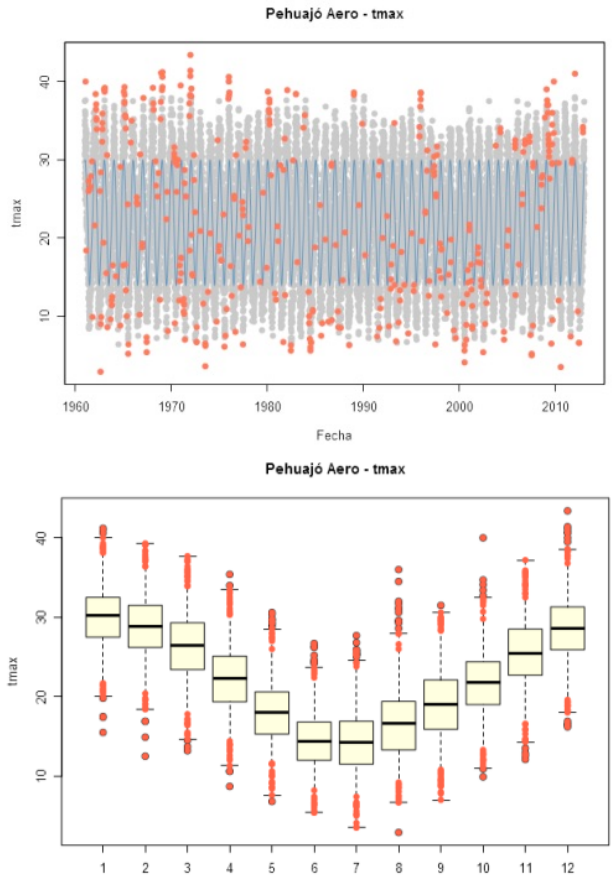


Figura 9. Diferencia entre la temperatura máxima diaria y el ciclo estacional en Pehuajó, año 1962. En tonos de grises se indican los distintos límites del rango aceptable para cada mes según el percentil escogido. El valor sospechoso (punto azul) cae fuera (debajo) del rango más estricto.



En el panel superior de la figura 10 se presentan todos los valores observados de Tmax en Pehuajó para 1961-2012 (puntos grises). El ciclo estacional ajustado se representa con una línea azul. Los puntos rojos señalan datos potencialmente erróneos, con desvíos extremadamente grandes (positivos y negativos) respecto al ciclo estacional. El panel inferior muestra *boxplots* (o diagramas de "cajas y bigotes") que describen la distribución estadística de los valores de Tmax diaria para cada mes del periodo 1961-2012. Los bordes inferiores y superiores de las cajas en color amarillo corresponden a los percentiles 25 y 75 de cada mes; es decir, las cajas contienen el 50 % central de los datos observados en cada mes. La línea gruesa horizontal dentro de cada caja indica la mediana de Tmax (el percentil 50). Los valores alrededor de los límites de las líneas verticales o "bigotes", por encima y debajo de cada caja son identificados como sospechosos.

Figura 10. Temperatura máxima diaria en Pehuajó, Argentina, 1961-2012. Panel superior: Ciclo estacional estimado (línea azul), datos (puntos grises) y datos sospechosos (puntos rojos). Panel inferior: *boxplots* mensuales (1961-2012) y datos sospechosos (puntos rojos).



6.2 Apartamientos respecto a múltiplos del rango intercuartil para ventanas de 3 o 5 días

Este control (RV02) también identifica desvíos sospechosamente extremos con respecto a un valor esperado. El valor esperado y su dispersión, se definen en este caso en base a dos parámetros estadísticos: la mediana (M) de los valores considerados y su pseudo desvío estándar (ssd). Tanto M como ssd se estiman para una ventana temporal centrada alrededor del día del año a analizar (día i) que incluye los valores de todos los años con observaciones disponibles para los días incluidos en la ventana. En este control se pueden utilizar dos opciones de ventanas centradas en el día i, la primera de 3 días (±1 día alrededor del día i) y otra de 5 días (± 2 días alrededor del día i). Por ejemplo, si consideramos el 5 de enero como día i, la ventana de 3 días contendrá los datos del 4 al 6 de enero de todos los años disponibles (por ejemplo, 1961-2018); en forma similar, la ventana de 5 días incluye los datos entre el 3 y el 7 de enero de todos los años disponibles.

El rango intercuartil (definido como la diferencia entre el percentil 75 y el percentil 25) es resistente a los valores extremos, por lo tanto, es un buen estadístico para utilizar en los controles de calidad que buscan identificar valores sospechosamente extremos. El pseudo desvío estándar ssd se calcula de la siguiente forma:

$$ssd = \frac{ri}{1.349}, \qquad (1)$$

donde ri indica el rango intercuartil. Para calcular ssd se divide ri por 1.349 ya que, para una distribución normal, el rango intercuartil es 1.349 veces el desvío estándar (Lanzante, 1996). Tanto la mediana como el rango intercuartil son parámetros conceptual y computacionalmente simples, y son recomendables para casos que no requieran alta eficiencia, por ejemplo, si hay una gran cantidad de datos disponibles o cuando el análisis es “altamente exploratorio.”

La naturaleza extrema de una observación se evalúa con el estadístico Z, que mide el apartamiento de los datos respecto de M en función del ssd. Es decir, el valor de *Z* para el día i del año j se calcula “estandarizando” los datos restándoles M y dividiéndolos por el ssd correspondientes al día i:

$$Z_{i,j} = \frac{|x_{i,j} - M_i|}{ssd_i}, \qquad (2)$$

donde $x_{i,j}$ es el valor de la variable analizada para un día de un año determinado, y M_i y ssd_i indican la mediana y el pseudo desvío estándar para ese mismo día, estimados para cada día del año i usando una ventana temporal de 5 días centrada alrededor de ese día. En el control de calidad, se identifican como sospechosos los datos cuyos valores de Z sean mayores que un cierto umbral. En este caso se usa un umbral de *Z* = 3. Más información sobre este control de calidad se puede encontrar en Lanzante (1996).

Para ilustrar el funcionamiento de este control, continuamos utilizando el ejemplo que se usó en la figura 9 (Tmax en Pehuajó, Argentina, año 1962). En este caso, la línea negra en la figura 11 indica la mediana de los valores de Tmax para cada día del año, estimada usando una ventana temporal

de 5 días y todos los valores para el periodo de referencia 1961-2102. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado nuevamente con un círculo rojo, indicando que también este control lo identifica como sospechoso (así como el RV01). El estadístico Z (figura 12) para ese día fue mayor que el límite de Z = 3. Cabe mencionar que la evaluación de los registros oficiales originales mostró que el valor correcto para este día fue de 12.9 °C en vez de 2.9 °C. En consecuencia, se confirma que el valor identificado por este control es efectivamente incorrecto.

Figura 11. Temperatura máxima diaria observada en Pehuajó (rojo) y mediana de valores para cada día del año (línea negra), año 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está indicado con un punto rojo, indicando que el control lo identifica como sospechoso.

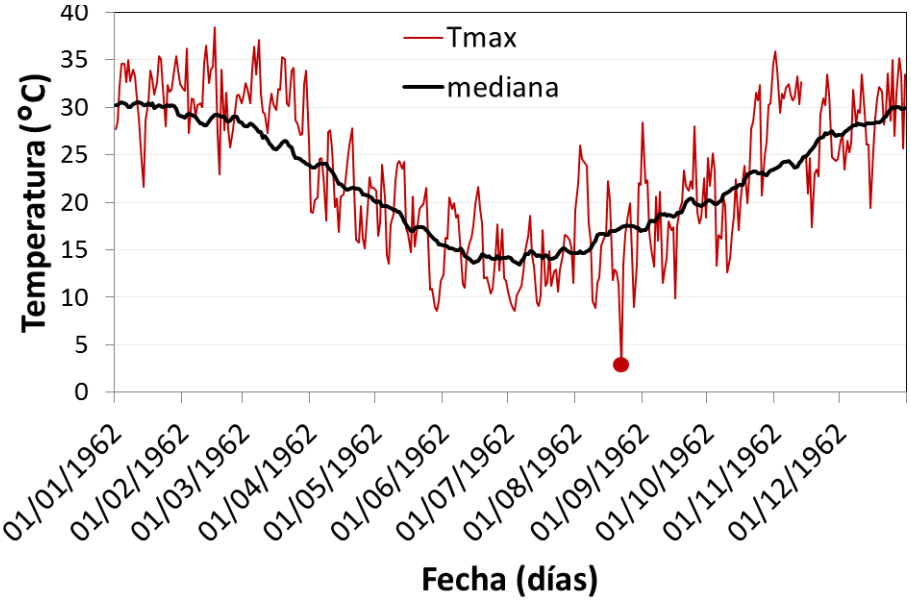
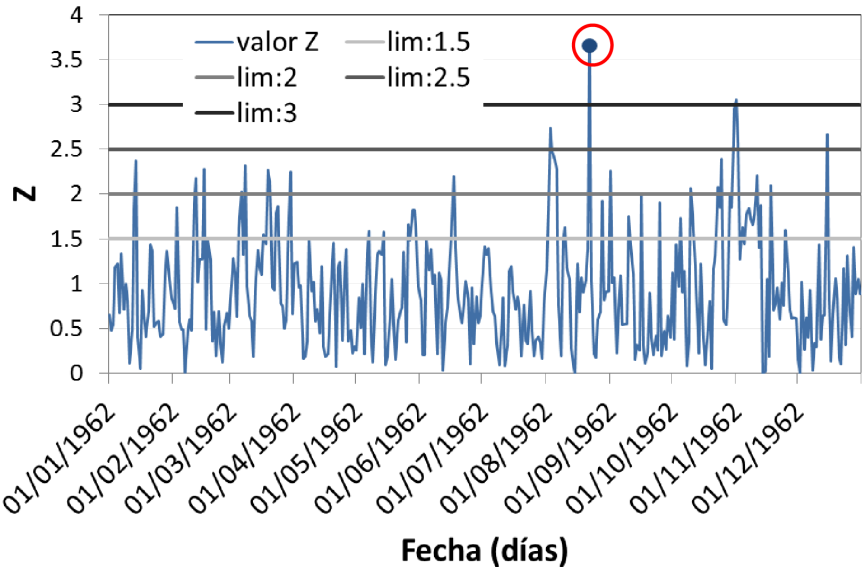


Figura 12. Estadístico Z calculado para cada día del año 1962 en Pehuajó. En tonos de grises se indican los distintos límites del rango aceptable. El valor sospechoso para el día 1962-08-22 (punto azul) cae fuera (encima) del rango más estricto (> z=3).



6.3 Apartamientos respecto a estadísticos resistentes (método *biweight*) para ventanas de 3 o 5 días

El cálculo de *scores* Z basado en la mediana y el rango intercuartil descrito en la sección 6.2 es una mejora con respecto al simple uso de medias y desvíos estándar para identificar valores extremos. Sin embargo, cuando se desea una resistencia mayor a la influencia de los valores extremos, se recomienda usar estimadores de tendencia central y dispersión más complejos, tales como la media y el desvío estándar calculados mediante la función *biweight* (Lanzante, 1996). Las estimaciones basadas en el método *biweight* involucran un cálculo en dos pasos. En el primer paso de este control (RV03) se estiman la tendencia central y la dispersión de los datos usando la mediana y el estadístico MAD (mediana de los desvíos absolutos). Estos estimadores se usan solamente para descartar valores extremos, a los que, en cálculos subsiguientes, se les asigna un peso igual a cero. En el segundo paso, se calculan la media y desvío estándar *biweight* ponderados: aquí los pesos utilizados disminuyen en forma no lineal con respecto al centro de la distribución de los datos (Lanzante, 1996). Los pesos $u_{i,j}$ para cada una de las observaciones $x_{i,j}$ para el día i del año j se calculan como

$$u_{i,j} = (x_{i,j} - M_i) / (\mathrm{c} \times MAD) \,, \qquad (3)$$

donde M_i y MAD_i son la mediana de los datos y la mediana de los desvíos absolutos estimados para el día i usando una ventana temporal de 5 días centrada alrededor de ese día y considerando todos los años en el registro histórico. El parámetro c define la distancia desde el centro de los datos a la que los pesos caen a 0. En estos controles se usa un valor de c = 7.5 que, en el caso

de una distribución normal, elimina valores mayores que ± 5 desvíos estándar (Hoaglin et al., 2000). Para $|u_{i,j}| \geq 1.0$, el peso se define como 1.0, de modo de eliminar la influencia de valores extremos.

A continuación, se calculan la media *biweight* $\bar{x}_i^{bwt}$ y el desvío estándar *biweight* $\bar{s}_i^{bwt}$ para cada día del año:

$$\bar{x}_i^{bwt} = M_i + \frac{\sum_{i=1}^{n}(x_{i,j} - M_i)(1-u_i^2)^2}{\sum_{i=1}^{n}(1-u_i^2)^2}, \text{y} \quad (4)$$

$$S_i^{bwt} = \frac{\sqrt{n\sum_{i=1}^{n}(x_{i,j} - M_i)^2(1-u_i^2)^4}}{|\sum_{i=1}^{n}(1-u_i^2)(1-5u_i^2)|}. \quad (5)$$

Los $\bar{x}_i^{bwt}$ y $\bar{s}_i^{bwt}$ están más influenciados por los valores cercanos al centro de la distribución que por los ubicados en las colas (en otras palabras, son más resistentes a la influencia de los valores extremos (Feng et al., 2004). Por esta razón, se los usa para calcular el estadístico Z mediante el cual se evalúa la calidad del dato:

$$Z_{i,j} = \frac{|x_{i,j} - \bar{x}_i^{bwt}|}{S_i^{bwt}}. \quad (6)$$

Como en el control anterior, se identifican como sospechosos los datos cuyos valores de Z son mayores que un cierto umbral. En este caso se usa un umbral de *Z* = 3. Este control fue utilizado en trabajos anteriores, como Feng et al. (2004) y Peterson et al. (1998), entre otros. Para ilustrar el funcionamiento de este control, continuamos utilizando el ejemplo de Tmax en Pehuajó. La figura 13 muestra con una línea negra los valores de la media *biweight*, estimada para cada día del año usando una ventana temporal de 5 días y considerando todos los valores en el período 1961-2102. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un punto rojo, indicando que este control lo identifica como sospechoso. El estadístico Z para ese día fue mayor que el límite de Z = 3 (figura 14). Este control de rango variable, al igual que los dos anteriores, logra identificar el valor erróneo.

Figura 13. Temperatura máxima diaria observada en Pehuajó (línea roja) y media *biweight* de valores para cada día de 1962 (línea negra). El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un punto rojo, indicando que el control lo identifica como sospechoso.

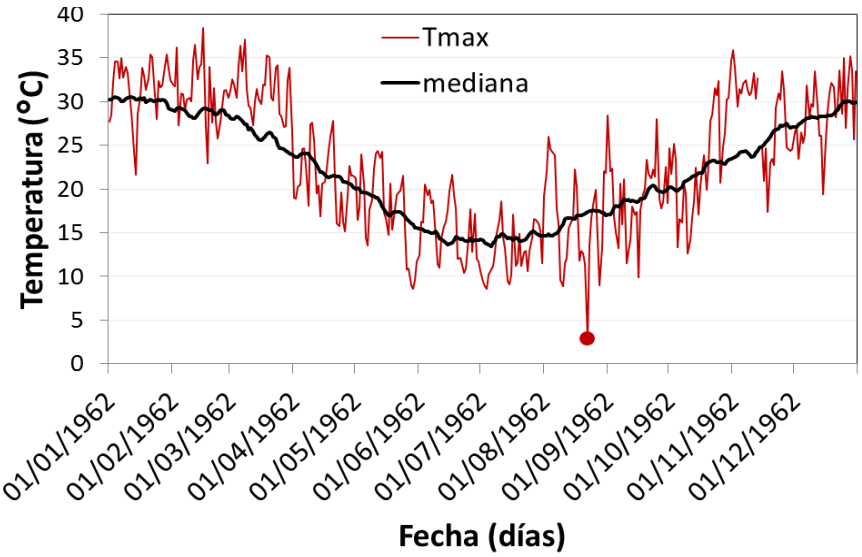
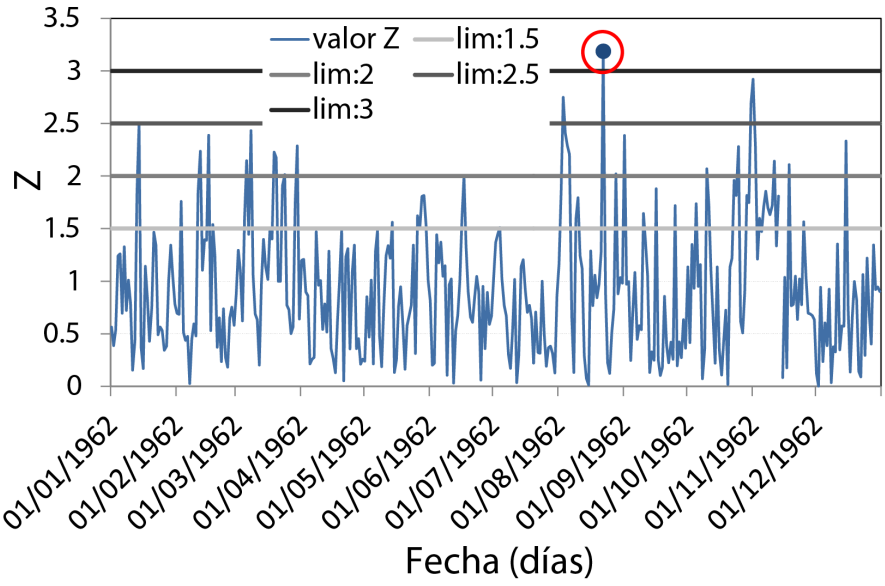


Figura 14. Estadístico Z calculado para cada día del año 1962 en Pehuajó. En tonos de grises se indican los distintos límites del rango aceptable. El valor sospechoso para el día 1962-08-22 cae fuera (encima) del rango más estricto (Z>3).

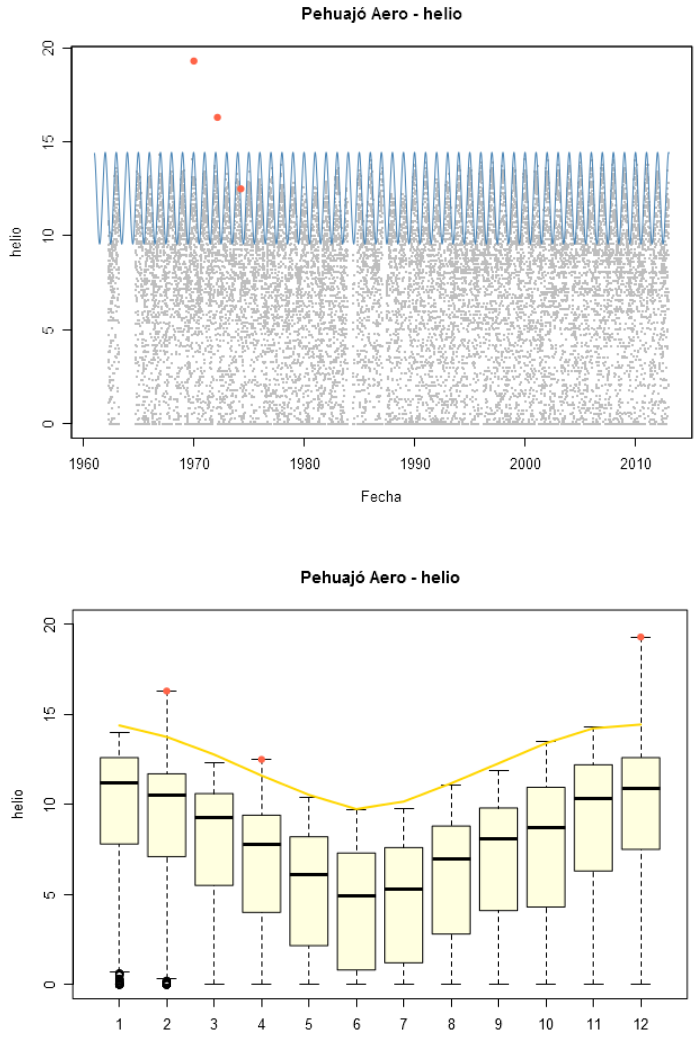


6.4 Relación entre la heliofanía medida y la heliofanía teórica astronómica

El máximo número de horas diarias de radiación solar –o heliofanía teórica astronómica– se puede calcular para una determinada latitud y día del año. Como este valor representa el máximo de la heliofanía observada en cada día, en este control (RV04) se identifican como sospechosos los registros que superen la heliofanía teórica astronómica calculada en cada estación (en base a su latitud) para cada día del año.

El cálculo de la heliofanía máxima para cada día del año puede realizarse con los paquetes de R `insol` (https://CRAN.R-project.org/package=insol) o `geosphere` (https://cran.r-project.org/package=geosphere). Estos paquetes utilizan, respectivamente, los métodos descriptos por Corripio (2003) y Forsythe et al. (1995). Para los controles descriptos aquí utilizamos el método de Corripio (2003) en el paquete `insol`. En función de la latitud de cada estación meteorológica se calculó la heliofanía máxima (expresada en horas) para cada día del año. Para evitar errores debido a redondeo de valores reportados, la heliofanía máxima calculada se multiplicó por 1.05 y este valor se usó como umbral en la identificación de valores sospechosos.

Figura 15. Heliofanía diaria en Pehuajó 1961-2012. Panel superior: Heliofanía máxima teórica para cada día del año (línea azul), valores de heliofanía reportados (puntos grises) y datos sospechosos (puntos rojos). Panel inferior: *Boxplots* mensuales de heliofanía (1961-2012) y datos sospechosos (puntos rojos). La línea amarilla representa la heliofanía máxima promedio para cada mes.



La figura 15 ejemplifica el control de los datos de heliofanía medidos en Pehuajó, Argentina, en 1961-2012. Los tres puntos marcados en rojo exceden el umbral definido en el párrafo anterior, y por lo tanto son considerados sospechosos. Dos de ellos (los desvíos más grandes) eran erróneos y se los pudo corregir con el dato que figuraba en la libreta meteorológica.

6.5 Apartamientos respecto al rango intercuartil de precipitación para ventanas mensuales

El diseño de controles de calidad para datos diarios de precipitación es una tarea difícil (Hubbard et ál., 2012). Este control (RV05) es análogo al descrito en la sección 5.2, pero se utiliza únicamente para valores de precipitación diaria. Dada la existencia de muchos días sin precipitación y la gran variabilidad entre meses en los montos diarios de lluvia, la estimación de la dispersión (el rango intercuartil) de los valores de lluvia se realiza con ventanas temporales más anchas (de un mes), en lugar de usar unos pocos días (3-5) como se hace con la temperatura.

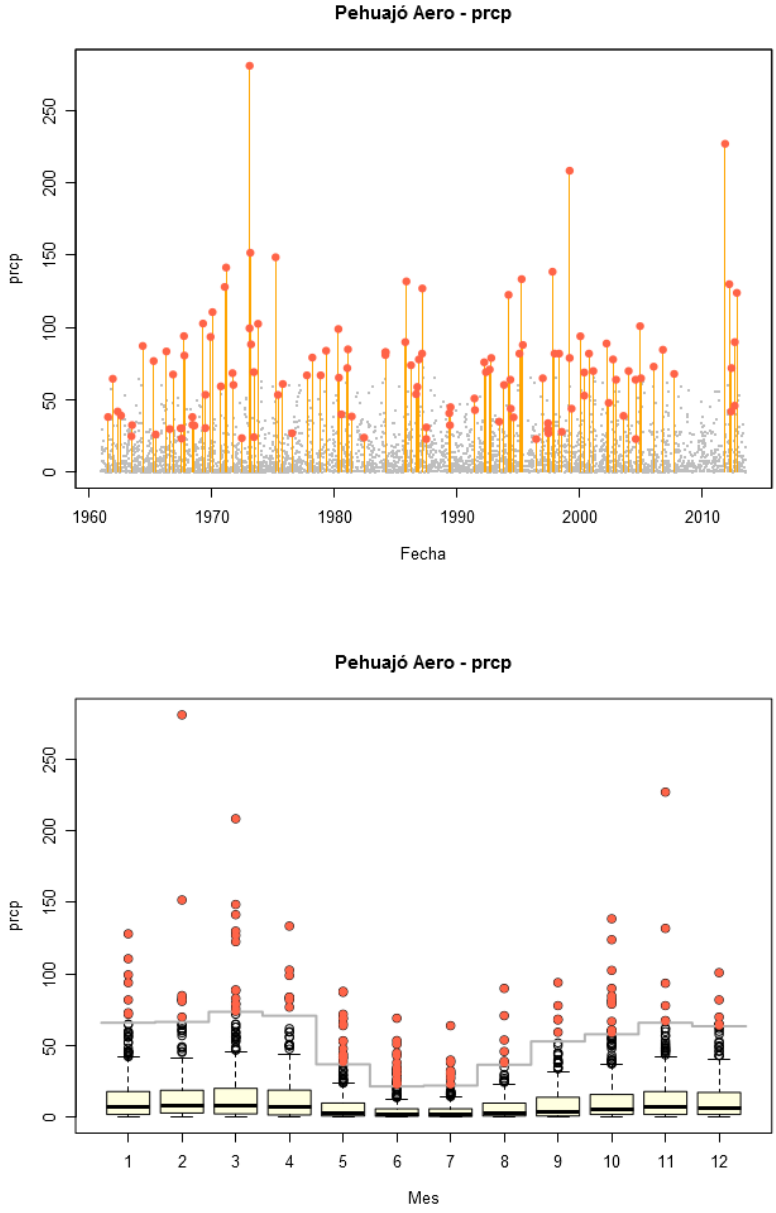
En este control, se identifican como sospechosos los valores diarios de precipitación que exceden un umbral PS_i estimado separadamente para cada mes i, que se calcula de la siguiente forma:

$$PS_i = p75_i + (n * ri_i)\,, \qquad (7)$$

donde $p75_i$ es el percentil 75 (es decir, el tercer cuartil) de los valores diarios de precipitación ≥ 0.1 mm (el umbral para la definición de un día de lluvia) en el mes i (estimado usando todos los años disponibles para ese mes en el registro histórico, por ejemplo, todos los febreros de 1961 a 2012), ri_i es el rango intercuartil para ese mes (estimado de la misma manera que el percentil 75) y n es un factor que multiplica a ri_i para definir cuántos mm de lluvia por encima del percentil 75 se consideran sospechosos.

Este control etiqueta como sospechosos todos los valores de lluvia diarios que superan el umbral PS para el mes correspondiente (ver ejemplo en la figura 16). Este tipo de control fue utilizado previamente por González-Rouco et ál. (2001) y por Aguilar y Prudhom en el software CLIMDEXEXTRAQC (documento disponible en http://www.c3.urv.cat/data/manual/Manual_rclimdex_extraQC.r.pdf).

Figura 16. Precipitaciones diarias en Pehuajó, Argentina en 1961-2012. Panel superior: Lluvias diarias (puntos grises) y datos sospechosos (puntos rojos con líneas naranja). Panel inferior: *boxplots* mensuales de precipitación diaria (1961-2012) y datos sospechosos (puntos rojos). La línea horizontal gris indica el umbral utilizado en cada mes para definir valores sospechosos. En ambos paneles se utilizan los valores diarios de precipitación ≥ 0.1 mm (el umbral que indica un día lluvioso).



6.6 Identificación de valores extremos mensuales de precipitación mediante ajuste de una distribución gamma

Como en la sección 6.5, este control (RV06) identifica como sospechosos los valores diarios de precipitación que exceden un cierto umbral. Sin embargo, el umbral no se define mediante

percentiles empíricos –como en el control anterior– sino en base a percentiles calculados usando un ajuste de una distribución de probabilidad teórica a los valores diarios de lluvia. Una de las distribuciones estadísticas que pueden utilizarse para representar datos de precipitación es la distribución gamma. Si bien puede haber otras distribuciones que presenten un mejor ajuste a los datos de precipitación, el objetivo de este control es establecer un umbral confiable a partir del cual se pueda identificar mediciones sospechosas de precipitación diaria.

En este control se agrupan los datos históricos para cada uno de los meses del año. Para cada mes, se estiman los parámetros α (forma) y β (escala) de una distribución gamma. Primero, se estiman valores iniciales de los parámetros de una distribución gamma usando el método de L-momentos (Vicente-Serrano, 2006). Con esos valores iniciales, luego se ajustan los parámetros finales usando el método de máxima verosimilitud. Para este paso se usa el paquete `fitdistr` de R (Delignette-Muller & Dutang, 2015). En los ajustes solo se incluyen precipitaciones diarias ≥ 0.1 mm, es decir que se descartan los días sin precipitación o con precipitaciones imperceptibles en el registro histórico.

Antes de avanzar con el control, se verifica que la distribución gamma ajuste correctamente los valores de precipitación usados. Para este fin se utiliza el paquete `KScorrect` de R (Novack-Gottshall & Wang, 2019). El paquete `KScorrect` implementa la corrección de Lilliefors al test de bondad de ajuste de Kolgomorov-Smirnov. Esta corrección es necesaria cuando los parámetros de la distribución son estimados a partir de la muestra (en este caso, los valores de precipitación). En este caso, los valores críticos P son estimados en base a simulaciones: aquí se utiliza un P de 0.1. Si la distribución gamma no ajusta bien los valores utilizados, no se puede realizar el control de precipitación (RV06).

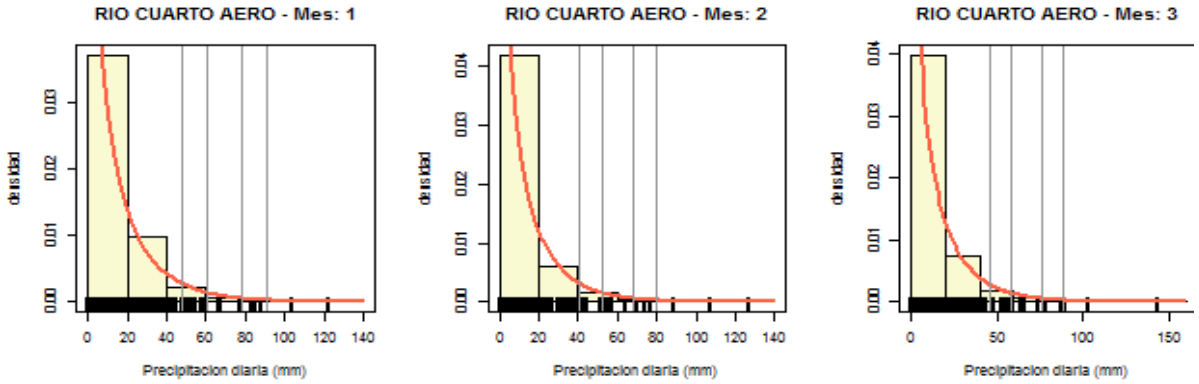
Luego de ajustar los parámetros de la distribución gamma y verificar su ajuste a los datos, se define el umbral a partir del cual las precipitaciones diarias se considerarán extremadamente altas y, en consecuencia, sospechosas. A diferencia de controles anteriores, el umbral puede definirse en base a un percentil calculado a partir de la distribución teórica (es decir, usando los parámetros estimados anteriormente). Específicamente, se identifican valores sospechosos de lluvia mediante

$$\text{Prcp}_{i,j} > perc_i^P \ , \qquad (8)$$

donde $prcp_{i,j}$ es la precipitación en el día i del año j y $perc_i^P$ es el valor de lluvia correspondiente a un percentil extremo P (donde P puede ser, por ejemplo, 0.975, 0.99, 0.995, etc.) para el mes en análisis. Este control fue previamente utilizado por Hubbard et al. (2012).

Un ejemplo se presenta en la figura 17, donde se muestran histogramas de precipitación diaria acumulada para enero, febrero y marzo durante 1961-2013 en Río Cuarto, provincia de Córdoba, Argentina. Las líneas cortas verticales perpendiculares al eje x de cada panel corresponden a los valores observados de lluvias diarias ≥ 0.1 mm. Las líneas verticales grises en cada panel indican posibles umbrales para la definición de precipitaciones extremas; de izquierda a derecha corresponden a los percentiles 0.950, 0.975, 0.990 y 0.995 respectivamente. Por ejemplo, para el mes de febrero se pueden definir como extremas las precipitaciones > 80 mm (es decir, el percentil 0.995). Al menos tres valores diarios de lluvia en febrero caen a la derecha de este percentil; estos días podrían considerarse sospechosos.

Figura 17. Histogramas de precipitaciones diarias en estación meteorológica Río Cuarto, Córdoba, Argentina. De izquierda a derecha, los histogramas corresponden a los meses de enero, febrero y marzo, respectivamente. La línea naranja en cada panel indica la distribución gamma ajustada a los valores de cada mes. Las líneas verticales grises corresponden a los percentiles 0.950, 0.975, 0.990 y 0.995 respectivamente.



6.7 Apartamientos respecto a medias y desvíos estándar resistentes (método *biweight*) para la amplitud térmica diaria

Este control es muy similar al expuesto en la sección 6.3. De hecho, el método de cálculo de los apartamientos de las medias y desvíos estándar resistentes mediante el método *biweight* es idéntico en ambos controles. Sin embargo, aquí no se analiza una variable individual, sino la relación entre dos de ellas, concretamente, la relación entre la temperatura máxima y la temperatura mínima de cada día. Es decir, este control se aplica a la amplitud térmica diaria, calculada como la diferencia entre la temperatura máxima y la temperatura mínima registradas en un día determinado. Este control fue sugerido por el Instituto de Clima y Agua del Instituto Nacional de Investigación Agropecuaria (INTA) de Argentina.

7. Familia de controles de continuidad temporal

La continuidad temporal de los valores diarios de las variables meteorológicas es un aspecto importante a la hora de analizar la consistencia de los datos climáticos. Con esta familia de controles se detectan secuencias de valores iguales a lo largo de varios días consecutivos, y también saltos o discontinuidades en las series analizadas.

7.1 Persistencia de valores constantes durante varios días consecutivos

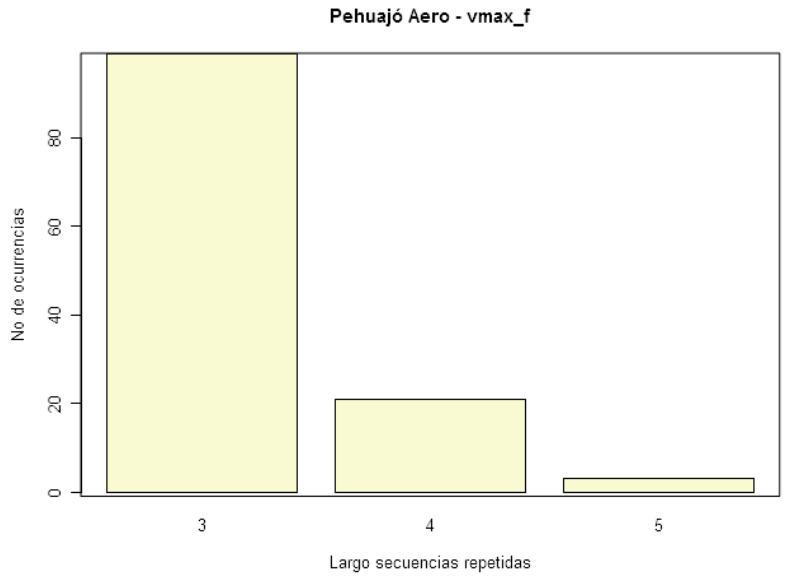
La persistencia de un mismo valor puede sugerir errores de transcripción o problemas de instrumental electrónico, por ejemplo, en estaciones meteorológicas automáticas –no consideradas en esta etapa (Estévez et al., 2015). Es posible que, en la realidad, una variable tenga un valor constante durante varios días consecutivos, pero la secuencia debe ser verificada para determinar su validez. Este tipo de controles fue utilizado por Meek & Hatfield (1994), Durre et al. (2010), Estévez et al. (2011), Estévez et al. (2015) y por Aguilar y Prudhom en el software CLIMDEX EXTRAQC (ver http://www.c3.urv.cat/data/manual/Manual_rclimdex_extraQC.r.pdf).

Este control (CT01) identifica secuencias de 3 o más días con valores idénticos de una determinada variable y se aplica a todas las variables de la base de datos del CRC-SAS. Como ejemplo de este tipo de controles, la figura 18 muestra la frecuencia de ocurrencias de valores idénticos consecutivos para la serie de velocidad de viento máximo diario en Pehuajó, Buenos Aires, Argentina.

Un caso especial de este control es la precipitación diaria. Para esta variable, los valores correspondientes a días lluviosos (días con precipitación mayor o igual al umbral definido en la configuración, generalmente 0.1 mm) se consideran de la misma forma que las otras variables. Para todas las secuencias con valores repetidos más de 3 días consecutivos, se marcan todos los valores en la secuencia como sospechosos y se controlan en los registros oficiales.

Para los días no lluviosos, en cambio, no se realiza este control, ya que puede haber secuencias muy largas de días sin lluvia, por ejemplo, en zonas áridas o semiáridas. Para los días secos, el valor de este control no está definido (es igual a NULL).

Figura 18. Número de ocurrencias de secuencias con valores de velocidad de viento máximo diario repetidos durante 3, 4 y 5 días en Pehuajó, Argentina. En los datos 1961-2012 no hay secuencias repetidas más largas (> 5 d) que las incluidas en la figura.

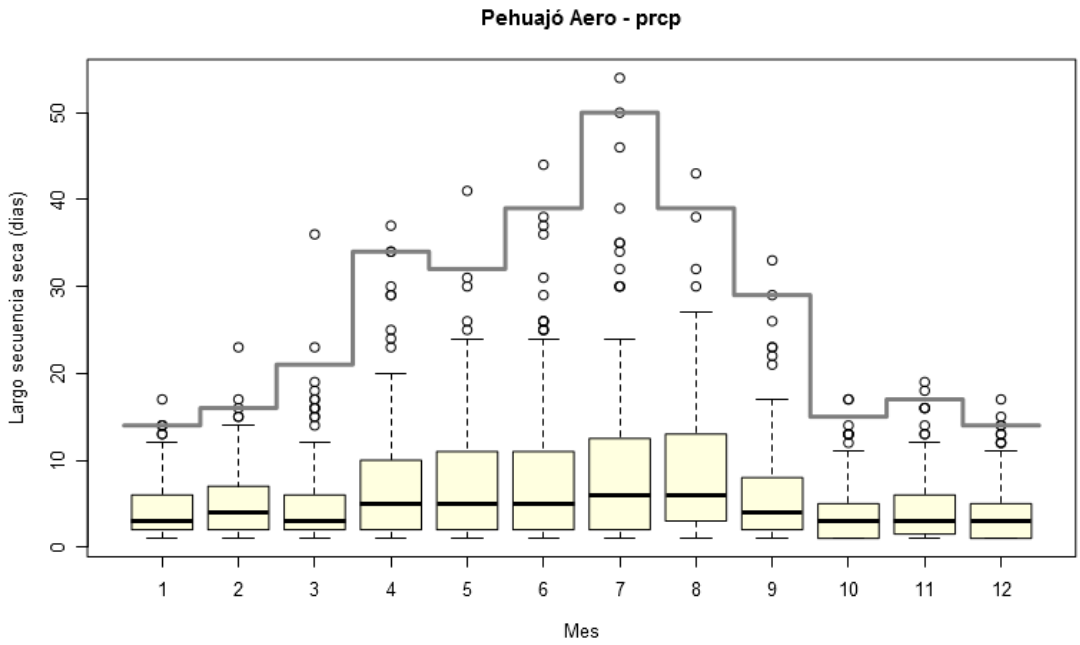


7.2 Persistencia extrema de días sin precipitación

El control descrito en la sección 7.1 excluye los valores de precipitación menores a un umbral parametrizable que define un día lluvioso (ej., 0.1 mm), ya que la persistencia de días sin lluvia no es necesariamente una secuencia incorrecta. El propósito de este control para precipitación (CT02) es detectar un error común al migrar o transcribir datos, que consiste en no incorporar datos existentes de precipitación y registrar, en cambio, un valor de cero. Otro error frecuente es reemplazar observaciones de precipitación faltantes –y que deberían identificarse como tales– por un valor de 0 mm.

Para identificar las secuencias sospechosas de días sin lluvia debe definirse un umbral a partir del cual una secuencia puede considerarse extrema y, por tanto, sospechosa. El umbral puede definirse en base a un cierto percentil (por ejemplo, 0.999, este valor es configurable) estimado a partir de la extensión observada de secuencias secas. Para poder acomodar posibles diferencias en la estacionalidad de las precipitaciones, en este control se estiman umbrales de secuencias secas extremas para cada mes del año. Usando todos los datos en el registro histórico, se calcula el largo de todas las secuencias secas que comiencen en un mes determinado. En base a esas extensiones, se estima el percentil usado como umbral. Todos los días en secuencias de días secos que excedan el umbral determinado se marcan como posibles sospechosos. Este control fue sugerido por Aizpuru & Leggieri (2008) y utilizado por Boulanger et al. (2010).

Figura 19. *Boxplots* de la distribución de extensión de secuencias de días sin precipitación en Pehuajó, Argentina, en el periodo 1961-2012. La línea gris escalonada indica los umbrales para identificación de largos extremos para cada mes (en este caso, el percentil 0.995). Durante los meses con mayor precipitación (por ejemplo, diciembre, enero), el umbral está alrededor de los 14-16 días, mientras que, en los meses más secos, las secuencias secas deben superar los 40 días para ser consideradas extremas.



Por ejemplo, en la figura 19 se observa la distribución mensual de la extensión de secuencias de días sin precipitación en Pehuajó, Argentina, en el periodo 1961-2012. Puede verse claramente que el umbral para identificación de extensiones extremas varía a lo largo del año. Aunque la estacionalidad de la precipitación en Pehuajó no es de las más marcadas, durante los meses más lluviosos (por ejemplo, diciembre, enero) el umbral está alrededor de los 14-16 días. En cambio, en los meses más secos, durante el invierno, el umbral debe establecerse en alrededor de 40 días.

7.3 Saltos excesivamente grandes entre días consecutivos

Este control (CT03) apunta a encontrar valores inusualmente altos o bajos de temperaturas (mínima, máxima, media y de rocío), humedad relativa y presión atmosférica (a nivel de la estación y a nivel del mar) respecto al registro del día anterior. El primer paso es la creación de una serie temporal de diferencias absolutas Δx de la variable analizada entre un día y el día inmediatamente anterior:

$$\Delta x = |x_i - x_{i-1}|\,, \qquad (9)$$

donde se usa la notación genérica x para todas las variables analizadas.

Para identificar saltos sospechosos, debe definirse un umbral a partir del cual una diferencia de valores entre días consecutivos puede considerarse extrema y, por tanto, potencialmente errónea. El umbral puede definirse en base a un percentil empírico $perc^x$ (por ejemplo, el percentil 0.995; este valor es configurable para cada variable analizada por este control) que se estima a partir de la distribución histórica de valores absolutos de diferencias observadas. Específicamente, se identifican diferencias extremas de temperatura mediante

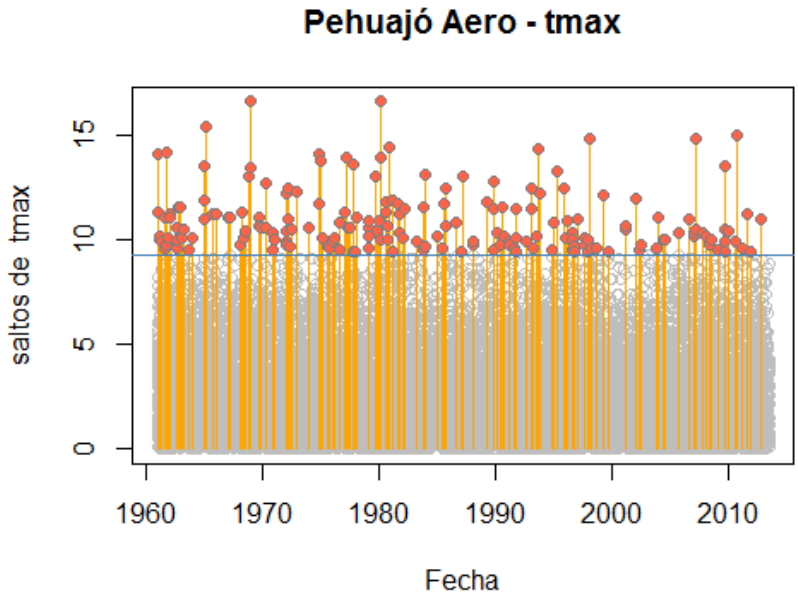
$$\Delta x > perc^x \ . \qquad (10)$$

La implementación del control se ilustra en la figura 20, que muestra el valor absoluto de diferencias en la temperatura máxima entre dos días consecutivos para el registro histórico en Pehuajó (1961-2012). La línea horizontal de la figura indica el valor del percentil 0.95, que se usa para separar los saltos de magnitud sospechosos. La figura 21 y la figura 22 ilustran el uso de este control para la identificación del valor erróneo del 22 de agosto de 1982. Otro ejemplo de la implementación de este control se puede encontrar en Kunkel et al. (1998).

7.4 Picos extremos I

Este control (CT04) se enfoca en la identificación de picos extremos en las series de temperaturas (mínima, máxima, media y de rocío), humedad relativa y presión atmosférica (a nivel de la estación

Figura 20. Diferencias absolutas de temperatura máxima en dos días consecutivos en Pehuajó, Argentina, en 1961-2012. La línea horizontal azul indica el percentil 0.99, utilizado para la identificación de saltos extremos. Los puntos rojos unidos al origen del eje x por líneas anaranjadas indican saltos extremos.



y a nivel del mar) respecto al registro del día anterior. Se define un “pico” como un valor de la variable mayor o menor a los valores de los dos días circundantes (el día anterior y el siguiente).

El primer paso en el control es la creación de dos series temporales de diferencias de variables entre un día i y los días inmediatamente anterior $(i-1)$ y posterior $(i+1)$:

$$\Delta x_1 = \left| x_i - x_{i-1} \right|, \text{ y}$$

$$\Delta x_2 = \left| x_{i+1} - x_i \right|,$$

donde se usa la notación genérica x para indicar cada una de las variables analizadas.

Figura 21. Panel superior: Temperatura máxima (Tmax) diaria observada en Pehuajó (línea roja) para cada día de 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un punto rojo, indicando que el control lo identifica como sospechoso. La escala de temperaturas debe leerse en el margen izquierdo de la figura. Panel inferior: diferencia absoluta entre $Tmax_i$ y $Tmax_{i-1}$; la escala de diferencias debe leerse en el margen derecho de la figura. Puede verse que la diferencia de temperatura entre el 22 y 21 de agosto de 1962 (punto negro) es extrema y coincide con el valor aparentemente bajo de Tmax en el panel superior.

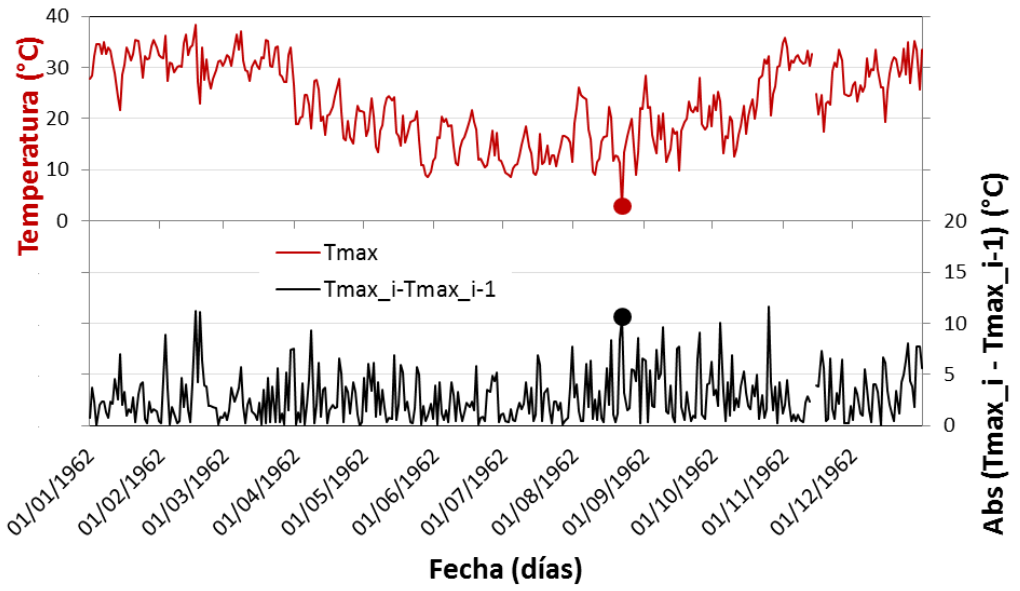


Figura 22. Valor absoluto de las diferencias entre Tmax en dos días consecutivos en Pehuajó para cada día de 1962. Las líneas horizontales indican diferentes umbrales posibles para la identificación de diferencias extremas basados en diferentes percentiles (0.900, 0.990 y 0.999). La diferencia de temperatura entre el 22 y 21 de agosto de 1962 excede el umbral definido por el percentil 0.990; el salto se debe al valor erróneo registrado el 22 de agosto.

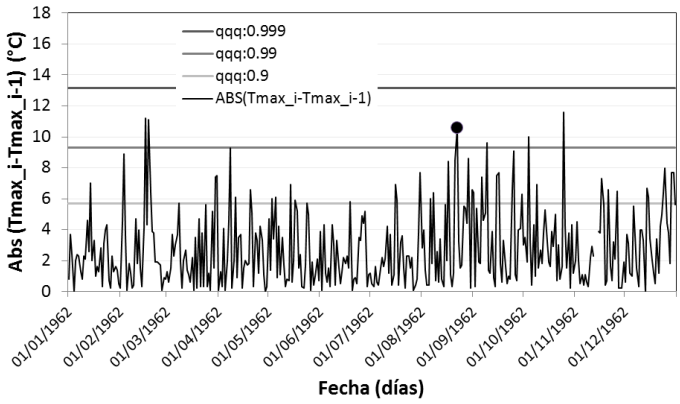
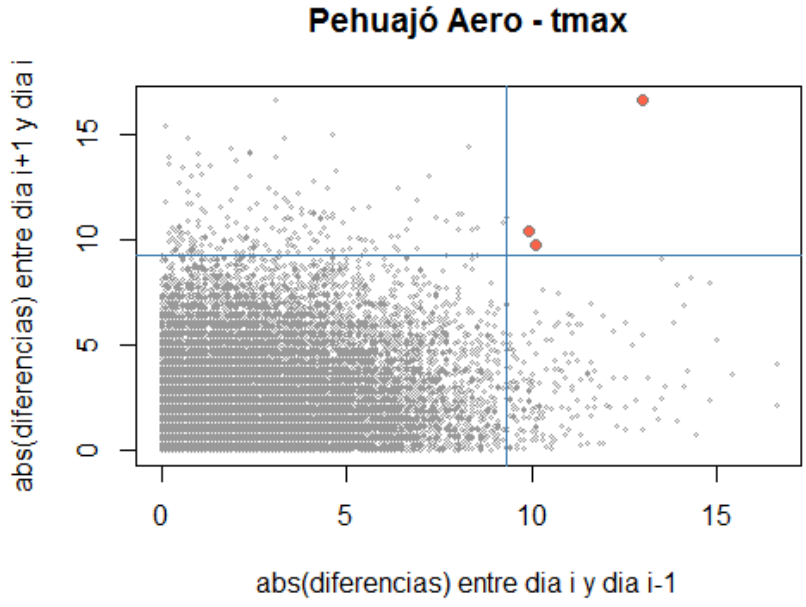


Figura 23. Diferencias absolutas de temperatura máxima entre un día y los días anterior (eje x) y siguiente (eje y) para Pehuajó, 1961-2012. En cada eje se indica el valor del umbral que define diferencias extremas (en este caso, el percentil 0.99). Los puntos para los cuales las dos diferencias de temperatura exceden los dos umbrales (marcados en rojo en el cuadrante superior derecho) se consideran "picos" extremos.



Para identificar picos sospechosos debe definirse un umbral a partir del cual una diferencia de valores entre un día y los valores circundantes puede considerarse extrema y, por tanto, potencialmente errónea. El umbral se define en forma similar a la usada en el control de "saltos" descrito en la sección 7.3. O sea, se calcula el valor absoluto de las diferencias entre un día y el día inmediatamente anterior

(ver ecuación 9). Ya que se considera el valor absoluto de las diferencias, es lo mismo calcular todas las diferencias de esta manera en lugar de hacer dos cálculos (diferencias con el día anterior y el siguiente). El umbral se define en base a un percentil empírico $perc^x$ (por ejemplo, el percentil 0.95; este valor es configurable para cada variable) que se estima a partir de la distribución histórica de valores absolutos de diferencias observadas.

Aquellos valores para los cuales ambas diferencias de temperatura Δx_1 y Δx_2 son mayores que el umbral, se identifican como sospechosos.

La implementación del control se ilustra en la figura 23, que muestra, a lo largo del eje x, el valor absoluto de diferencias en la temperatura máxima entre los días i e $i-1$ para el registro histórico en Pehuajó (1961-2012). En el eje y se indica el valor absoluto de diferencias en la temperatura máxima entre los días i e $i+1$. En cada eje se indica el valor del umbral (línea azul) que define diferencias extremas (en este caso, el percentil 0.99). Los puntos para los cuales las dos diferencias de temperatura exceden el umbral (marcados en rojo en el cuadrante superior derecho) se consideran “picos” extremos. Otro ejemplo de la implementación de este control se puede encontrar en Kunkel et al. (1998).

7.5 Picos extremos II

Este control –como el anterior– también apunta a identificar picos extremos (*bleeps* o *spikes*) de un día de duración en las series de los valores de una variable determinada (se usa para Tmax, Tmin, Tmed Trocío, HR, presión en estación y presión a nivel del mar). Como en la sección 7.4, para datos diarios se define un “pico” como un valor diferente (mucho mayor o menor) a los valores de los dos días circundantes (el día anterior y el siguiente). Este control también se puede utilizar para series con frecuencias de observación más altas, como datos sub-diarios observados por estaciones meteorológicas automáticas. Estas series frecuentemente presentan los picos breves (positivos o negativos) que se intentan detectar en este control (Fiebrich et al., 2010).

El control CT05 involucra dos etapas. En la primera etapa se determina si x_i –el valor de la variable para el día i– es un pico. Por ejemplo, si $x_{i-1} < x_i > x_{i+1}$, el valor x_i es un pico positivo. A la inversa, si $x_{i-1} > x_i < x_{i+1}$, el valor x_i es un pico negativo. En el caso de que el valor x_i sea faltante, no se puede determinar si es un pico, por lo que el resultado final del test es nulo (valor NA). Obviamente, tampoco se podrá definir si x_{i-1} y x_{i+1} son picos, por lo que el resultado del control para esos días también será nulo.

En la segunda etapa del control se determina si un pico detectado en la primera etapa se considera sospechoso. Para identificar picos sospechosos, el control utiliza ventanas temporales móviles de ancho impar (de modo de tener igual número de valores antes y después del día central de la ventana); actualmente se utiliza una ventana de 7 días, pero este ancho de ventana es configurable. Usando todos los valores diarios dentro de la ventana –salvo el valor para el día central x_i– se calculan la mediana (M) y la dispersión (MAD) de la variable analizada. Estos estimadores resistentes de tendencia central —M y MAD— se utilizan para definir umbrales inferior

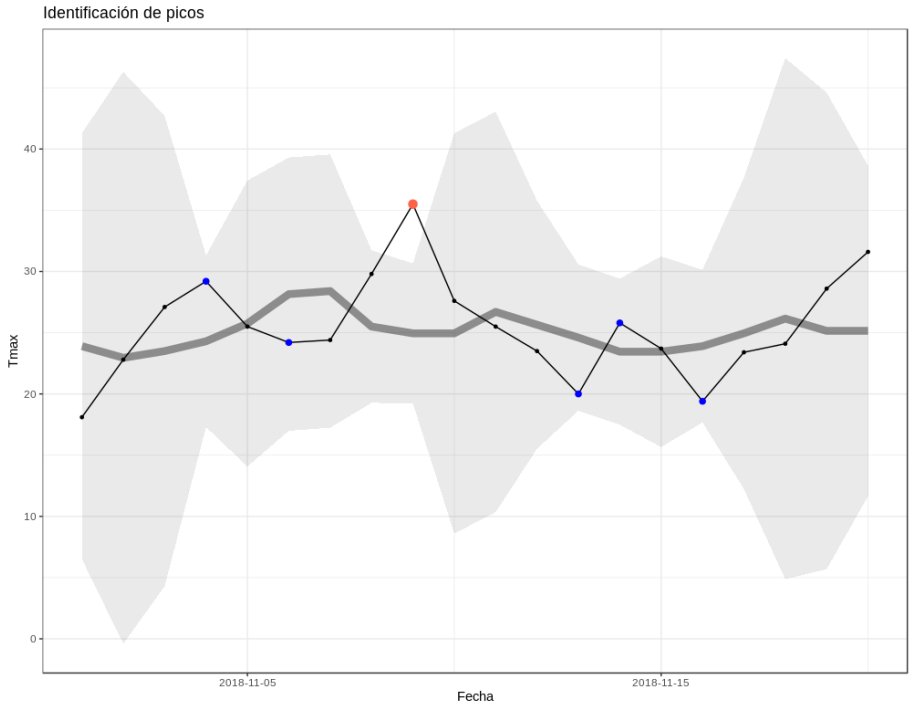
y superior más allá de los cuales una diferencia de valores entre el día central y los dos valores circundantes puede considerarse extrema y, por tanto, potencialmente errónea. Primero, el MAD de todos los valores en la ventana se multiplica por un factor definido en el archivo de configuración para cada variable analizada. Este producto se suma (o se resta) a la M de valores en la ventana para definir los umbrales superior e inferior más allá de los cuales un pico se considera sospechoso. En el control se utiliza 5.0 como factor que multiplica al MAD (si este factor es menor, por ejemplo 4.0, habrá más valores identificados como sospechosos). Como la función `mad()` en R (que se utiliza para calcular esta métrica) devuelve un valor escalado para coincidir con la desviación estándar de una distribución normal, el producto puede interpretarse como el número de desvíos aceptables con respecto a la diferencia absoluta entre el valor x_i y la mediana de la ventana.

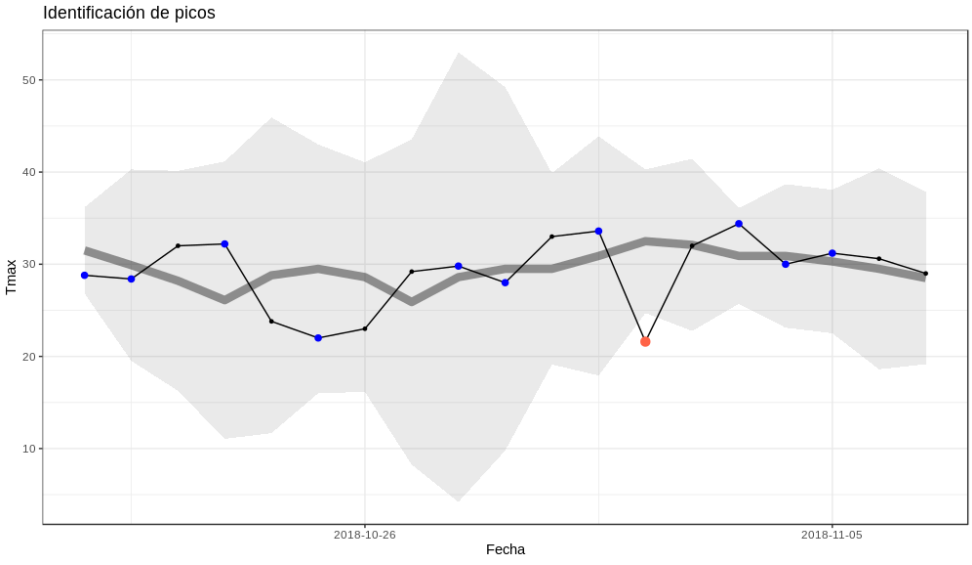
En algunos casos ocurre que la dispersión de valores en una ventana temporal es muy baja y por lo tanto el valor del umbral calculado también es muy bajo –posiblemente generando muchas falsas alarmas. En estos casos, el control especifica un umbral mínimo para cada variable en el archivo de configuración (este control usa el umbral mínimo 4.0, pero este valor es configurable). El umbral finalmente utilizado por este control surge de la comparación entre el umbral calculado y el umbral mínimo: se selecciona el valor mayor de estas dos cantidades.

En suma, para ser identificado como un pico extremo, el valor de una variable en una estación meteorológica para un día determinado debe cumplir dos condiciones: (i) el valor debe ser un pico (positivo o negativo) y (ii) el valor absoluto de la diferencia entre la variable analizada para el día y la mediana de los valores en la ventana temporal debe ser mayor que (a) el umbral calculado (MAD * factor) o (b) el umbral mínimo, el que sea mayor de ambas cantidades.

La figura 24 muestra ejemplos de la detección de picos extremos en las estaciones meteorológicas Pehuajó, Argentina (panel A) y Londrina, Brasil (panel B).

Figura 24. Ejemplos de la detección de picos extremos en las estaciones meteorológicas Pehuajó, Argentina (panel A) y Londrina, Brasil (panel B). La línea fina negra indica la serie temporal de temperatura máxima diaria para una parte del registro en cada estación. La línea gris más gruesa representa la mediana de los valores dentro de ventanas móviles de 7 días. El área gris representa la zona entre el umbral superior e inferior de valores sospechosos (ver descripción del cálculo en el texto). Los días indicados con puntos azules indican picos positivos o negativos de cualquier magnitud. Los puntos rojos son picos de una magnitud sospechosa (se extienden más allá de la zona gris).





7.6 Valores sospechosos en una serie temporal desestacionalizada y con tendencia de baja frecuencia eliminada

Este control (CT06) identifica valores sospechosos en series a las cuales se les ha removido tanto una estimación de variaciones estacionales o cíclicas como una tendencia de baja frecuencia. El control utiliza los paquetes `stlplus` y `anomalize` del lenguaje R.

El primer paso en el control es la descomposición de la serie temporal de la variable de interés. La serie observada se descompone en tres componentes aditivos: (i) un componente cíclico o estacional, (ii) una tendencia de baja frecuencia y (iii) un remanente o residuo, que es la serie que se utiliza para detectar valores sospechosos. La suma de los tres componentes produce la serie observada original. El control utiliza el método STL (el acrónimo de *Seasonal and Trend decomposition using Loess*) para la descomposición estacional (Cleveland et al., 1990). Una ventaja del método STL es que la variabilidad cíclica puede incluir varias escalas temporales (por ejemplo, variaciones semanales o un ciclo anual). Además, la estimación mediante el método STL es resistente a la presencia de valores inusuales o extremos.

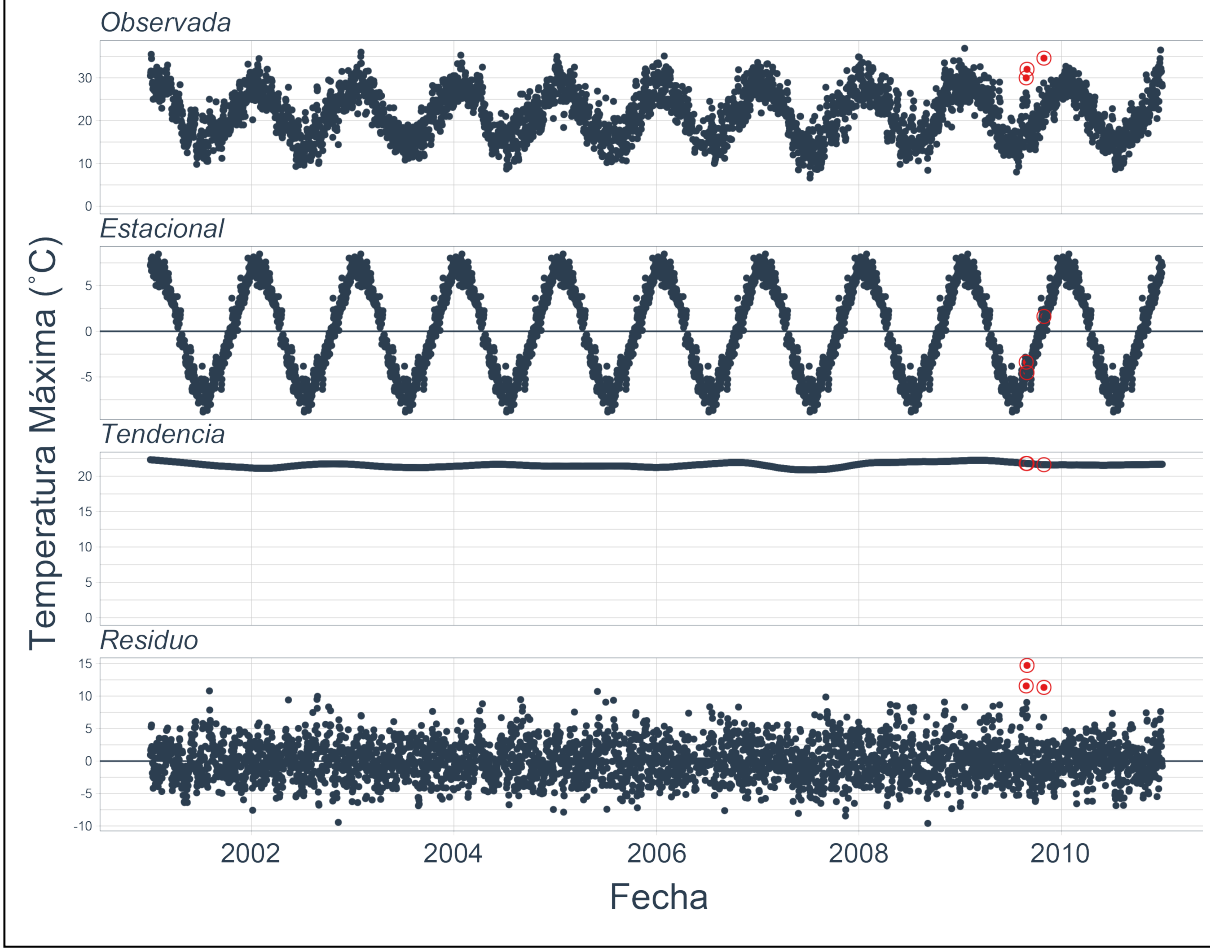
En este control, la descomposición estacional se realiza mediante el paquete `stlplus` de R (Hafen, 2016). Este paquete permite un mejor control de la descomposición estacional y además permite faltantes en la serie a descomponer. Alternativamente, se puede utilizar la descomposición incluida como parte del paquete `anomalize`, que permite utilizar varios métodos para obtener una serie remanente. Por ejemplo, el paquete puede utilizar el método STL original, implementado en la función `stl()` de R, pero esta función no permite que haya faltantes, por lo que antes habría que imputar valores faltantes en la serie a descomponer. Otro método de descomposición en el paquete `anomalize` está basado en el cálculo de la mediana de una ventana temporal móvil (ver documentación del paquete para más detalles).

El segundo paso involucra la detección de valores sospechosos o inusuales en la serie remanente (es decir, una vez removidos los componentes periódicos y de baja frecuencia). Nuevamente, el paquete `anomalize` ofrece varios métodos para la detección de posibles anomalías. En este control se utiliza el método llamado GESD, una implementación iterativa del test *Generalized Extreme Studentized Deviate* (Lau, 2015), que evalúa posibles anomalías progresivamente: las desviaciones más grandes son eliminadas paulatinamente y se recalculan valores críticos para el test que se reducen a medida que se eliminan los desvíos grandes. Por su naturaleza iterativa este método requiere más esfuerzo computacional, pero su mayor ventaja es que es resistente a la influencia de valores muy extremos. El método GESD se utiliza para todas las variables a las cuales se les aplica este control.

El control se ilustra en la figura 25, que muestra una serie de temperatura máxima diaria para Aeroparque Buenos Aires, Argentina, 2001-2010. La serie observada se descompone usando el paquete `stlplus` en tres componentes: (a) un componente estacional (con un período especificado de 365.25 días), (b) una tendencia de baja frecuencia y (c) una serie residual a la que se le aplica la detección de valores sospechosos. La suma de los tres componentes reconstruye la serie observada. A continuación, el paquete `anomalize` se utiliza para aplicar el método GESD

al componente residual, utilizando el parámetro α = 0.05. Este parámetro determina el ancho de la banda que define valores no sospechosos; un valor más bajo de α aumenta el ancho de la banda dentro de la cual se aceptan valores, haciendo más difícil que un valor sea considerado una anomalía. Los círculos rojos en cada panel de la figura 19 indican tres valores identificados como sospechosos; una revisión manual posterior indicó que estos valores sospechosos eran correctos.

Figura 25. Ejemplo de la identificación de valores sospechosos en base a la descomposición de una serie de temperatura máxima diaria para Aeroparque Buenos Aires, Argentina, 2001-2010. La serie original (panel superior) se descompone en (a) un componente estacional periódico, (b) una tendencia de baja frecuencia (que parece ser constante durante el período ilustrado) y (c) una serie residual a la que se le aplica la detección de valores sospechosos. Los círculos rojos en cada panel indican tres valores identificados como sospechosos; una revisión manual posterior indicó que los valores eran correctos.



8. Familia de controles de consistencia entre variables

En las secciones subsiguientes se presenta una familia de controles basados en las relaciones teóricas que existen entre las distintas variables que conforman la base de datos del CRC-SAS. Los controles de consistencia de variables encuentran amplia aplicación en diversos sistemas de control de calidad, dado que permiten detectar inconsistencias o incoherencias entre las variables incluidas en las bases de datos. Más información puede hallarse en Meek & Hatfield (1994), Kunkel et al. (1998), y Estévez et al. (2011).

8.1 Consistencia entre temperaturas

En este caso se busca que los valores de las temperaturas diarias (máxima, media, mínima) cumplan una serie de criterios basados en las definiciones de estas variables. Por ejemplo, se requiere que la temperatura mínima del día *i* sea inferior a la temperatura máxima de ese mismo día. También se pueden relacionar las temperaturas máximas y mínimas de días consecutivos para detectar inconsistencias. Los valores diarios que no cumplen con las relaciones propuestas son marcados como sospechosos. Los criterios utilizados para identificar inconsistencias en valores de temperaturas se listan abajo.

Consistencia entre temperaturas mínima, media y máxima diarias (test CEV01). Por definición, la temperatura mínima es la más baja de cada día, y la máxima es la más alta, mientras que la temperatura media es el promedio diario de varias temperaturas medidas a lo largo del día. Por lo tanto, se identifican como sospechosos los valores para el día i de temperaturas máxima, media y mínima que no cumplan con la siguiente relación:

$$\mathrm{Tmin}_i < \mathrm{Tmed}_i < \mathrm{Tmax}_i \ . \qquad (13)$$

Consistencia entre temperaturas medias calculadas de diferentes formas (test CEV02). La temperatura media diaria del aire puede calcularse de diferentes maneras. El valor de temperatura media $Tmed_i$ almacenado en la base de datos del CRC-SAS generalmente es el promedio de 3-4 observaciones a lo largo del día i[1]. Para este control, sin embargo, se calcula un nuevo valor de temperatura media basado en el promedio de las temperaturas mínima y máxima diarias: $Tmed_i^* = \frac{Tmax_i + Tmin_i}{2}$. En general, ambos valores son bastante parecidos, pero la existencia de diferencias marcadas entre $Tmed_i$ y Tmed_i^* puede ser una indicación de errores. Para permitir una cierta tolerancia, la temperatura media se considera sospechosa cuando la diferencia absoluta entre $Tmed_i$ y Tmed_i^* es superior al percentil 0.999 de todas las diferencias diarias.

Consistencia entre las temperaturas máximas y mínimas de días consecutivos (test CEV03 y CEV04). Dada la continuidad que generalmente presentan los valores de temperatura en días

[1] El número de observaciones que se utiliza para calcular variables agregadas para un día (como la temperatura media) también se lista en la base de datos del CRC-SAS, ya que este número puede variar entre países, entre estaciones meteorológicas e, incluso, en el tiempo para una misma estación.

sucesivos, se pueden plantear los siguientes criterios para evaluar la consistencia de los registros de temperatura de 3 días consecutivos: *i-1, i,* e *i+1*.

$$\text{Tmin}_{i-1} \leq \text{Tmax}_i \geq \text{Tmin}_{i+1} \text{ , y} \qquad (14)$$

$$\text{Tmax}_{i-1} \geq \text{Tmin}_i \leq \text{Tmax}_{i+1} \text{ .} \qquad (15)$$

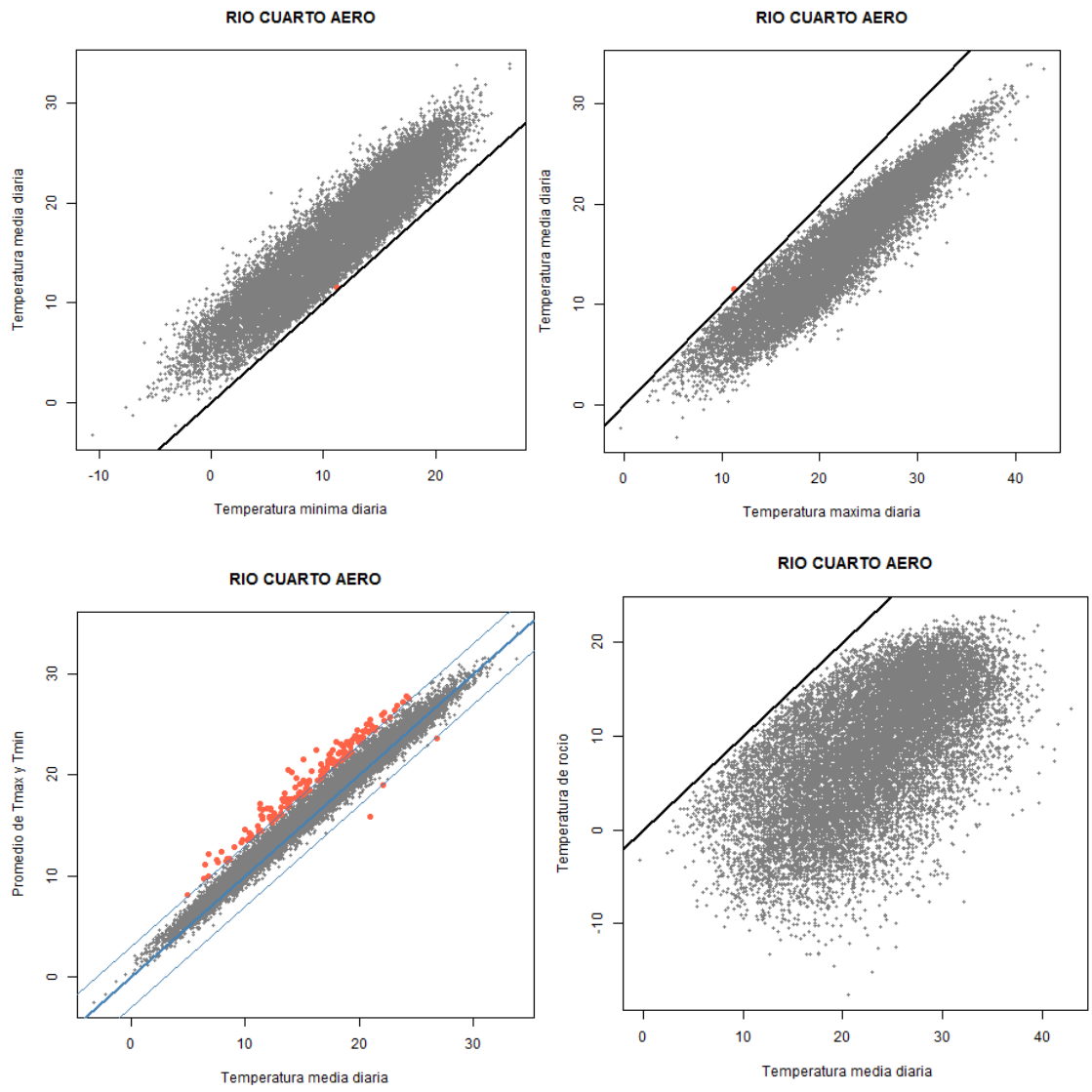
Cada una de las relaciones anteriores se verifica en un control separado (test CEV03 y CEV04, respectivamente). Si alguna de las dos relaciones no se cumple, se marcan como sospechosas las temperaturas involucradas de los tres días en cuestión.

Consistencia entre la temperatura media diaria y la temperatura media de rocío (test CEV05). Por su definición, la temperatura de rocío Td es menor o igual que la temperatura del aire T. Por lo tanto, el promedio diario de la temperatura de rocío tiene que ser menor o igual que la temperatura media del aire para el día *i*:

$$\underline{Td_i} \leq Tmed_i \text{ .} \qquad (16)$$

Cuando la relación (16) no se cumple, se marcan ambas variables como sospechosas. Los diferentes controles de consistencia en temperaturas máximas, mínimas, medias y de rocío se ilustran en la figura 26 con datos de Río Cuarto, Argentina.

Figura 26. Visualización de los controles de consistencia entre temperaturas diarias. Arriba, izquierda: relación entre temperatura mínima (eje x) y media (eje y). La línea 1:1 se indica en negro; hay un punto (rojo) identificado como sospechoso; aunque este punto está por encima de la línea 1:1 (o sea, Tmed ≥ Tmin), se marca el punto porque falla la relación entre Tmed y Tmax (ver panel de arriba a la derecha). Arriba, derecha: relación entre temperatura máxima (eje x) y media (eje y); se puede ver un punto rojo en el cual Tmed > Tmax. Abajo, izquierda: relación entre temperatura media calculada con varias observaciones diarias (eje x) y la semisuma de temperaturas extremas (eje y). Son puntos sospechosos los que están por encima o por debajo de una tolerancia determinada con respecto a la línea 1:1. Abajo, derecha: relación entre temperatura media (eje x) y temperatura de rocío (eje y).



Consistencia en la amplitud térmica diaria (CEV11). Para estudiar conjuntamente las temperaturas máximas y mínimas se calcula la amplitud térmica diaria (diferencia entre las temperaturas máxima y mínima diarias) aplicando distintos controles para identificar datos

erróneos de temperatura máxima y/o mínima, como por ejemplo que la diferencia entre la temperatura máxima y la mínima esté en el rango [0.01 – 30.00 °C].

8.2 Consistencia entre datos de presión atmosférica

Como la presión atmosférica es la fuerza que la columna de aire ejerce sobre una determinada superficie, se puede inferir que cuanto más alto esté ese punto, menor será la presión, dado que también es menor la cantidad de aire que hay por encima de él.

Todas las estaciones meteorológicas de la región del CRC-SAS se encuentran sobre el nivel del mar, por lo que son sospechosos los registros en los que la presión atmosférica al nivel de la estación $Pres_{est}$ es mayor o igual que la presión reducida al nivel del mar $Pres_{nm}$, o sea cuando no se cumple el siguiente criterio (CEV_06):

$$\mathrm{Pres}_{est} < Pres_{nm} \qquad (1)$$

8.3 Consistencia entre datos de viento

Los datos de viento en la base de datos del CRC-SAS incluyen tres variables: (a) la dirección media, (b) la velocidad media del viento a lo largo del día y (c) la velocidad máxima del viento observada en el día. Las direcciones del viento en cada observación se redondean a la decena de grado más próxima.

La información de viento es consistente cuando se cumplen las convenciones de registro de las observaciones. Una de estas convenciones es que la dirección 0 indica calma, o sea que la velocidad del viento es igual a 0 m s-1. Por lo tanto, si un registro incluye velocidades mayores que 0, la dirección no puede ser 0 (el viento norte se indica con la dirección 36, en decenas de grados). Por lo tanto, los dos criterios siguientes deben cumplirse simultáneamente, o los datos serán marcados como sospechosos (CEV_08):

$$vmax.d \geq 0 \ \wedge \ vmax.f \geq 0\,. \qquad (2)$$

Asimismo, el viento medio (o promedio diario de la velocidad del viento) debe ser menor o igual que el viento máximo diario (máxima velocidad del viento observada en el día). En este caso se busca detectar problemas en la digitalización de la información, tanto de los datos diarios de velocidad media como en la velocidad máxima del viento a lo largo del día. La condición que debe cumplirse es (CEV_07):

$$vmed \leq vmax.f, \, o \qquad (3)$$

$$vmed \ = 0 \ \wedge \ vmax.f = 0\,. \qquad (4)$$

8.4 Consistencia entre nubosidad y precipitación

La nubosidad y la precipitación se relacionan directamente. Para que se registren precipitaciones es necesario que el cielo presente cobertura nubosa. En este caso se identifican como sospechosos los días en los cuales se hayan observado precipitaciones (precipitación > 0.1 mm), pero cuyo promedio de nubosidad es igual a 0 (indicando cielo totalmente despejado (CEV_09). El error puede presentarse tanto en los datos de precipitación como en los de nubosidad, por lo que ambas variables se marcan como sospechosas.

Un problema asociado a este control es que los valores de estas dos variables corresponden a períodos distintos. La nubosidad diaria se calcula como el promedio de la nubosidad en las horas 6, 12, 18 y 24 UTC, mientras que la precipitación en un día es el valor acumulado desde las 12 UTC del día *i* hasta las 12 UTC del día *i+1*. Entonces, por ejemplo, si la precipitación ocurre solo durante la noche (entre las 0 y las 12 UTC del día *i+1)* y el cielo estuvo despejado (nubosidad 0) durante el día *i*, los datos no parecen consistentes, pero son válidos.

8.5 Consistencia entre nubosidad y heliofanía

La heliofanía representa la cantidad de horas diarias en las que la radiación solar incide directamente sobre la estación meteorológica. Cuando el cielo permanece completamente cubierto por nubes (ocho octavos de nubosidad se codifica como nubosidad = 8) a lo largo de todo el día, la cantidad de horas de sol debería ser igual a 0. Por lo tanto, se identifican como dudosos los días en los cuales la nubosidad promedio del día analizado es igual a 8 (cielo completamente cubierto) y la heliofanía es mayor que 0 (CEV_10). Como el error puede estar en la nubosidad o en la heliofanía, ambas variables se marcan como sospechosas.

9. Familia de controles de consistencia espacial

Todos los controles descritos hasta ahora analizan series temporales registradas en una única estación meteorológica. En la familia de controles de consistencia espacial, en cambio, se comparan los datos de una determinada estación meteorológica –denominada "estación central"– con datos de estaciones vecinas. La vecindad entre estaciones se define según (a) una distancia máxima a la estación central (por ejemplo, 200 km) y (b) una diferencia absoluta máxima de altitud con respecto a la estación central (por ejemplo, 100 m).

Los criterios de vecindad intentan garantizar que las estaciones incluidas en la comparación tengan condiciones similares a las de la estación central. Una desventaja es que diferencias grandes entre valores de estaciones vecinas puede deberse a causas como variaciones considerables de altitud en la región o de características geográficas, o al pasaje de frentes atmosféricos que afectan a una estación, pero no a la otra (Hubbard et al., 2012; Kunkel et al., 2005). Además de permitir el análisis de valores sospechosos, el método de comparación con estaciones vecinas puede servir para completar datos faltantes (Hubbard et al., 2012).

9.1 Control de regresión espacial ponderada

Este control (CES01) verifica que el valor de una variable meteorológica medida en la estación central y en un día determinado caiga dentro de un intervalo de confianza calculado a partir de ajustes estadísticos basados en datos de estaciones vecinas. El control se utiliza para temperatura máxima, mínima, media y de rocío. Los intervalos de confianza se estiman mediante una serie de regresiones simples entre los valores de la variable de interés en la estación central y los valores medidos en cada estación vecina. Para estimar los coeficientes de cada regresión se utiliza una ventana temporal de 91 días, centrada en el día cuyo valor está siendo controlado. Para incluir una estación vecina en el control de la estación central, se requiere que dentro de la ventana de 91 días (a) existan al menos 30 pares de valores (o sea, pares de mediciones para el mismo día en la estación central y la vecina) para la variable considerada, y (b) que la correlación con los valores de la estación central sea mayor que un umbral (por ejemplo, 0.8). Este control ha sido denominado *spatial regression test* o SRT por Hubbard et al. (2005), Hubbard et al. (2007), Hubbard et al. (2012) y Kunkel et al. (2005). Para ilustrar el uso del método, supongamos que estamos controlando el valor de temperatura máxima del aire Tx en la estación meteorológica c (central) medida el 22 de agosto de 1962.

1. El primer paso es extraer las temperaturas máximas diarias en cada estación i de las N estaciones meteorológicas vecinas a la estación c para la ventana temporal de 91 días centrada en el 22 de agosto de 1962; esta ventana incluye el período entre el 8 de julio y el 6 de octubre de 1962.

2. A continuación, se ajusta una serie de regresiones lineales simples entre la temperatura máxima Tx_ien cada estación vecina i (la variable independiente) y la temperatura máxima en la estación central (Tx_c, la variable dependiente). Cada regresión produce (a) una estimación Tx_i' de la temperatura máxima en la estación c para el centro de la ventana

temporal y (b) el error estándar s_i de la regresión (o valor cuadrático medio, o *rms* por las siglas en inglés de *root mean square*). Ambos valores se calculan a partir de los datos en la estación vecina i y la regresión estimada entre las dos estaciones.

3. Usando las regresiones para todas las estaciones vecinas, se calcula una estimación no sesgada Tx_c^* de la temperatura máxima en

$$\mathrm{Tx}_\mathrm{c}^* = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\frac{\mathrm{Tx}_\mathrm{i}'}{\mathrm{s}_\mathrm{i}^2} \right)}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\frac{1}{\mathrm{s}_\mathrm{i}^2} \right)}. \quad (21)$$

También usando todas las regresiones, se calcula un valor ponderado del error estándar de la estimación:

$$\mathrm{s}_\mathrm{c}^{*2} = N / \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{s}_\mathrm{i}^2} \quad (22)$$

Este método da mayor peso a las estimaciones de las estaciones vecinas que tienen una mayor asociación estadística con la estación central. A diferencia de otros métodos basados en distancias espaciales, aquí no se asume que la mejor estación para comparar con la estación central es la más cercana geográficamente. En cambio, el método explora las asociaciones entre los datos de cada estación vecina y la estación central para definir (a) qué estaciones deben incluirse en el análisis y (b) qué ponderación debe darse a cada estación vecina (Hubbard et al., 2005).

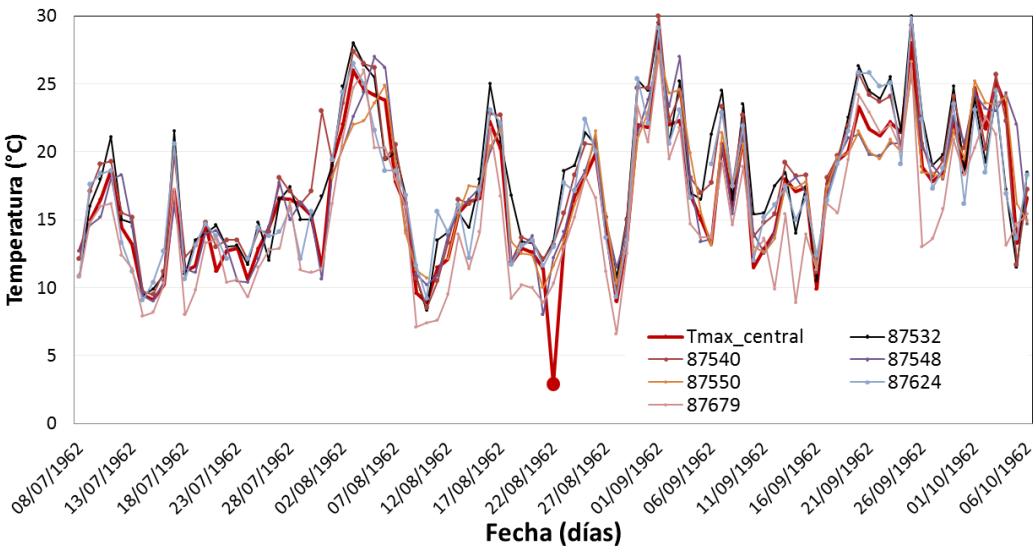
1. Considerando la estimación ponderada del valor en la estación central y el error estándar correspondiente –ecuaciones 21 y 22– se construye un intervalo de confianza alrededor de Tx_c, la temperatura máxima observada en la estación central en el día central de la ventana temporal:

$$Tx_c^* - f \;\leq\; Tx_c \;\leq\; Tx_c^* + f\, s_c^{*2}\,,$$

donde f es un factor de multiplicación. Si el valor observado Tx_c no cae dentro del intervalo de confianza, entonces se considera sospechoso.

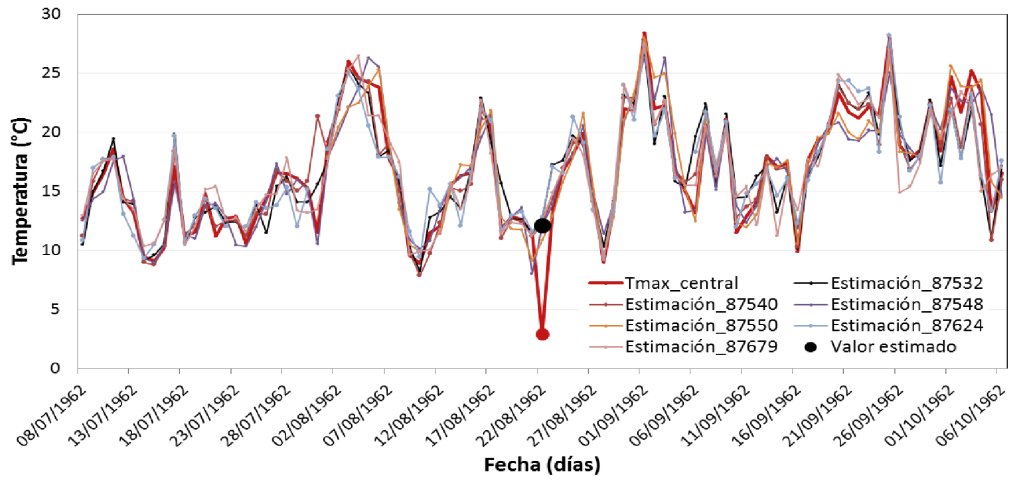
El desempeño de este control se ejemplifica en la figura 27 con datos de temperatura máxima medidos en Pehuajó, Argentina y en las seis estaciones más cercanas. La ventana temporal que se muestra está centrada en el 22 de agosto de 1962, que es el día para el cual se está controlando esta variable en Pehuajó; la ventana incluye 91 días entre el 8 de julio y el 6 de octubre de 1962. En general, las trazas observadas de temperatura máxima en todas las estaciones caen dentro de una banda bastante estrecha. Una excepción es la observación en Pehuajó para el 22 de agosto de 1962, que es sensiblemente más baja que los valores registrados ese día en las estaciones vecinas.

Figura 27. Series observadas de temperatura máxima en la estación central (Pehuajó, Argentina) y seis estaciones vecinas. Las series corresponden a una ventana temporal de 91 días centrada en el 22 de agosto de 1962; la ventana incluye el período entre el 8 de julio y el 6 de octubre de 1962. En general, las trazas de la temperatura máxima observadas en todas las estaciones caen dentro de una banda relativamente estrecha. Una excepción es la observación del 22 de agosto de 1962 en Pehuajó, que es sensiblemente más baja que los registros de ese día en las estaciones vecinas; la estimación ponderada del valor en la estación central (Pehuajó) para el 22 de agosto de 1962 tiene un valor de 12.1 °C y un intervalo de confianza entre 5.03 y 19.17 °C. La observación identificada como sospechosa (2.9 °C) cae fuera del intervalo.



Como parte del control de regresión espacial, se estiman valores de temperatura máxima para Pehuajó (la estación central) a partir de observaciones en las estaciones vecinas; las series de valores estimados se muestran en la figura 28. La estimación ponderada del valor en la estación central para el 22 de agosto de 1962 tiene un valor de 12.1 °C. El error estándar ponderado es de

Figura 28. Series de temperatura máxima observada (línea roja gruesa) y estimadas a partir de regresiones con estaciones vecinas para Pehuajó (estación central), Argentina. La estimación ponderada de la temperatura máxima del 22 de agosto de 1962 en (12.1 °C) se indica con un punto negro. La temperatura máxima registrada en Pehuajó (2.9 °C) se indica con un punto rojo. Este valor se confirmó como erróneo.



2.02 °C. Con un factor de multiplicación de 3.5, el intervalo de confianza – calculado según la ecuación 23 – para la observación estudiada se extiende desde 5.03 °C a 19.17 °C. El valor registrado de 2.9 °C (punto rojo en la figura 22) claramente cae fuera del intervalo y, en consecuencia, debe considerarse sospechoso. Una verificación posterior reveló que el dato correcto era 12.9 °C –es decir, al digitalizar la información se omitió el número 1.

9.2 Control de regresión espacial mediante un índice de concordancia

Para este control (CES02), se utilizan coeficientes de regresión e índices de concordancia calculados a partir de los valores medidos en una estación central (aquella que está siendo controlada) y un número de estaciones vecinas. Durre et al. (2010) presentan una descripción detallada de este control, que es relativamente similar al control de regresión espacial utilizado por Hubbard et al. (2005) y presentado en la sección anterior (sección 9.1). En el control usado por Hubbard et al. (2005), la ponderación de las estaciones vecinas se basa en el error estándar de las regresiones para cada una de ellas. En este caso, sin embargo, la ponderación se realiza mediante un *índice de concordancia* d, propuesto por Legates & McCabe Jr. (1999). Los pasos involucrados en este control se describen en los párrafos siguientes.

Tabla 4. Temperaturas máximas (°C) en Pehuajó, Argentina, y en cuatro estaciones meteorológicas vecinas. Los datos se muestran para una ventana temporal de 3 días entre el 21 y el 23 de agosto de 1962. El valor que se está controlando es el valor para Pehuajó, que por lo tanto se denomina *estación central* y el día del centro de la ventana temporal (22 de agosto de 1962); la celda correspondiente al valor siendo controlado está sombreada en azul. Las celdas en gris claro corresponden a los valores de los vecinos con menor diferencia absoluta respecto a la estación central (ver paso 2).

	TEMPERATURAS MÁXIMAS DIARIAS (° C)				
Fecha	**Estación central**	**Estación vecina 1**	**Estación vecina 2**	**Estación vecina 3**	**Estación vecina 4**
1962-08-21	11.4	9.1	8.0	12.0	10.7
1962-08-22	2.9	10.6	12.2	13.4	10.7
1962-08-23	16.7	16.6	17.2	19.0	16.1

1. El primer paso consiste en seleccionar, para cada estación vecina, los datos de la variable analizada (por ejemplo, temperatura máxima Tx) para una ventana de tres días centrada en el día que se está controlando en la estación central. Por ejemplo, si se está analizando la temperatura máxima en el día i en la estación central c que tiene cuatro estaciones vecinas, se reunirán 12 valores de temperatura máxima (tres observaciones correspondientes a los días $i-1$, i, e $i+1$ para cada una de las cuatro estaciones vecinas). Un ejemplo se ilustra en la tabla 4, que contiene datos de temperatura máxima en Pehuajó, Argentina, para el período de 3 días entre el 21 y 23 de agosto de 1962.

2. En segundo lugar, se calcula, para cada estación vecina, el valor absoluto de las diferencias entre (i) los tres valores dentro de la ventana temporal y (ii) el valor de la estación central en el día i (centro de la ventana). Para cada estación vecina, se reemplaza el dato original del día i por el valor correspondiente al día en la ventana que tenga la menor diferencia absoluta. En la tabla 4, los valores seleccionados para reemplazar los datos originales del 22 de agosto corresponden en realidad al 21 de agosto, y se indican con un color gris. Este paso se repite para cada día t de la serie original de datos, y así se forma una "serie nueva" de datos.

3. Para cada mes/año (por ejemplo, agosto de 1962) de la serie nueva creada en el paso 2, se define una ventana temporal que se extiende 15 días antes y 15 días después del comienzo del mes/año. Por ejemplo, para agosto de 1962, la ventana comienza el 17 de julio y termina el 15 de septiembre de 1962.

4. Para cada estación vecina, se realiza una regresión tomando como variable independiente las observaciones de esa estación y los valores en la estación central como variable dependiente. Se utiliza una estación vecina solamente (i) si hay un mínimo P de pares de valores dentro de

la ventana (es decir, datos para el mismo día en la estación central y la vecina) y (ii) si la correlación r entre la estación central y la vecina es mayor que un umbral determinado. En este caso utilizamos P = 30 y r = 0.80. Si no se cumplen estas condiciones, los coeficientes de la regresión para esa ventana y estación se definen como faltantes y no se puede hacer el control.

5. Para cada estación vecina, se calcula el índice de concordancia (*index of agreement*) d propuesto por Legates & McCabe Jr. (1999), que se calcula como

$$\mathrm{d} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} |\mathrm{y(i)} - \mathrm{x(i)}|}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \left[\, |\mathrm{x(i)} - \bar{\mathrm{y}}| + |\mathrm{y(i)} - \bar{\mathrm{y}}| \,\right]}, \qquad (24)$$

donde m es el número de pares de datos válidos dentro de la ventana, $x(i)$ e $y(i)$ son las observaciones en la estación vecina y la estación central, respectivamente, para el día i de la ventana, e $\bar{y}$ denota un promedio sobre todas las observaciones dentro de la ventana temporal. Valores altos de d indican una alta correlación así como diferencias absolutas pequeñas entre x e y (Durre et al., 2010). Se ordenan las estaciones vecinas de mayor a menor por índice d y se excluyen las vecinas cuyo orden sea mayor que el número máximo de estaciones especificado.

6. Con los resultados de los pasos (4) y (5), el valor de la estación central para cada día en la ventana temporal se estima a partir de las regresiones y el índice de concordancia para cada estación vecina. La estimación se calcula como

$$\hat{\mathrm{y}}_{\mathrm{i}} = \frac{\sum_{\mathrm{k}=1}^{\mathrm{n}} \left[\mathrm{a(k)} + \mathrm{b(k)x'(i,k)}\right] \mathrm{d(k)}}{\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{d(k)}}, \qquad (25)$$

donde $\hat{y}_i$ es la estimación para la estación central en el día i, n es el número válido de estaciones vecinas, $a(k)$ y $b(k)$ son respectivamente la ordenada al origen o intercepto y la pendiente de la regresión para la estación k, y $x'(i,k)$ es la observación para la estación vecina k en el día i; este valor corresponde a la serie nueva (ver paso 2) derivada a partir de los 3 días centrados en el día i.

7. El último paso consiste en la determinación de valores sospechosos. Para hacer esta determinación, Durre et al. (2010) definen como sospechosos los valores que cumplen las dos condiciones siguientes:
   - El valor absoluto de la diferencia entre el valor de la estación central y la estimación de ese valor (que se denomina residuo) es > 8 °C para el control de temperaturas; y
   - El valor absoluto del residuo estandarizado es > 4; el residuo estandarizado se calcula restando la media de todos los residuos dentro de cada mes/año y dividiendo por el desvío estándar de esos residuos.

Para hacer el control más flexible, se modifican aquí los criterios originales: se define como sospechoso todo valor cuyo residuo y residuo estandarizado sean mayores que un cierto percentil (en este caso, 0.99).

Según Durre et al. (2010), este control incluye varias diferencias con respecto a la regresión espacial descrita por Hubbard & You (2005) que apuntan a reducir la cantidad de falsas alarmas (valores identificados como sospechosos que son en realidad correctos). Primero, en lugar de la correlación o el error estándar de la regresión para ponderar la asociación con cada vecino, se usa el índice de concordancia d de Legates & McCabe Jr. (1999) que, además de medir la covarianza entre vecino y estación central, mide las diferencias absolutas entre ambas series. En consecuencia, la selección de estaciones vecinas con valores altos del índice d debería reducir el riesgo de tener residuos extremos cuyos valores sean causados por errores en el cálculo de la estimación, más que por el valor observado en la estación central. Segundo, al usar una ventana de tres días, se reducen los errores en la estimación que puedan ser causados por diferencias asociadas con eventos meteorológicos como el pasaje de un frente (Hubbard et al., 2012). Finalmente, el uso simultáneo de los residuos y los residuos estandarizados reduce el riesgo de emitir muchas falsas alarmas cuando el desvío estándar de los residuos es pequeño.

La figura 29 (panel superior) muestra una asociación bastante cercana entre las temperaturas máximas observadas en Pehuajó (“estación central”) y los valores estimados para esa estación a partir de 4 estaciones vecinas, y en el panel inferior se indican los residuos. En la figura 30 se observa que los valores observados y los estimados a partir de estaciones vecinas son muy similares, excepto para el 22 de agosto, fecha para la cual el valor observado ya fue reportado como erróneo. En el panel inferior de la figura 30, se muestran residuos absolutos y estandarizados. Claramente, los residuos del 22 de agosto superan ambos umbrales establecidos.

Figura 29. Panel superior: relación entre los valores observados de temperatura máxima en Pehuajó, Argentina y los valores estimados en base a 4 estaciones vecinas. Los puntos identificados como sospechosos se muestran en rojo y la línea 1:1 en azul. Panel inferior: Serie temporal de residuos para la estación central. El umbral para la identificación de valores sospechosos es la línea horizontal azul.

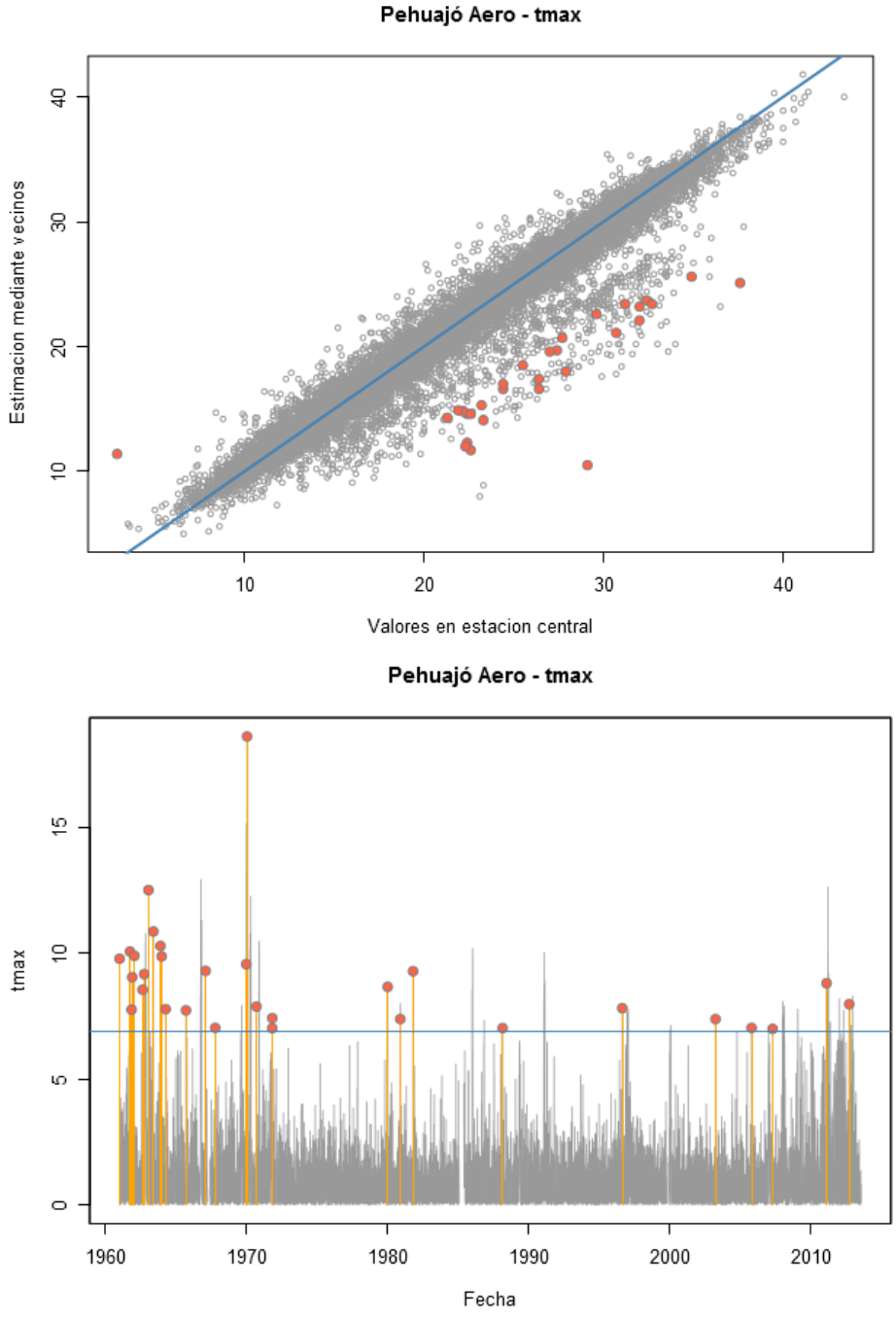
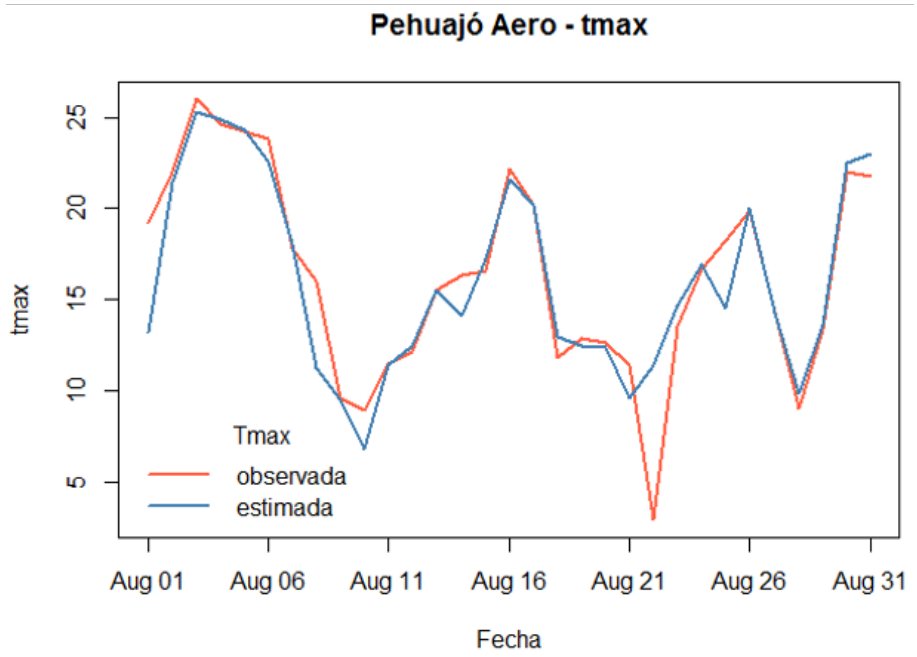
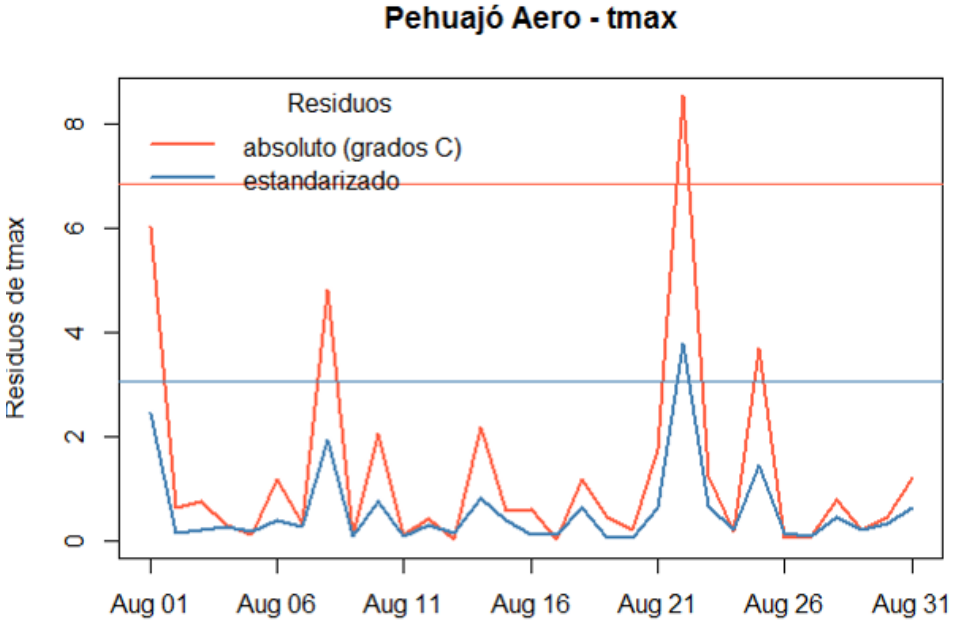


Figura 30. Panel superior: Valor observado de temperatura máxima para una estación central (Pehuajó, Buenos Aires, Argentina) y valor estimado a partir de 4 estaciones vecinas. Las series corresponden a agosto de 1962. Panel inferior: Series temporales de diferencias o residuos entre valores observados y estimados para Pehuajó. Los umbrales considerados para considerar como sospechosos a los residuos absolutos (en °C) y estandarizados (sin unidades) se muestran como líneas horizontales en los colores respectivos. El valor para el 22 de agosto de 1962 muestra residuos por encima de los dos umbrales.





9.3 Corroboración espacial de registros de temperatura

Este control (CES03) se aplica a las temperaturas diarias mínimas, máximas, medias y de rocío y se basa en el método de corroboración espacial descrito en la página 1625 de Durre et al. (2010) para determinar si un valor cae fuera de un rango de valores reportados en estaciones vecinas. A diferencia de los controles anteriores, las estaciones vecinas se seleccionan puramente según un criterio de distancia: las estaciones vecinas deben estar a menos de 300 km de la estación central y tener una diferencia de elevación de menos de 100 m respecto a la estación central; estos valores son configurables. Para poder realizar este control se requieren al menos 3 estaciones vecinas con datos y un máximo de 5 vecinas con datos (este valor es configurable). La principal ventaja de este control es su aplicabilidad en áreas donde la alta variabilidad espacial o la falta de series completas impiden estimar regresiones entre estaciones (Durre et al., 2010).

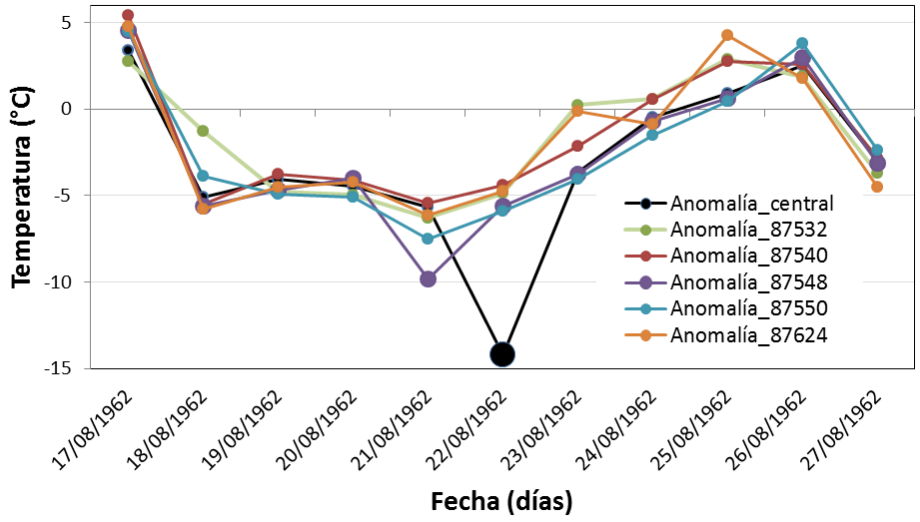
El control CES03 utiliza anomalías de temperaturas en la estación central y en estaciones vecinas. Estas anomalías se calculan respecto a un valor medio estimado para cada estación y día del año. El primer paso del control, entonces, es el cálculo del promedio de la variable analizada (por ejemplo, temperatura máxima diaria) para cada día del año y cada estación. Para calcular este promedio se utiliza un procedimiento resistente a valores extremos (la función *biweight*) y todos los valores históricos para la variable y estación consideradas. Para el análisis, se considera una ventana temporal centrada en el día para el que se realiza el cálculo; por ejemplo, Durre et al. (2010) consideran una ventana de 15 días. En el control que se realiza en la base de datos CRC-SAS se utilizará una ventana de 21 días: por ejemplo, la media para el 11 de enero se calcula con los valores observados entre el 1 y el 21 de enero de todos los años disponibles para la estación. Deben existir al menos el 10 % de valores posibles (ancho de la ventana por número de años disponibles para hacer el cálculo de la media *biweight*). A continuación, se construye una serie temporal de anomalías –respecto al valor promedio– de cada día del año.

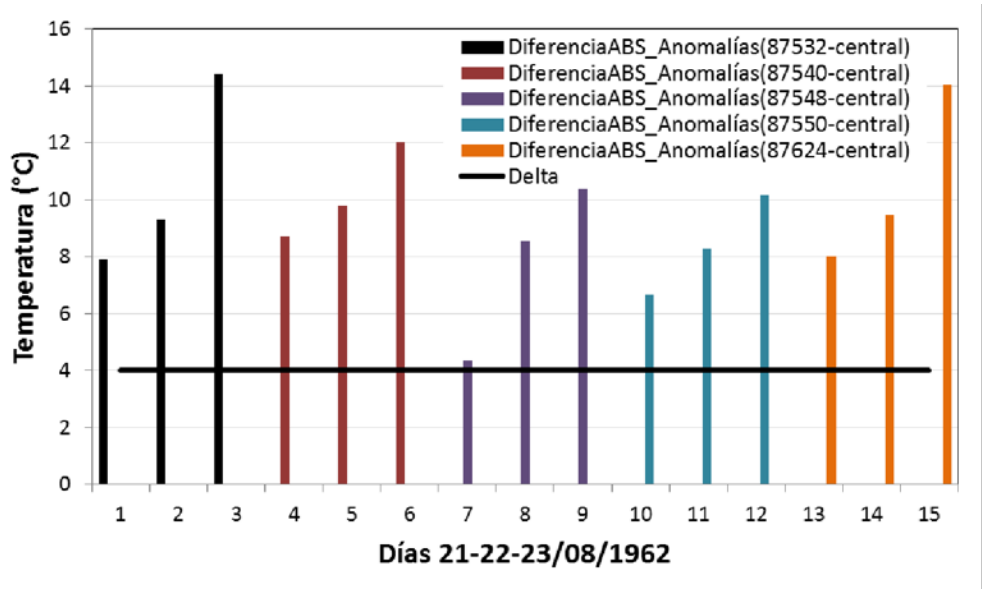
El segundo paso consiste en seleccionar –para cada estación meteorológica vecina– las anomalías de la variable analizada (por ejemplo, temperatura máxima) para una ventana de 3 días centrada en el día que se está controlando en la estación central. Este paso es similar al descrito en la sección 9.2. Si se está analizando la temperatura máxima del día t en la estación central c, que tiene 5 estaciones vecinas, se reunirán 15 valores de anomalías de temperaturas máximas (3 observaciones correspondientes a los días $t-1$, t, y $t+1$ para cada una de las 5 estaciones vecinas). Para realizar el control, se requiere que las estaciones vecinas (como mínimo 3 vecinas) tengan un número mínimo de anomalías disponibles (por ejemplo, 9).

En tercer lugar, se calculan las diferencias absolutas entre (i) las anomalías de las estaciones vecinas y (ii) la anomalía de la estación central en el día t (centro de la ventana). Si todas estas diferencias son superiores a un umbral definido, se considera que las anomalías vecinas no corroboran la anomalía de la estación central, y por lo tanto el valor en la estación central no pasa el control. El umbral utilizado se define en términos de temperatura: por ejemplo, Durre et al. (2010) usan un umbral de 10 °C; en el control que se realiza en la base de datos CRC-SAS se usa un umbral de 2 °C. Si no se pasa el control, se marca como sospechoso el valor de la variable considerada para el día t en la estación central c.

El control de corroboración espacial examina la asociación entre valores observados en una ventana temporal relativamente corta (3 días). En consecuencia, se puede utilizar en situaciones en las que es imposible realizar el control de regresión espacial, ya sea por datos incompletos dentro de la ventana temporal usada para estimar la regresión –generalmente más ancha– o porque la correlación entre estación central y sus vecinas es muy baja. La desventaja, sin embargo, es que el control de corroboración no puede detectar inconsistencias espaciales de tan baja magnitud como aquellas identificadas por el control de regresión. En consecuencia, los controles de regresión y corroboración se complementan mutuamente (Durre et al., 2010). La figura 30 ilustra la corroboración espacial de la temperatura máxima diaria en Pehuajó.

Figura 30. Panel superior: Temperatura máxima observada en la estación central (Pehuajó, Argentina) y cinco estaciones vecinas (General Pico (87532), Trenque Lauquén (87540), Junín (87548), Nueve de Julio (87550) y Anguil (87624)). Las series corresponden al período entre el 17 y el 27 de agosto de 1962. Panel inferior: Diferencias absolutas entre anomalías de temperaturas en la estación central y en estaciones vecinas, 21 al 23 de agosto de 1962. Las anomalías se calculan respecto a un valor promedio estimado para cada estación y día del año. Si todas las diferencias absolutas consideradas están por encima de un umbral determinado (la línea horizontal en la figura) se considera que el valor de la estación central es sospechoso.





9.4 Corroboración espacial de la precipitación

Este control (CES04) –basado en el control de corroboración espacial de precipitación propuesto por Durre et al. (2010)– se aplica solamente a las precipitaciones diarias. El control intenta determinar si un valor en la estación central es muy diferente del rango de valores reportados en estaciones geográficamente vecinas dentro de una ventana temporal de 3 días. Las estaciones vecinas deben estar a menos de 300 km de la estación central y tener una diferencia de elevación de menos de 100 m respecto a la estación central; estos valores son configurables. Para poder realizar este control se requieren al menos 3 estaciones vecinas con datos y un máximo de 5 vecinas con datos (este valor es configurable). El control se basa en la comparación de (i) la precipitación en la estación central *c* y el día *t* con (ii) el rango de lluvias observadas para las estaciones vecinas en los días $t-1$, t, y $t+1$. Si la lluvia para la estación central en el día t cae *dentro* de ese rango, el valor en la estación central pasa el control y no se considera sospechoso. En cambio, si la lluvia registrada en la estación central cae *fuera* del rango definido por las precipitaciones en las estaciones vecinas, entonces se realiza una verificación adicional para decidir si el valor se considera sospechoso.

En la verificación adicional, la diferencia entre la lluvia en la estación central y el siguiente valor más alto o más bajo debe exceder un umbral determinado. Para definir este umbral, primero se calcula una cantidad llamada MATD, por las siglas de *Minimum Absolute Target–Neighbor Difference* (ver Apéndice C de Durre et ál. (2010)). El cálculo se realiza primero en base a las diferencias entre el valor central (a saber, la lluvia registrada el día t en la estación central) y las lluvias en estaciones vecinas dentro de una ventana de 3 días (días $t-1$, t y $t+1$). Si el valor central es mayor que el valor vecino más alto, o menor que el valor vecino más bajo, el MATD calculado con las lluvias se define como el valor absoluto de la diferencia más pequeña entre el valor central y los valores vecinos. Si no es así, el MATD se define como cero (0).

El MATD se calcula también para los rankings climatológicos de los valores de precipitación descritos en el párrafo anterior, y se denomina $MATD_{ranking}$. Los *rankings* (orden de menor a mayor) se estiman para cada estación usando todos los valores de precipitación mayores a 0.1 mm (la definición de día lluvioso) en una ventana (parametrizable) de 29 días centrada en el día analizado (día t) y todos los años observados. Los *rankings* se expresan en porcentaje del valor máximo observado dentro de la ventana. Para el cálculo del *ranking* se requiere que dentro de la ventana analizada existan al menos 20 valores superiores a 0.1 mm. Si existen suficientes valores para estimar $MATD_{ranking}$, se estima un umbral para el control usando una formulación similar a la ecuación C1 (página 1631) de Durre et al. (2010):

$$\mathrm{U} = -45.72\ \ln\left(\mathrm{MATD}_{\mathrm{ranking}}\right) + 180.00\,. \qquad (5)$$

Se hace notar que en el trabajo original de Durre et al. (2010), la constante que se suma en el último término es de 180.00 mm en lugar de 269.24 mm, como en el trabajo original. Si el mínimo valor absoluto de las diferencias entre rankings porcentuales excede el umbral U, el valor central de precipitación se marca como sospechoso. Si por algún motivo no se pueden calcular los

rankings porcentuales para la estación central o para un número suficiente de estaciones vecinas, el umbral U se define como el máximo de la función en la ecuación –en este caso ≈ 180 mm.

9.5 Diferencias con valores interpolados a partir de datos vecinos

Este control (CES05) verifica que el valor de una variable meteorológica medida en la estación central y en un día determinado se encuentre dentro de un intervalo de confianza calculado a partir de la interpolación de datos de estaciones vecinas. El control se utiliza para temperaturas máxima, mínima, media y de rocío, humedad relativa, presión al nivel del mar, viento medio y viento máximo.

Para poder realizar este control se requieren al menos 5 estaciones vecinas con datos (valor configurable). Estas estaciones deben estar a menos de 250 km de la estación central y tener una diferencia de elevación de menos de 200 m respecto a la estación central; estos valores son configurables.

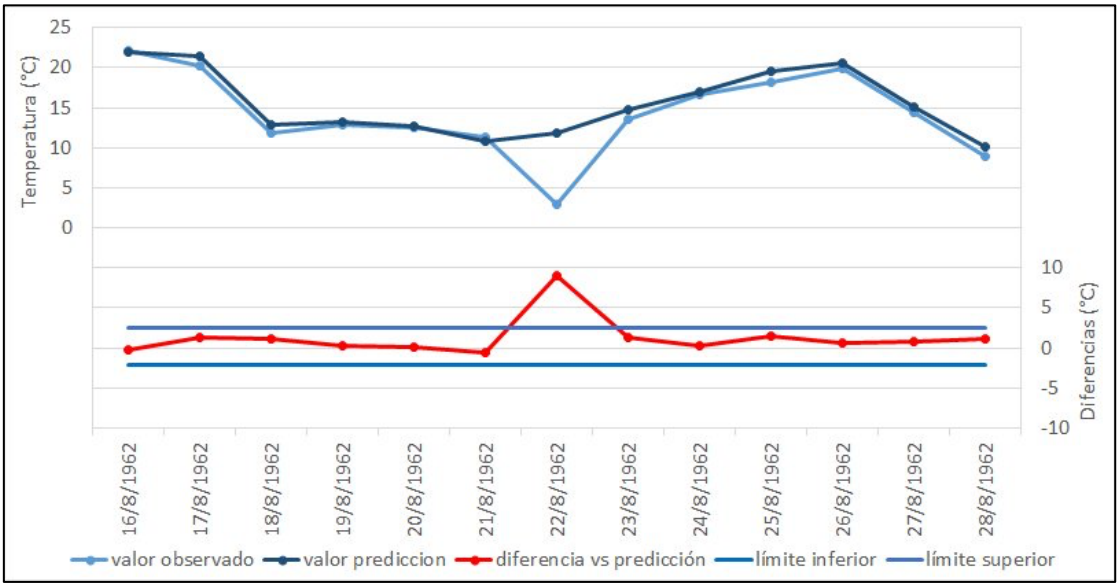
El control CES05 compara al valor registrado en la estación central con una estimación basada en los datos de las estaciones vecinas utilizando distancia inversa ponderada (o IDW por las siglas en inglés de inverse distance weighted). La diferencia entre ambos valores se compara con un intervalo generado a partir de valores estadísticos de un periodo suficientemente largo.

Los estadísticos se calculan utilizando los registros de todos los días de cada mes del año en el periodo 1971-2010. Se calculan las diferencias entre el valor observado en la estación central y el estimado con IDW utilizando solo datos de las estaciones vecinas. Si se cuenta con al menos 100 diferencias diarias, entonces se calcula un estimador resistente de tendencia central (mediana, o cuantil 50, “Q50”) y otro de dispersión de los datos (estadístico MAD o mediana de desvíos absolutos). Para cada mes, utilizando un factor.mad de 3 (valor configurable), el intervalo de confianza para evaluar los valores de la estación central se construye de la siguiente forma:

$$[Q50_m - MAD_m * \text{factor.mad} \; ; \; Q50_m + MAD_m * \text{factor.mad}] \quad (27)$$

Si la diferencia entre valor de la variable para la estación central y la estimación cae dentro del rango, el valor en la estación central pasa el control y no se considera sospechoso. En cambio, si la diferencia cae fuera del rango definido, entonces se considera que el valor de la estación central es sospechoso. La figura 31 muestra el caso de la temperatura máxima en Pehuajó y se observa que el valor del 22 de agosto de 1962 resulta sospechoso.

Figura 31. Panel superior: Temperatura máxima observada en la estación central (Pehuajó, Argentina) y valor de la predicción utilizando IDW. Las series corresponden al período entre el 16 y el 28 de agosto de 1962. Panel inferior: diferencias entre el valor de la temperatura máxima y el cálculado con IDW, junto con los límites superior e inferior, para los mismos días. Si la diferencia está por fuera de los umbrales determinados se considera que el valor de la estación central es sospechoso.



10. Estaciones automáticas

En la región de CRS-SAS se cuenta con 626 estaciones automáticas. La distribución de la totalidad de estaciones (convencionales y automáticas) se muestra en la figura 32. De ellas, 475 estaciones corresponden a Brasil, 70 a Argentina y 81 a Chile.

A esta información en alta resolución temporal se le aplican controles de calidad específicos para tratar de separar aquellos datos válidos de los que aparentemente no lo son y se realiza una agregación de variables a nivel diario para algunas de las variables reportadas. Las variables agregadas a valor diario tienen un único proceso de control de calidad, sin distinguir si provienen de observaciones manuales o automáticas, pero en el caso de las automáticas no se realiza un control manual.

Por lo tanto, se han implementado una serie de controles de calidad específicos para las observaciones meteorológicas de estaciones automáticas. Algunos de los controles implementados son similares a aquellos desarrollados para el control de datos diarios. Los controles de calidad implementados para datos meteorológicos sub-diarios se han organizado en cinco grandes grupos o “familias”:

- **Controles generales.** Estos controles verifican la integridad general de los datos. Por ejemplo, se controla que no haya fechas duplicadas o fuera de secuencia en las observaciones diarias. Estos controles son realizados antes de que los datos entren al servidor.

- **Controles de rango fijo.** Estos controles aseguran que no existan valores físicamente imposibles o nunca antes observados en el registro histórico de una variable meteorológica (por ejemplo, temperaturas mayores a 55 °C).

- **Controles de rango variable.** En esta familia, los rangos o umbrales usados para “marcar” valores sospechosos varían con el tiempo, tomando valores específicos para cada día o mes del año, por lo que los controles son más finos o sensibles que los controles de rango fijo.

- **Controles de continuidad temporal.** Estos controles estudian las secuencias de valores de cadavariable en días consecutivos. Algunos de los controles en esta familia detectan la presencia de saltos o picos inusuales en las series de datos.

- **Controles de consistencia entre variables.** Una serie de controles en esta familia o grupo evalúan la consistencia entre valores de pares o tríadas de variables que deben guardar cierta consistencia. Un ejemplo obvio es la verificación de que la temperatura mínima diaria sea menor o igual que la temperatura máxima diaria.

El resultado de la etapa de control de calidad es una serie de datos marcados como *sospechosos* o *inválidos*. Los controles de calidad están agrupados (además de las cinco familias listadas arriba)

en dos categorías, basadas en la certeza con que un dato puede ser considerado erróneo: (a) controles *principales* y (b) controles *complementarios*. Si un dato falla al menos un control principal, indefectiblemente es considerado como *inválido* (por ejemplo, un test de rango fijo al cual se le presenta una temperatura de 87 °C). Sin embargo, si un dato pasa todos los controles principales pero falla al menos uno de los controles complementarios, entonces es considerado como *sospechoso*.

Finalmente, si el dato no falla ninguno de los controles (tanto principales como complementarios) se lo considera *válido*.

En estaciones automáticas, si hay datos que no pasan algún control, no hay manera de controlar la veracidad de los datos (como sí puede hacerse conlos datos de estaciones convencionales, para los cuales existen formularios de registro de datos) y los datos “sospechosos” o “inválidos” son marcados como tales en la base de datos y pueden ser excluidosde análisis subsiguientes. Sin embargo, existe una mayor certeza de que los datos marcados como inválidos son realmente erróneos, dado el mayor rigor de los controles principales.

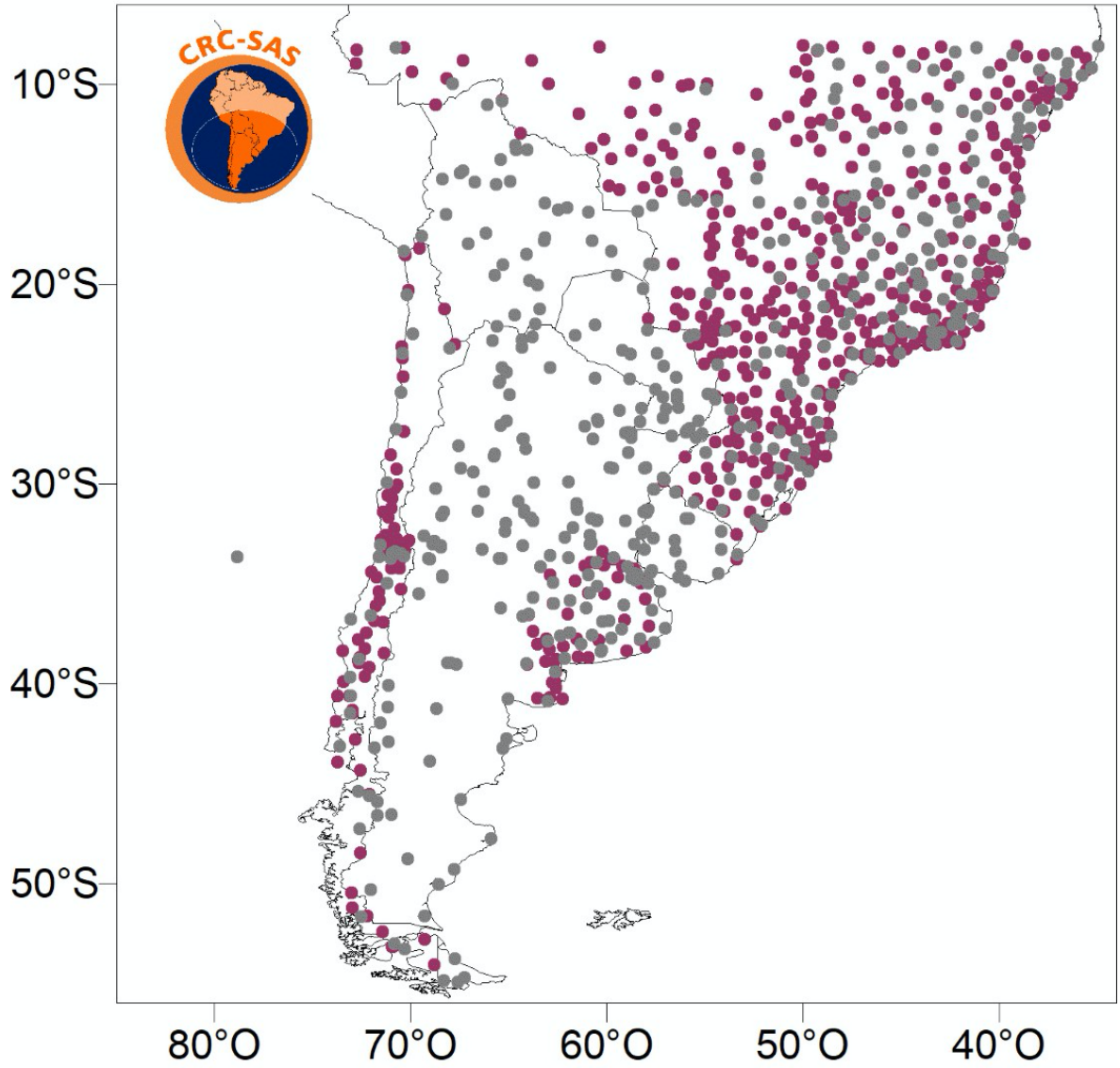


Figura 32. Estaciones automáticas (violeta) y convencionales (gris) en la región CRC-SAS.

Referencias

Aizpuru, J. & Leggieri, L., 2008. Predicción de indicadores de cambio climático para Argentina durante el siglo XXI, Universidad de Buenos Aires, Buenos Aires, Argentina.

Boulanger, J.-p., Aizpuru, J., Leggieri, L. & Marino, M., 2010. A procedure for automated quality control and homogenization of historical daily temperature and precipitation data (APACH): part 1: quality control and application to the Argentine weather service stations. Climatic Change, 98(3-4): 471-491.

Corripio, J.G., 2003. Vectorial algebra algorithms for calculating terrain parameters from DEMs and solar radiation modelling in mountainous terrain. International Journal of Geographical Information Science, 17(1): 1-23.

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G. & Vose, R.S., 2010. Comprehensive automated quality assurance of daily surface observations. Journal of Applied Meteorology and Climatology, 49(8): 1615-1633.

Estévez, J., Gavilán, P. & Giráldez, J.V., 2011. Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology, 402(1–2): 144-154.

Feng, S., Hu, Q. & Qian, W., 2004. Quality control of daily meteorological data in China, 1951-2000: a new dataset. International Journal of Climatology, 24: 853-870.

Forsythe, W.C., Rykiel Jr., E.J., Stahl, R.S., Wu, H.-i. & Schoolfield, R.M., 1995. A model comparison for daylength as a function of latitude and day of year. Ecological Modelling, 80: 87-95.

Freires Lúcio, F.D. & Grasso, V.F., 2016. The global framework for climate services (GFCS). Climate Services, 2–3: 52-53.

González-Rouco, J.F., Jiménez, J.L., Quesada, V. & Valero, F., 2001. Quality control and homogeneity of precipitation data in the southwest of Europe. Journal of Climate, 14(5): 964-978.

Hastie, T. & Tibshirani, R., 1990. Generalized additive models. Monographs on Statistics & Applied Probability. Chapman & Hall / CRC, Boca Raton, Florida, USA.

Hoaglin, D., F. Mosteller, J. Tukey, 1983. Understanding robust and exploratory data analysis. Wiley Classic Library.

Hoaglin, D.C., Mosteller, F. & Tukey, J.W., 2000. Understanding robust and exploratory data analysis. Wiley Classics Library. John Wiley & Sons, New York.

Hubbard, K., You, J. & Shulski, M., 2012. Toward a better quality control of weather data. In: M.S.F. Nezhad (Editor), Practical concepts of quality control. InTech.

Kunkel, K.E. et al., 1998. An expanded digital daily database for climatic resources applications in the midwestern United States. Bulletin of the American Meteorological Society, 79(7): 1357-1366.

Lanzante, J.R., 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data. International Journal of Climatology, 16: 1197-1226.

Meek, D.W. & Hatfield, J.L., 1994. Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology, 69(1–2): 85-109.

Peterson, T.C., Vose, R., Schmoyer, R. & RazuvŠev, V., 1998. Global historical climatology network (GHCN) quality control of monthly temperature data. International Journal of Climatology, 18: 1169-1179.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org.

Vicente-Serrano, S., 2006. Differences in spatial patterns of drought on different time scales: An analysis of the Iberian Peninsula. Water Resources Management, 20(1): 37-60.

Aizpuru, J. & Leggieri, L., 2008. Predicción de indicadores de cambio climático para Argentina durante el siglo XXI, Universidad de Buenos Aires, Buenos Aires, Argentina.

Boulanger, J.-p., Aizpuru, J., Leggieri, L. & Marino, M., 2010. A procedure for automated quality control and homogenization of historical daily temperature and precipitation data (APACH): part 1: quality control and application to the Argentine weather service stations. Climatic Change, 98(3-4): 471-491.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. & Terpenning, I., 1990. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1): 3-73.

Corripio, J.G., 2003. Vectorial algebra algorithms for calculating terrain parameters from DEMs and solar radiation modelling in mountainous terrain. International Journal of Geographical Information Science, 17(1): 1-23.

Delignette-Muller, M.L. & Dutang, C., 2015. fitdistrplus: An R package for fitting distributions. Journal of Statistical Software; 64(4).

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G. & Vose, R.S., 2010. Comprehensive automated quality assurance of daily surface observations. Journal of Applied Meteorology and Climatology, 49(8): 1615-1633.

Estévez, J., Gavilán, P., García-Marín, A.P. & Zardi, D., 2015. Detection of spurious precipitation signals from automatic weather stations in irrigated areas. International Journal of Climatology, 35(7): 1556-1568.

Estévez, J., Gavilán, P. & Giráldez, J.V., 2011. Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology, 402(1–2): 144-154.

Feng, S., Hu, Q. & Qian, W., 2004. Quality control of daily meteorological data in China, 1951-2000: a new dataset. International Journal of Climatology, 24: 853-870.

Fiebrich, C.A., Morgan, C.R., McCombs, A.G., Hall, P.K. and McPherson, R.A., 2010. Quality assurance procedures for mesoscale meteorological data. Journal of Atmospheric and Oceanic Technology, 27(10): 1565-1582.

Forsythe, W.C., Rykiel Jr., E.J., Stahl, R.S., Wu, H.-i. & Schoolfield, R.M., 1995. A model comparison for daylength as a function of latitude and day of year. Ecological Modelling, 80: 87-95.

González-Rouco, J.F., Jiménez, J.L., Quesada, V. and Valero, F., 2001. Quality control and homogeneity of precipitation data in the southwest of Europe. Journal of Climate, 14(5): 964-978.

Hafen, R., 2016. stlplus: Enhanced seasonal decomposition of time series by loess. https://CRAN.R-project.org/package=stlplus.

Hastie, T. & Tibshirani, R., 1990. Generalized additive models. Monographs on Statistics & Applied Probability. Chapman & Hall / CRC, Boca Raton, Florida, USA.

Hoaglin, D.C., Mosteller, F. & Tukey, J.W., 2000. Understanding robust and exploratory data analysis. Wiley Classics Library. John Wiley & Sons, New York.

Hubbard, K., You, J. & Shulski, M., 2012. Toward a better quality control of weather data. In: M.S.F. Nezhad (Editor), Practical Concepts of Quality Control. InTech.

Hubbard, K.G., Goddard, S., Sorensen, W.D., Wells, N. & Osugi, T.T., 2005. Performance of quality assurance procedures for an applied climate information system. Journal of Atmospheric and Oceanic Technology, 22(1): 105-112.

Hubbard, K.G., Guttman, N.B., You, J. & Chen, Z., 2007. An improved QC process for temperature in the daily cooperative weather observations. Journal of Atmospheric and Oceanic Technology, 24(2): 206-213.

Hubbard, K.G. & You, J., 2005. Sensitivity analysis of quality assurance using the spatial regression approach—A case study of the maximum/minimum air temperature. Journal of Atmospheric and Oceanic Technology, 22(10): 1520-1530.

Kunkel, K.E. et al., 1998. An expanded digital daily database for climatic resources applications in the midwestern United States. Bulletin of the American Meteorological Society, 79(7): 1357-1366.

Kunkel, K.E. et al., 2005. Quality control of pre-1948 cooperative observer network data. Journal of Atmospheric and Oceanic Technology, 22(11): 1691-1705.

Lanzante, J.R., 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data. International Journal of Climatology, 16: 1197-1226.

Lau, A.T.C., 2015. GESD – A robust and effective technique for dealing with multiple outliers. ASTM Standardization News: 40-41.

Legates, D.R. & McCabe Jr., G.J., 1999. Evaluating the use of ‘‘goodness-of-fit’’ measures in hydrologic and hydroclimatic model evaluation. Water Resour. Res., 35: 233-241.

Meek, D.W. & Hatfield, J.L., 1994. Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology, 69(1–2): 85-109.

Novack-Gottshall, P. & Wang, S.C., 2019. KScorrect: Lilliefors-corrected Kolmogorov-Smirnov goodness-of-fit test.

OMM, 2017. Guía del Sistema Mundial de Observación de la Organización Meteorológica Mundial. OMM-No. 488. https://library.wmo.int/?lvl=notice_display&id=12763#.Yrm3P3bMKUk

Peterson, T.C., Vose, R., Schmoyer, R. & RazuvŠev, V., 1998. Global historical climatology network (GHCN) quality control of monthly temperature data. International Journal of Climatology, 18: 1169-1179.

Vicente-Serrano, S., 2006. Differences in spatial patterns of drought on different time scales: An analysis of the Iberian Peninsula. Water Resources Management, 20(1): 37-60.

Apéndice A

A1. Formato de los archivos de transferencia de datos climáticos

Como se ha discutido en el texto, los miembros del CRC-SAS contribuirán datos de variables meteorológicas archivadas a nivel diario a la base de datos del CRC. Estos datos se importarán a través de (a) archivos de texto organizados por estación meteorológica o, alternativamente, (b) un archivo único con datos de varias estaciones meteorológicas. Esta última opción quizás sea más recomendable para actualizaciones periódicas de datos, mientras que la importación de registros históricos largos (durante la carga inicial de la base de datos del CRC) será más fácil si los archivos están separados por estación. En cualquier caso, el formato de los archivos de texto debe ser el mismo.

A continuación se dan algunos detalles del formato que los países deberían observar para minimizar los errores en la importación de datos.

A. Si se transfieren datos para una única estación meteorológica, el nombre del archivo debería ser el ID internacional de la estación y la extensión ".csv" (aunque las columnas del archivo estén separados por tabulaciones). Por ejemplo el archivo "87544.csv" contendrá datos para la estación cuyo identificador es 87544 (Pehuajó, Argentina).

B. Si, en cambio, se transfieren datos para múltiples estaciones, por ejemplo como parte de una actualización periódica de datos, entonces el archivo debe tener el nombre abreviado de la institución y la extensión ".csv". Por ejemplo, un archivo que contenga los datos transferidos por el Servicio Meteorológico Nacional de Argentina para los últimos 10 días debería llamarse "SMN.csv". Los nombres abreviados para otras instituciones son "INMET" para Brasil, "DMH" para Paraguay, "INUMET" para Uruguay y "SENAMHI" para Bolivia.

C. El archivo debe contener un encabezado (o sea, la primera línea en el archivo) con los nombres abreviados de las variables en cada columna en el orden citado en la tabla A-1. El encabezado se usa para verificar la integridad de los datos durante la importación a la base de datos relacional, por lo que es necesario respetar estos nombres y su orden. Los nombres a utilizar en el encabezado se muestran en la columna "nombre abreviado" de la tabla A-1 ("omm_id", "fecha", etc.). Es indistinto usar mayúsculas o minúsculas en el encabezado: "FECHA" y "fecha" son ambos válidos.

D. Cada fila subsiguiente en el archivo representa las observaciones para un día y una estación meteorológica.

E. Cada columna corresponde a una variable. Las columnas deben estar separadas por una tabulación (tab o "\t"). El orden de las columnas debe ser el indicado en la tabla A-1. Aunque no se incluyan todas las variables listadas en la tabla A-1 (por ejemplo, algunos miembros solamente transfieren datos de temperaturas máxima y mínima y precipitación), todas las columnas sin datos deben llenarse con valores faltantes (ver ítem siguiente).

F. Los valores faltantes para una variable se deben indicar como "\N" (sin las comillas), que será interpretado por la base de datos como un valor NULL o faltante. También es válido dejar en blanco (vacío) el valor faltante de una variable. En el archivo de texto, un espacio vacío se indica como dos tabulaciones consecutivas "\t\t" (que separan la columna anterior y la posterior a la columna en blanco). Sin embargo, NO debe incluirse un espacio en blanco como dato faltante –es decir, debe evitarse "\t \t" (notar el espacio entre las dos tabulaciones).

G. Debe utilizarse el punto "." (no la coma ",") como separador de decimales en los valores de las variables. No debe usarse coma para separar miles (por ejemplo, NO usar "1,013.7" sino "1013.7").

H. Las fechas pueden listarse (a) en formato ISO 8601, o sea, el año, mes y día están separados por un guión medio (AAAA-MM-DD); o (b) como día, mes y año separados por una barra (DD/MM/AAAA).

Apéndice B

B1. Información incluida en la versión *preliminar* de los metadatos para cada estación meteorológica en la base de datos del Centro Regional Climático para el Sur de América del Sur

Tabla B-1. Listado de datos a incluir en el archivo de metadatos para cada estación meteorológica a incluirse en la base de datos del CRC-SAS.

NOMBRE ABREVIADO	NOMBRE ABREVIADO
omm_id	Número internacional de la estación (código OMM)
nombre	Nombre de la estación
lon_grad	Longitud (grados sexagesimales)
lon_min	Longitud (minutos sexagesimales)
lon_seg	Longitud (segundos sexagesimales)
lon_hem	Hemisferio longitudinal (E/O)
lat_grad	Latitud (grados sexagesimales)
lat_min	Latitud (minutos sexagesimales)
lat_seg	Latitud (segundos sexagesimales)
lat_hem	Hemisferio latitudinal (N/S)
lon_dec	Longitud (grados decimales)
lat_dec	Latitud (grados decimales)

elev	Elevación sobre el nivel del mar (m)
pais_iso2c	Código alfabético del país (estándar ISO 3166-1 alpha-2)
nivel_adm_1	Nivel administrativo 1 (estado, provincia)
nivel_adm_2	Nivel administrativo 2 (departamento, partido)
fecha_inicio	Fecha de inicio de operación de la estación
fecha_fin	Fecha de fin de operación de la estación
activa	¿La estación está activa en la actualidad? (sí/no)
operador	Código de institución que opera la estación (ej., “1” por INMET)
id_interno	Número interno de la estación dentro de la organización
tipo_estacion	Código de tipo de estación meteorológica (C: convencional, A: automática, etc.)

Algunas consideraciones sobre el archivo de metadatos:

- El archivo de metadatos debe tener codificación UTF-8, de modo de poder almacenar acentos y otros caracteres especiales;

- Valores negativos de latitud y longitud decimal indican, respectivamente hemisferios S y W;

- La fecha de inicio de las observaciones en una estación meteorológica no necesariamente debe coincidir con la primera fecha para la cual se transfieren datos. La estación puede haber comenzado a funcionar en 1947, por ejemplo, pero se contribuyen datos desde principios de 1961; y

- La fecha de fin de las observaciones solo debe llenarse en el caso de estaciones meteorológicas que hayan dejado de tomar observaciones. Esta fecha indica cuándo ha dejado de operar una estación meteorológica. Por lo tanto, para estaciones que estén en actividad este campo debe estar vacío (o lleno con “\N”). Este campo NO debe llenarse con la fecha de la última observación transferida a la base de datos del CRC.