Centro Regional del Clima para el Sur de América del Sur

Centro Regional do Clima para o Sul da América do Sul

Serie Reportes Técnicos – Reporte Técnico CRC-SAS-2014-001

Descripción de controles de calidad de datos climáticos diarios implementados por el Centro Regional del Clima para el Sur de América del Sur

Hernán Veiga

Natalia Herrera

María de los Milagros Skansi

Servicio Meteorológico Nacional, Buenos Aires, Argentina

Guillermo Podestá

University of Miami, Rosenstiel School of Marine and Atmospheric Science, Miami, USA

Última actualización: 2015-12-15

1 Introducción

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

Uno de los principales objetivos a la hora de construir una base de datos climáticos regionales es desarrollar un esquema de control de calidad de los datos. A través de este esquema de controles se intenta identificar datos erróneos o dudosos posiblemente relacionados con (a) problemas en la observación realizada por el observador meteorológico y/o (b) errores de digitalización o transcripción de la información. El propósito de este documento es describir la compilación y control de calidad de una base de datos de variables climáticas diarias. Los datos son compilados por las instituciones participantes en el Centro Regional del Clima para el sur de América del Sur.

El Centro Regional del Clima para el sur de América del Sur (en adelante, CRC-SAS) es un esfuerzo liderado por los servicios meteorológicos e hidrológicos de Brasil y Argentina como parte del Marco Mundial para los Servicios Climáticos (High-level taskforce for the Global Framework for Climate Services, 2011) de la Organización Meteorológica Mundial (OMM) de la Organización de Naciones Unidas. El CRC-SAS involucra además la participación de Paraguay, Uruguay, Bolivia y Chile. 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 cambio climático.

2 Datos climáticos

Una de las primeras actividades del CRC-SAS es la recopilación de datos diarios para una serie de variables meteorológicas (Tabla 1) para el período desde el 1 de enero de 1961 hasta el presente. Como un paso inicial, se están compilando datos provenientes de estaciones meteorológicas convencionales operadas por los seis países miembros del CRC-SAS. Las estaciones convencionales deben cumplir con ciertos requisitos: deben estar equipadas con instrumentos manuales operados por un observador meteorológico, el cual se debe encargar de efectuar mediciones meteorológicas y puede también ser el encargado del funcionamiento y mantenimiento del instrumental si se le proporciona la formación adecuada (ftp://ftp.wmo.int/Documents/MediaPublic/Publications/WMO488_GOSguide/488_2012_es.pdf).

La cobertura geográfica de la base de datos del CRC-SAS incluye las estaciones convencionales de Uruguay, Argentina, Paraguay, Bolivia y Chile, y las estaciones de Brasil ubicadas al sur del paralelo 10°S. Las variables meteorológicas que se incluyen en la base de datos se muestran en la Tabla 1. Hay tres variables (temperatura máxima y mínima diaria, precipitación acumulada) que son contribuidas por *todos* los miembros del CRC-SAS que operan estaciones meteorológicas convencionales. Las restantes variables incluidas en la tabla representan cantidades útiles para el cálculo de variables climáticas derivadas (por ej., 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.

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	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

3 Organización general de los 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 los datos originales. Por esta razón, el CRC-SAS– en colaboración con un proyecto de investigación financiado por el Instituto Interamericano para la Investigación del Cambio Global y por el Banco Interamericano de Desarrollo – ha implementado una serie de procedimientos publicados en la literatura científica para el control de calidad de 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 el Reporte Técnico CRC-SAS-2013-001. En este documento, 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 de distintos tipos (ver Sección 3.1). Estos controles identifican registros que contienen variables meteorológicas con valores sospechosos. Una segunda etapa del control de calidad involucra la verificación manual de todos los valores sospechosos identificados en la etapa anterior. Un operador verifica los datos sospechosos utilizando los registros originales, registrados en papel o en formato digital pero con mayor resolución temporal (por ejemplo, datos horarios a partir de los cuales se calculan los valores diarios que están siendo controlados). Por ejemplo, si es necesario corroborar el valor de temperatura máxima para un día determinado, se puede analizar el valor diario anotado en la libreta meteorológica (el registro diario de la estación, generalmente un documento en papel) o se pueden controlar las observaciones horarias para ese día (sea en papel o en formato digital) para determinar la veracidad del dato dudoso.

3.1 Grupos de controles de calidad

Los controles de calidad se organizaron en seis familias o grupos que agrupan controles de características similares (Figura 1). Los detalles de los controles incluidos en estas familias se discuten en secciones subsiguientes, pero en esta sección se presenta una breve descripción de cada grupo.

- **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. Otra verificación que se realiza es 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 se han implementado como parte del proceso de actualización periódica de datos (por ejemplo, se identificará la repetición de fechas antes de actualizar la información para una estación en la base de datos del CRC-SAS).
- **Controles de rango fijo.** Estos controles aseguran que no existan valores físicamente imposibles o nunca antes observados en el registro histórico. Los límites propuestos son fijos para cada variable durante todo el periodo de datos y todas las estaciones meteorológicas. Por ejemplo, una temperatura máxima diaria de 99°C está por encima del record mundial. Es posible que un valor así corresponda a un código de valor faltante que no se ha definido apropiadamente como tal.
- **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

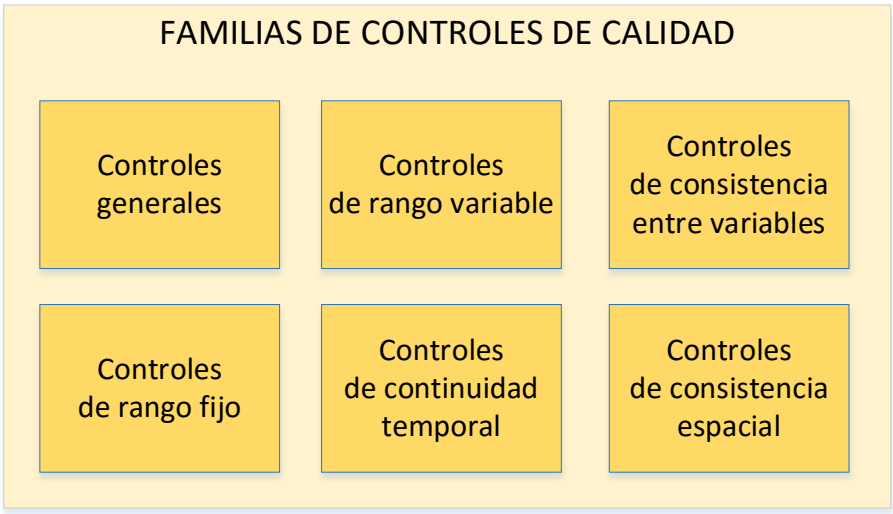


Figura 1. Familias de controles de calidad. Dentro de cada familia existen varios controles que se aplican a diferentes variables.

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.

- **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 a 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.
- **Controles de consistencia entre variables.** Una serie de controles en esta familia o grupo evalúan la consistencia entre valores de pares o grupos 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.
- **Controles de consistencia espacial.** Todos los controles descritos 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, sin embargo, los valores de una variable para una estación determinada (que generalmente se denomina la "estación central") se comparan con valores de esa variable registrados en estaciones geográficamente cercanas (o "estaciones vecinas").

Dentro de cada familia de controles puede haber varios controles basados en distintos cálculos. Además, no todos los controles se aplican a todas las variables. Hay controles, por ejemplo, que se utilizan solamente para datos de precipitación. La Tabla 2 lista los controles incluidos hasta el momento en cada familia de controles, y las variables a las cuales se aplica cada control.

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

	Variables Meteorológicas Diarias												
	tmax	tmin	tmed	td	pres.est	pres.mar	Prcp	hr	helio	nub	vmax.d	vmax.f	vmed
1. Controles de rango fijo													
1.1 Límites inferiores y superiores constantes	●	●	●	●	●	●	●	●	●	●	●	●	●
2. Controles de rango variable													
2.1 Desviaciones respecto al ciclo estacional	●	●	●	●	●	●		●					
2.2 Desviaciones respecto a múltiplos del rango intercuartil para ventanas de 3 o 5 días	●	●	●	●	●	●							
2.3 Desviaciones respecto a estadísticos robustos (biweight) para ventanas de 3 o 5 días	●	●	●	●	●	●							
2.4 Heliofanía menor o igual a la heliofanía teórica astronómica									●				
2.5 Desviaciones respecto al rango intercuartil para ventanas mensuales							●						
2.6 Valores extremos de precipitación (cuantiles de distribución gamma)							●						
3. Controles de continuidad temporal													
3.1 Persistencia de valores constantes por 3 o más días consecutivos	●	●	●	●	●	●	●	●	●	●	●	●	●
3.2 Persistencia extrema (definida climatológicamente) de días sin precipitación							●						
3.3 Saltos excesivos entre días consecutivos	●	●	●	●		●							
3.4 Picos de corta duración (1 día)	●	●	●	●		●							

	Variables Meteorológicas Diarias												
	tmax	tmin	tmed	td	pres.est	pres.mar	Prcp	hr	helio	nub	vmax.d	vmax.f	vmed
4. Controles de consistencia entre variables													
4.1 Tmin(día i) < Tmed(día i) < Tmax (día i)	●	●	●										
4.2 Tmed ≈ (tmax+tmin) / 2	●	●	●										
4.3 Tmin (día i-1) ≤Tmax (día i) ≥ Tmin (día i+1)	●	●											
4.4 Tmax (día i-1) ≥ Tmin (día i) ≤ Tmax (dia i+1)	●	●											
4.5 Td ≤ Tmed	●			●									
4.6 Tmax - Tmin (amplitud térmica diaria)	●	●											
4.7 pres.est ≤ pres.mar					●	●							
4.8 vmax.d = 0 y vmax.f = 0											●	●	
4.9 vmax.d ≠ 0 y vmax.f ≠ 0											●	●	
4.10 vmed ≤ vmax												●	●
4.11 prcp = 0 y nub ≠ 0							●			●			
4.12 helio > 0 y nub = 8									●	●			
5. Controles de consistencia espacial (entre estaciones)													
5.1 Control de regresión espacial ponderada	●	●	●	●									
5.2 Control de regresión espacial basado en índice de concordancia	●	●	●	●									
5.3 Control de corroboración espacial para temperatura	●	●	●	●									
5.4 Control de corroboración espacial para precipitación							●						

3.2 Implementación de los controles de calidad

Todos los controles de calidad se implementaron en el lenguaje R (R Core Team, 2013), un entorno de programación diseñado para realizar análisis estadísticos y visualizar datos. El R es 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 ahorran tener que programar todos los cálculos deseados, reduciendo el tiempo de implementación y la probabilidad de errores de programación. Por ejemplo, existen al menos dos paquetes que permiten 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 3.1. 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 una tarea repetida en una función (por ejemplo, calcular la media y desvío estándar robusta de una variable usando la función de ponderación *biweight*) permite evitar la duplicación de código y, por lo tanto, disminuir la chance de errores y facilitar el mantenimiento de los scripts.

3.3 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 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 sospechosas y por lo tanto ambas variables reciben el valor FALSE para el 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 buscó una detección eficiente de errores, buscando balancear la relación entre el tiempo que se utiliza en la verificación manual de los datos dudosos y la tasa de detección de registros realmente erróneos. Estos aspectos se discuten en detalle más adelante en este documento.

4 Descripción detallada de los controles de rango fijo

La primera familia de controles – controles de rango fijo – compara el valor de una variable meteorológica con valores extremos pre-establecidos. 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 -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 utilizaron valores extremos históricos para Argentina (-39.0°C, 49.1°C), y 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.

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

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

Variable	**Intervalo válido**
Temperaturas	-39°C ≤ temperaturas ≤ 49.1°C
Presión	530 hPa ≤ presión ≤ 1060 hPa
Precipitación	0 ≤ prcp ≤ 200 mm
Humedad relativa	0 ≤ hr ≤ 100
Heliofanía	0 ≤ Heliofanía ≤ 24
Nubosidad	0 ≤ Nubosidad ≤ 9
Dirección de viento máximo diario	0 ≤ vmax.d ≤ 36 (Direcciones en decenas de grado, redondeados a valores enteros)
Velocidad de viento máximo diario	0 ≤ vmax.f ≤ 120 nudos (≈ 62 m s^{-1})
Velocidad de viento medio	0 ≤ vmed ≤ 50 nudos (≈ 26 m s^{-1})

5 Descripción detallada de los controles de rango variable

A diferencia de los controles de rango fijo que utilizan un intervalo constante para detectar datos posiblemente erróneos, los intervalos usados en los controles de rango variable para aceptar o rechazar datos 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. La diferencia es que el rango válido de valores se define para cada día o mes del año. Las secciones siguientes discuten en detalle los diferentes controles dentro de esta familia.

5.1 Desviaciones respecto al ciclo estacional

Dado que muchas variables meteorológicas presentan un comportamiento regular o repetido estacionalmente, se implementó 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 (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 identifica como sospechosos a los valores de una variable que muestren desvíos extremos con respecto al ciclo estacional ajustado. Los desvíos o residuos extremos se definen en términos de percentiles estimados a partir de estos desvíos. Por ejemplo, para valores excepcionalmente bajos y altos de una variable se pueden utilizar los percentiles 1 y 99, respectivamente ($perc_{01}$ y $perc_{99}$). Los percentiles pueden estimarse (a) para cada día del año, o (b) para cada mes. En este caso, los percentiles de los desvíos se calculan para cada mes.

Para describir el funcionamiento de este control, se presenta un ejemplo basado en observaciones de temperatura máxima diaria (Tmax) en Pehuajó (Provincia de Buenos Aires, Argentina). En el panel izquierdo de la Figura 2 se presentan los valores observados de Tmax en Pehuajó para 1961-2012 (puntos grises). El ciclo estacional ajustado se muestra como una línea azul. Los puntos rojos señalan datos potencialmente erróneos, ya que muestran desvíos muy grandes (positivos y negativos) respecto al ciclo estacional. El panel derecho muestra boxplots (o diagramas de "cajas y bigotes") con la distribución de Tmax diarias para cada mes. Los bordes inferiores y superiores de las "cajas" en color amarillo corresponden a los percentiles 25 y 75 para cada mes; es decir, las cajas contienen el 50% central de los datos observados para cada mes. La línea gruesa horizontal dentro de cada caja indica la mediana de Tmax (el percentil 50). Los valores en los extremos de las líneas verticales o "bigotes" por encima y debajo de cada caja son identificados como sospechosos.

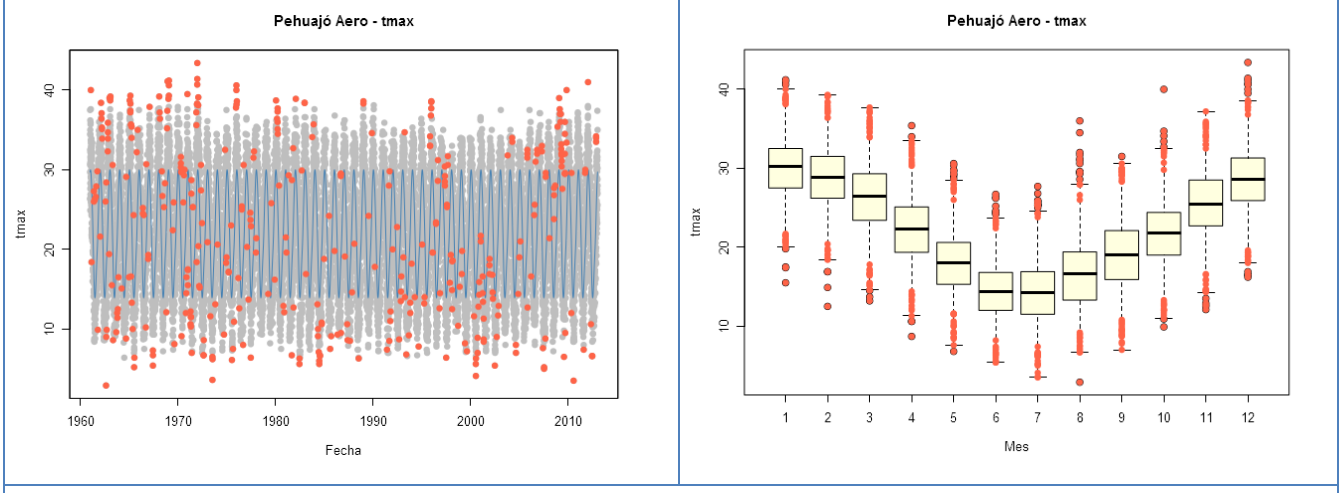


Figura 2. Temperatura máxima diaria en Pehuajó 1961-2012. Izquierda: Ciclo estacional estimado (línea azul), datos (puntos grises) y datos sospechosos (puntos rojos). Derecha: Boxplots mensuales (1961-2012) y datos sospechosos (puntos rojos).

Para ver más detalle, en la Figura 3 se presenta un ejemplo específico para el año 1962 en Pehuajó. Los valores de Tmax se representan con una línea roja. La curva negra representa el ciclo estacional ajustado. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que el control lo identifica como sospechoso. La Figura 4 muestra la serie de diferencias entre valores observados y el ciclo estacional para la serie en la Figura 3. La Figura 4 muestra, además, los límites de los rangos para este control definidos por tres pares de percentiles: [0.100, 0.900], [0.010, 0.990] y [0.001, 0.999]. Puede verse que los límites de cada rango varían temporalmente, ya que se estiman separadamente para cada mes. El valor previamente identificado como sospechoso cae fuera del intervalo más ancho (que deja 1/1000 de los desvíos fuera de cada extremo del rango aceptado). En realidad, el valor correcto para este día (verificado en los registros oficiales) fue de 12.9°C en vez de 2.9°C. En consecuencia, el valor sospechoso puede considerarse como realmente incorrecto.

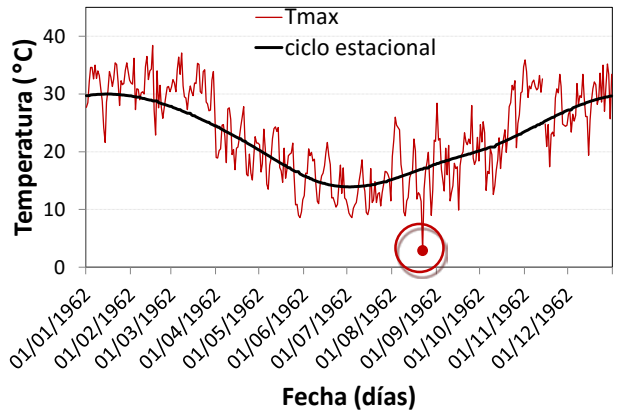


Figura 3. Temperatura máxima diaria observada en Pehuajó (rojo) y ciclo estacional estimado (negro) para el año 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que el control lo identifica como sospechoso.

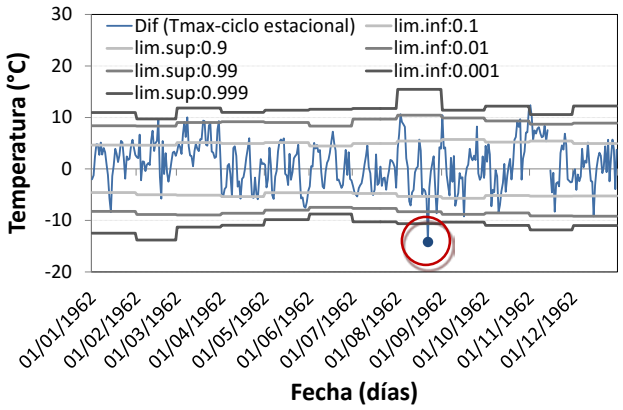


Figura 4. Diferencia entre la temperatura máxima diaria y el ciclo estacional en Pehuajó, año 1962. En tonos de grises se indican los distintos límites del rango aceptable para cada mes según el percentil escogido. El valor sospechoso cae fuera (debajo) del rango más amplio.

5.2 Desviaciones respecto al rango intercuartil para ventanas diarias de 3 o 5 días

Este control también identifica desvíos extremos con respecto a un intervalo aceptado. El intervalo, sin embargo, no se define en función de un ciclo estacional, sino en base a dos parámetros estadísticos: la mediana (*M*) de valores considerados y su seudo-desvío estándar (*SSD*). Tanto la mediana como el seudo-desvío estándar se estiman para una ventana temporal centrada alrededor del día del año a analizar (día *i*) y que incluye los valores para 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 como día *i* al 5 de enero, la ventana de 3 días considera a los datos del 4 al 6 de enero de todos los años disponibles (por ejemplo, 1961-2012); 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 extremos. El seudo-desvío estándar ssd se calcula de la siguiente forma:

$$ssd = ri / 1.349 \qquad (1)$$

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 conceptualmente y computacionalmente simples, y son recomendables para casos

en que no se necesite alta eficiencia, por ejemplo si hay una gran cantidad de datos disponibles o cuando el análisis es “altamente exploratorio” (Lanzante, 1996).

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

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

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

La Figura 5 presenta el mismo ejemplo que se usó en la Figura 3 (Tmax en Pehuajó, Argentina, para el año 1962). En este caso, sin embargo, la línea negra indica la mediana de valores de Tmax para cada día del año, estimada usando una ventana temporal de 5 días y 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 círculo rojo, indicando que también este control lo identifica como

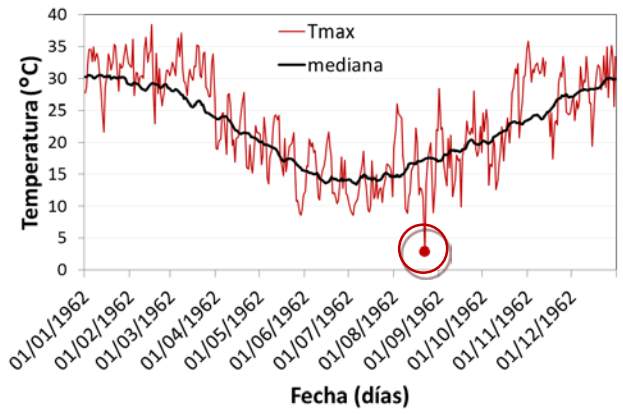


Figura 5. Temperatura máxima diaria observada en Pehuajó (rojo) y mediana de valores para cada día del año (negro), año 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que el control lo identifica como sospechoso.

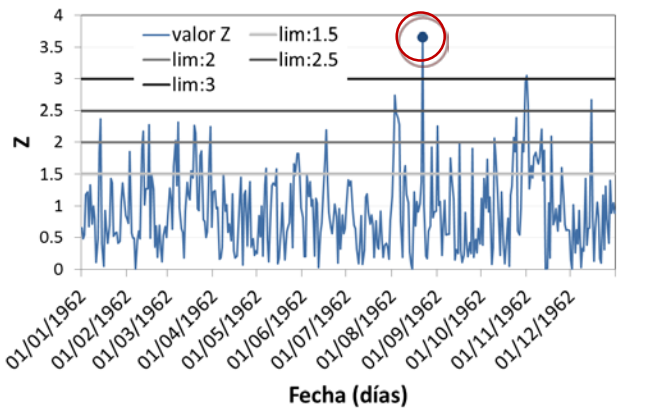


Figura 6. Estadístico Z calculado para cada día del año 1962 en Pehuajó. En tonos de grises se indican los distintos límites del rango aceptable. El valor sospechoso para el día 1962-08-22 cae fuera (encima) del rango más amplio.

sospechoso. El estadístico Z (Figura 6) para ese día fue mayor al límite de Z = 3. Recordamos al lector que el valor correcto para este día (verificado en los registros oficiales) fue de 12.9°C en vez de 2.9°C. En consecuencia, el valor identificado por este control es realmente incorrecto.

5.3 Desviaciones respecto a medias y desviaciones estándar robustas (método *biweight*) para ventanas diarias de 3 o 5 días

El cálculo de scores Z basado en la mediana y el rango intercuartil descrito en la Sección 5.2 es una mejora con respecto al simple uso de medias y desvíos estándar para identificar valores extremos. Sin embargo, cuando se desea mayor resistencia a la influencia de valores extremos se recomienda usar estimadores de tendencia y central y dispersión más complejos, tales como la media y desvío estándar calculados mediante la función *biweight* (Hoaglin, 1983). Las estimaciones basadas en el método *biweight* involucran un cálculo en dos pasos. En el primer paso se estiman la tendencia central y dispersión de los datos usando la mediana y el estadístico MAD (mediana de desvíos absolutos). Estos estimadores se usan solamente para descartar valores extremos, asignándoles un peso igual a cero en cálculos subsiguientes. En el segundo paso, se calculan la media y desvío estándar ponderados, en el cual los pesos utilizados disminuyen en forma no lineal con respecto al centro de la distribución de datos (Lanzante, 1996).

Los pesos $u_{i,j}$ para cada una de las observaciones $x_{i,j}$ para el día del año i y el año j se calculan como

$$u_{i,j} = (\mathrm{x}_{i,j} - \mathrm{M}_i) / (\mathrm{c} \times \mathrm{MAD}) \ , \qquad (3)$$

donde M_i y MAD_i indican la mediana de los datos y la mediana de desvíos absolutos estimados para el día del año i usando una ventana temporal de 5 días centrada alrededor de ese día. El parámetro c define a qué distancia del centro de los datos 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 a ± 5 desvíos estándar (Hoaglin, 1983). Para $\left| u_{i,j} \right| \geq$ 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 $\overline{x}_i^{bwt}$ y el desvío estándar biweight $\overline{s}_i^{bwt}$ para cada día del año:

$$\overline{x}_i^{bwt} = M_i + \frac{\sum_{i=1}^{n} \left(\mathrm{x}_{i,j} - M_i \right) \left(1 - u_i^2 \right)^2}{\sum_{i=1}^{n} (1 - u_i^2)^2} \ , \text{ y} \qquad (4)$$

$$S_i^{bwt} = \frac{\sqrt{n \sum_{i=1}^{n} \left(\mathrm{x}_{i,j} - M_i \right)^2 \left(1 - u_i^2 \right)^4}}{\left| \sum_{i=1}^{n} \left(1 - u_i^2 \right) \left(1 - 5u_i^2 \right) \right|} \ . \qquad (5)$$

Las estimaciones de $\overline{x}_i^{bwt}$ y $\overline{s}_i^{bwt}$ están más influenciadas por valores cercanos al centro de la distribución que por valores en las colas o extremos. Por esta razón, los valores estimados son más resistentes a la influencia de valores extremos (Feng et al., 2004). Con estos valores se calcula el estadístico Z a partir del cual se evalúa la calidad del dato:

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

Como en el control anterior, se identifican como sospechosos a los datos cuyos valores de Z sean mayores a un cierto umbral. En este caso se usa un umbral de *Z* = 3. Este control fue previamente utilizado en trabajos anteriores, por ejemplo los de 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 7 muestra con una línea negra la media biweight, estimada para cada día del año usando una ventana temporal de 5 días y 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 círculo rojo, indicando que este control lo identifica como sospechoso. El estadístico *Z* (El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que también este control lo identifica como sospechoso. La Figura 8 muestra que el valor del estadístico *Z* para ese día fue mayor al límite de Z = 3. Como en los dos controles de rango variable anteriores, este control también identifica correctamente el valor erróneo.

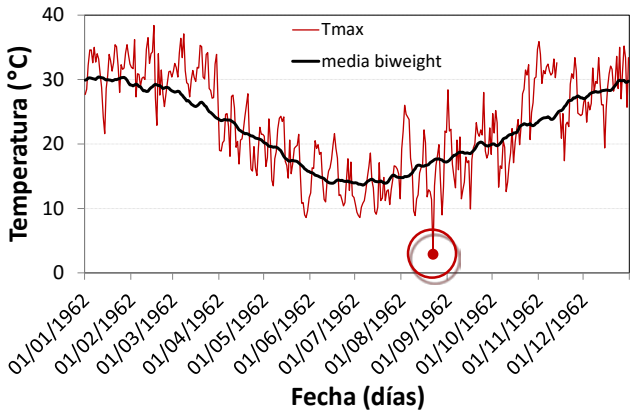


Figura 7. Temperatura máxima diaria observada en Pehuajó (rojo) y media *biweight* de valores para cada día de 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que el control lo identifica como sospechoso.

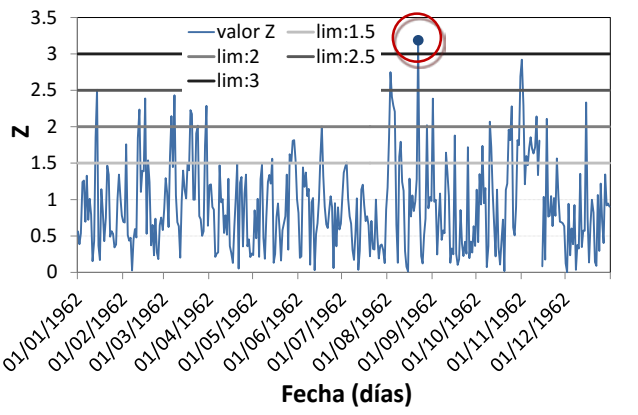


Figura 8. Estadístico Z calculado para cada día del año 1962 en Pehuajó. En tonos de grises se indican los distintos límites del rango aceptable. El valor sospechoso para el día 1962-08-22 cae fuera (encima) del rango más amplio.

5.4 Heliofanía menor o igual a 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 cada día del año. Como este valor representa un máximo para la heliofanía observada en cada día, se identifican como sospechosos a los registros de heliofanía que superen dicho umbral.

El cálculo de la heliofanía máxima para cada día del año puede realizarse con los paquetes de R `insol` o `geosphere`. Estos paquetes utilizan, respectivamente, los métodos descritos por Corripio (2003) y Forsythe et al. (1995). Para los controles descritos aquí utilizamos el método de Corripio (2003). En función de la latitud de cada estación meteorológica se calculó la heliofanía máxima para cada día del año. Para evitar errores debido a redondeo de valores reportados, la heliofanía máxima se multiplicó por 1.05 y este valor se usó como umbral en la identificación de valores sospechosos.

La Figura 9 ejemplifica el control de valores de heliofanía usando datos para Pehuajó, 1961-2012. Los tres puntos marcados en rojo exceden el umbral definido en el párrafo anterior, y por lo tanto considerados sospechosos. Dos de ellos (los desvíos más pronunciados) eran erróneos y se los pudo corregir con el dato que figuraba en la libreta meteorológica. El tercer punto, si bien excede la duración teórica del día, está dentro del margen tolerado.

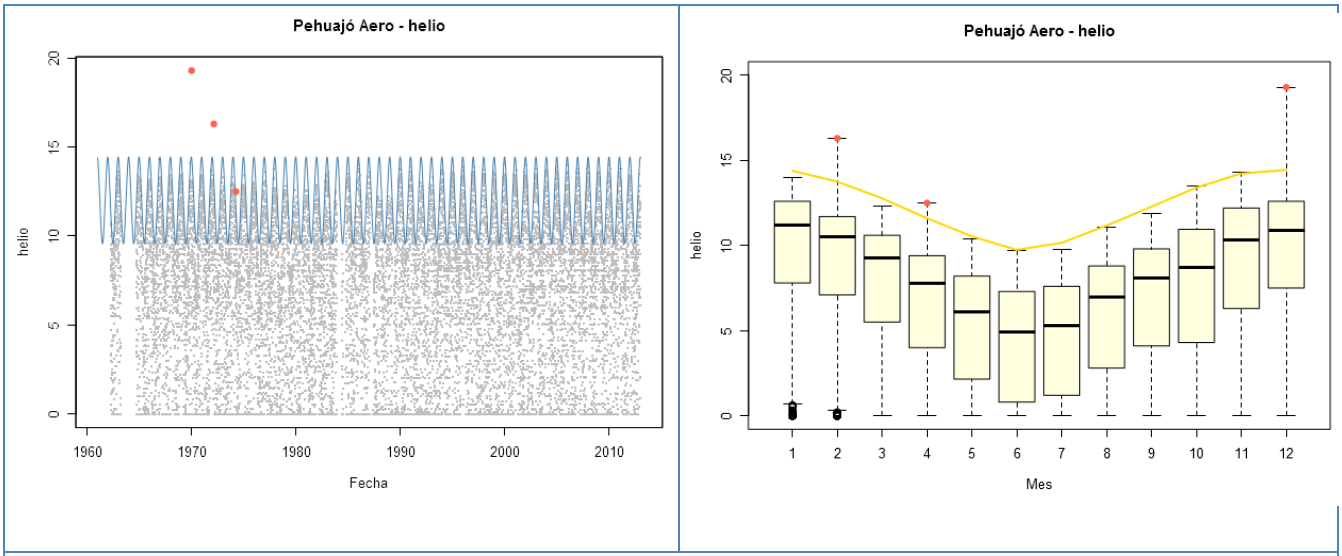


Figura 9. Heliofanía diaria en Pehuajó 1961-2012. Panel izquierdo: Heliofanía máxima teórica para cada día del año (línea azul), valores de heliofanía reportados (puntos grises) y datos sospechosos (puntos rojos). Panel derecho: Boxplots mensuales de heliofanía (1961-2012) y datos sospechosos (puntos rojos). La línea amarilla representa la heliofanía máxima promedio para cada mes.

5.5 Desviaciones respecto al rango intercuartil de precipitación para ventanas mensuales

El diseño de controles de calidad para datos diarios de precipitación ha sido tradicionalmente difícil (Hubbard et al., 2012). Este control es análogo al descrito en la Sección 5.2, pero se utiliza solamente para el control de valores de precipitación. Dada la existencia de muchos días sin precipitación y la posible gran variabilidad intra-mensual en los montos diarios de lluvia, la estimación de la dispersión (el rango intercuartil) de 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 a los valores diarios de precipitación que excedan un umbral PS_i estimado separadamente para cada mes i. El umbral para cada mes se calcula de la siguiente forma:

$$PS_i = perc75_i + (\mathrm{n} \times ri_i) \quad , \qquad (7)$$

donde $perc75_i$ es el percentil 75 (es decir, el tercer cuartil) de los valores diarios de precipitación > 0.1 mm para 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 como sospechosos.

Este control etiqueta como sospechosos a todos los valores de lluvia diarios que superan el umbral PS para el

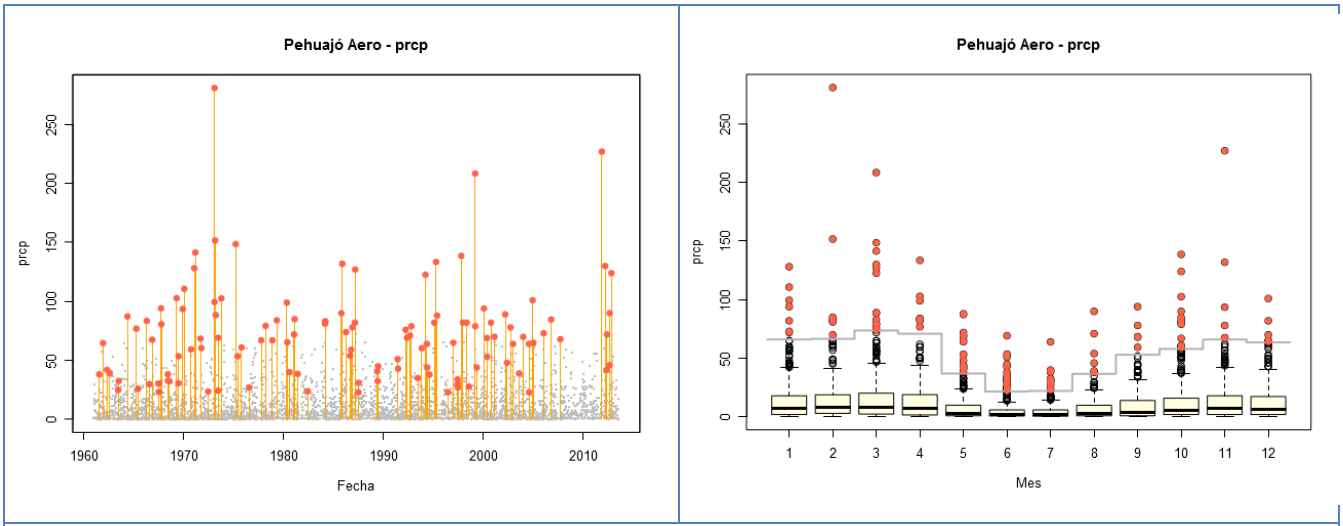


Figura 10. Precipitaciones diarias en Pehuajó 1961-2012. Panel izquierdo: Lluvias diarias (puntos grises) y datos sospechosos (puntos rojos con líneas naranja). Panel derecho: Boxplots mensuales de precipitación diaria (1961-2012) y datos sospechosos (puntos rojos). La línea horizontal gris indica el umbral utilizado en cada mes para definir valores sospechosos. En ambos paneles, solamente se utilizan los valores diarios de precipitación > 0.1 mm.

mes correspondiente. Un ejemplo de este control se ilustra en la ajuste del control se presenta en la Figura 10. Este tipo de control fue utilizado por previamente por González-Rouco et al. (2001) y por Aguilar y Prudhom en el software CLIMDEX EXTRAQC (documento disponible en http://www.c3.urv.cat/data/manual/Manual_rclimdex_extraQC.r.pdf).

5.6 Identificación de valores extremos mensuales de precipitación mediante ajuste de una distribución Gamma

Como en la Sección 5.5, este control identifica como sospechosos a los valores diarios de precipitación que excedan un umbral. Sin embargo, el umbral no se define a través de percentiles empíricos – como en el control anterior, sino en base a percentiles calculados en base al 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 para determinar el valor a partir del cual se pueda identifican a las precipitaciones diarias como sospechosas.

En este control se separan todos 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 usando el método de L-momentos (Vicente-Serrano, 2006). Las distribuciones se ajustan usando solamente precipitaciones diarias mayores a 0.1 mm, o sea se descartan los días sin precipitación o con precipitaciones imperceptibles en el registro histórico.

Luego de ajustar los parámetros de la distribución gamma, se define un umbral a partir del cual las precipitaciones diarias se consideran extremadamente altas y, en consecuencia, sospechosas. El umbral puede definirse en base a un percentil calculado a partir de la distribución teórica (es decir, usando los parámetros estimados anteriormente). Este percentil, entonces, no se deriva de la distribución empírica de valores, como se ha hecho en controles anteriores. Específicamente, se identifican valores sospechosos de lluvia mediante

$$\mathrm{prcp}_{i,j} > Q_i^P \quad , \qquad (8)$$

donde $\mathrm{prcp}_{i,j}$ es la precipitación en el día i del año j y Q_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 cuestión. Este control fue previamente utilizado por Hubbard et al. (2012).

Un ejemplo de este control se ilustra en la ajuste del control se presenta en la Figura 11 utilizando datos para los meses de enero, febrero y marzo en Río Cuarto, provincia de Córdoba, Argentina. Las líneas cortas verticales a lo largo del eje x de cada panel corresponden a los valores observados de lluvias diarias > 0.1 mm. Las líneas verticales grises indican posibles umbrales para la definición de precipitaciones extremas; las líneas corresponden, de izquierda a derecha, 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 a precipitaciones aproximadamente mayores que 80 mm (el percentil 0.995). Las líneas verticales cortas a lo largo del eje (que representan los valores observados) sugieren que hay por lo menos 3 días que podrían considerarse sospechosos.

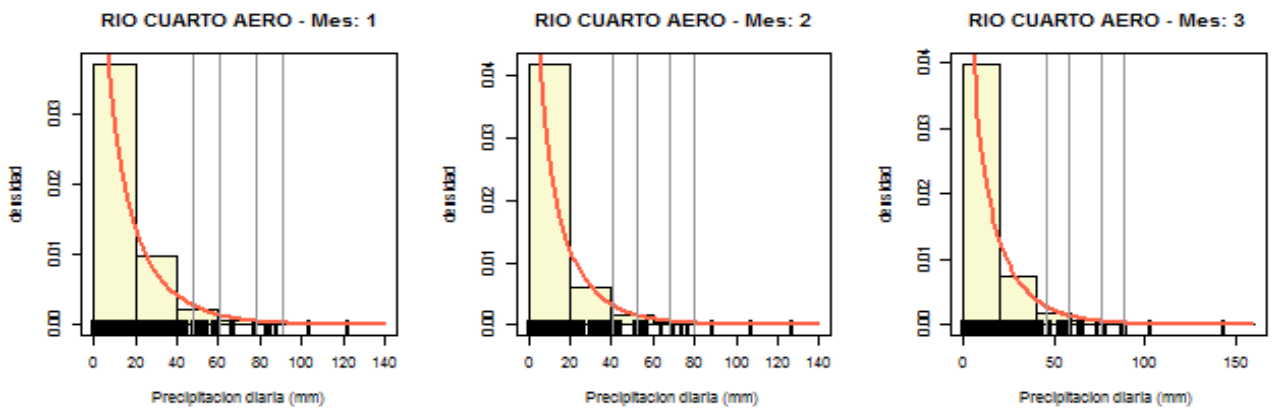


Figura 11. Histogramas de precipitaciones diarias para enero, febrero y marzo en Río Cuarto, Córdoba, Argentina. La figura muestra histogramas de valores diarios de precipitación > 0.1 mm para cada mes. Los valores de lluvia también se indican como líneas verticales cortas a lo largo del eje x de cada figura. La línea naranja corresponde a la distribución gamma ajustada. Las líneas verticales en gris indican posibles umbrales de precipitaciones extremas basados en percentiles 0.950, 0.975, 0.990 y 0.995.

6 Descripción detallada de los controles de continuidad temporal

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

6.1 Persistencia de valores constantes por N días consecutivos

Este control intenta detectar secuencias de valores repetidos a lo largo de varios días consecutivos. La persistencia del mismo valor puede sugerir errores de transcripción o problemas en el caso de instrumentos con registro electrónico de datos, por ejemplo en estaciones meteorológicas automáticas – no consideradas en esta etapa (Estévez et al., 2011). La persistencia de valores constantes por varios días consecutivos es posible, pero necesita ser verificada para determinar si esa secuencia es válida o inválida. Este tipo de controles fue utilizado por Meek y Hatfield (1994), Durre et al. (2010), Estévez et al. (2011) y por Aguilar y Prudhom en el software CLIMDEX EXTRAQC (ver http://www.c3.urv.cat/data/manual/Manual_rclimdex_extraQC.r.pdf).

En este control se identifican secuencias de 3 o más días con valores idénticos. Este control se aplica a todas las variables en la base de datos del CRC-SAS, excluyendo las precipitaciones iguales a 0.0 mm (sin precipitación). Para todas las secuencias con valores repetidos más de 3 días consecutivos, se marcan todos los valores en la secuencia como sospechosos. Como ejemplo de este tipo de controles, la Figura 12 muestra la frecuencia de

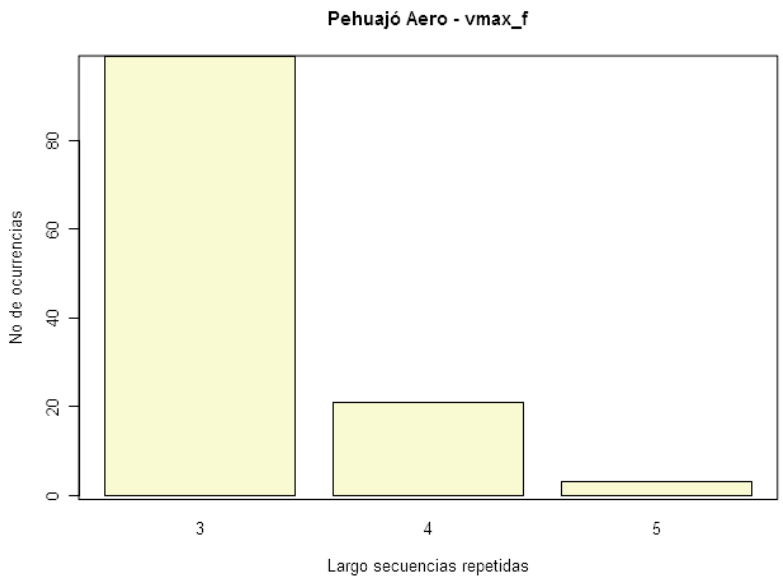


Figura 12. Número de ocurrencias de secuencias con valores de velocidad de viento máximo diario repetidos durante 3, 4 y 5 días en Pehuajó, Argentina. En los datos 1961-2012 no hay secuencias repetidas más largas que las incluidas en la figura.

ocurrencias de valores idénticos consecutivos para la serie de velocidad de viento máximo diario en Pehuajó, Buenos Aires, Argentina.

6.2 Persistencia extrema de días sin precipitación

El control descrito en la Sección 6.1 excluye valores de precipitación < 0.1 mm ya que la persistencia de días sin lluvia no puede necesariamente considerarse una secuencia incorrecta. Este control, por el contrario, se enfoca en secuencias de días sin lluvia (precipitación diaria < 0.1 mm). El propósito de este control es detectar un error común al migrar o transcribir datos: el error consiste en no incorporar datos existentes de precipitación y registrarlos, en cambio con 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 de días sin precipitación puede considerarse extrema y, por tanto, sospechosa. El umbral puede definirse en base a un cierto percentil (por ejemplo, 0.995) estimado a partir de los largos observados 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. Usando todos los datos en el registro histórico, se calcula el largo de todas las secuencias secas que comiencen en un mes determinado. En bases a esos largos, 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 y Leggieri (2008) y aplicado por Boulanger et al. (2010).

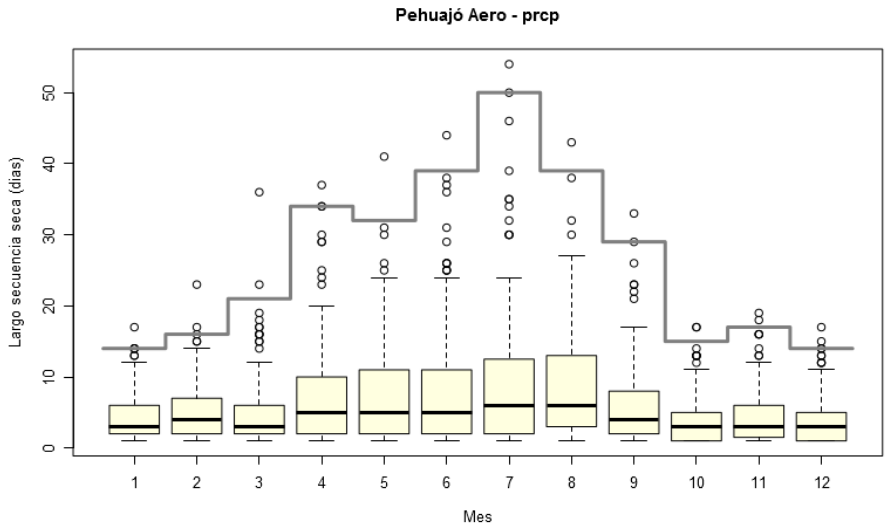


Figura 13. Boxplots de la distribución de largos de secuencias de días sin precipitación en Pehuajó en el periodo 1961-2012. La línea escalonada gris indica que el umbral para identificación de largos extremos varía a lo largo del año. Durante los meses con mayor precipitación (por ej., diciembre, enero), el umbral está alrededor de los 14-16 días, mientras que en los meses más secos, las secuencias secas deben exceder alrededor de 40 días para ser consideradas extremas.

Por ejemplo en la Figura 13 se observa la distribución mensual del largo de secuencias de días sin precipitación en Pehuajó en el periodo 1961-2012. Puede verse claramente que el umbral para identificación de largos extremos varía a lo largo del año. Aunque Pehuajó no es una de las estaciones con estacionalidad de la precipitación más marcada, durante los meses más lluvioso (por ej., diciembre, enero) el umbral está alrededor de los 14-16 días, mientras que en los meses más secos durante el invierno, las secuencias sin lluvia deben exceder alrededor de 40 días para ser consideradas extremas.

6.3 Saltos excesivos entre días consecutivos

Este control está diseñado para identificar saltos o diferencias extremas entre los valores de días consecutivos. Es decir, el control apunta a encontrar valores inusualmente altos o bajos respecto al registro del día anterior. El primer paso en el control es la creación de una serie temporal de diferencias absolutas ΔT de temperaturas entre un día y el día inmediatamente anterior:

$$\Delta T = \left| T_i - T_{i-1} \right| , \qquad (9)$$

donde se usa la notación genérica T para todas las temperaturas incluidas en la base de datos (mínima, máxima media y de rocío).

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 $perc^T$ (por ejemplo, el percentil 0.995) que se estima a partir de la distribución histórica de valores absolutos de diferencias observadas. Específicamente, se identifican diferencias extremas de temperatura mediante

$$\Delta T > \text{perc}^T \quad . \tag{10}$$

La implementación del control se ilustra en la Figura 14, 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 sospechosa. La Figura 15 y la Figura 16 ilustran el uso de este control para la identificación del valor erróneo para el 22 de agosto de 1982. Otro ejemplo de la implementación de este control se puede encontrar en Kunkel et al. (1998).

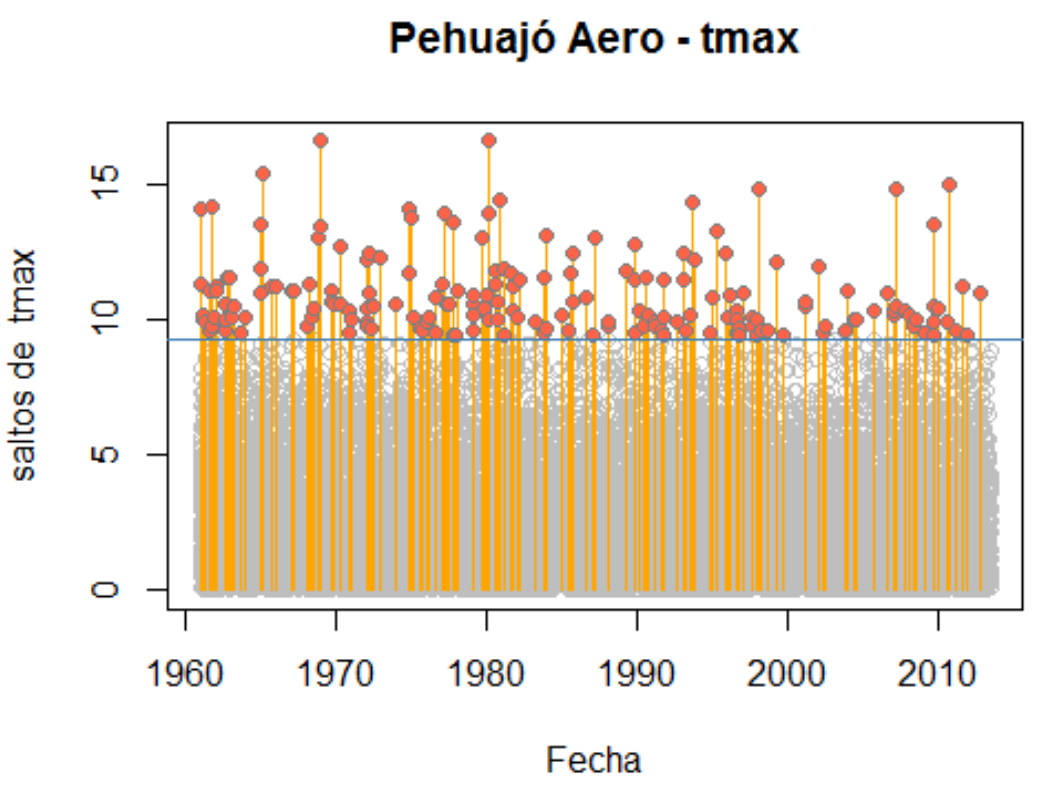


Figura 14. Diferencias absolutas de temperatura máxima en dos días consecutivos para Pehuajó, 1961-2012. La línea horizontal azul indica el percentil 0.99, utilizado para la identificación de saltos extremos. Los puntos rojos unidos al origen del eje x por líneas anaranjadas indican saltos extremos.

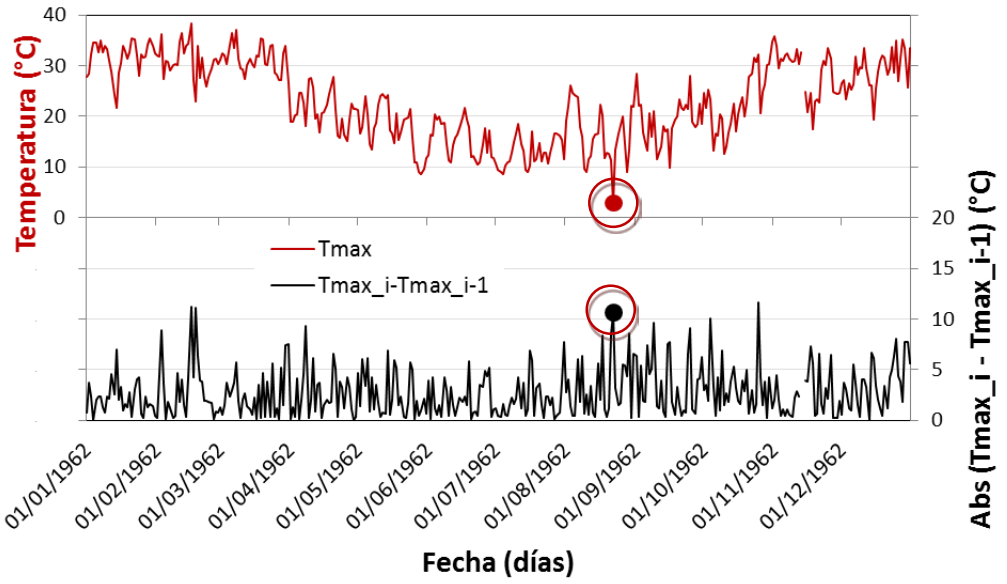


Figura 15. Panel superior: Temperatura máxima (Tmax) diaria observada en Pehuajó (rojo) para cada día de 1962. El valor de Tmax para el día 1962-08-22 (2.9 °C) está resaltado con un círculo rojo, indicando que el control lo identifica como sospechoso. La escala de temperaturas debe leerse en el margen 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 es extrema y coincide con el valor aparentemente bajo de Tmax en el panel superior.

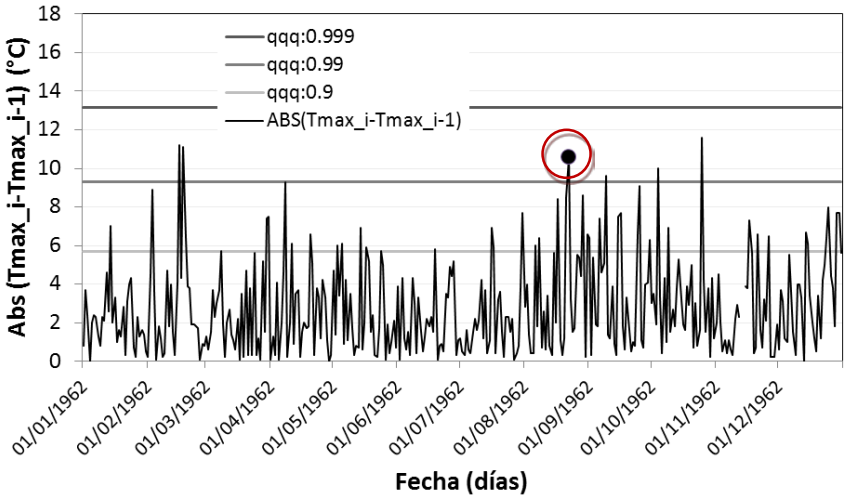


Figura 16. Valor absoluto de las diferencias entre Tmax en dos días consecutivos en Pehuajó para cada día de 1962. Las líneas horizontales indican diferentes umbrales posibles para la identificación de diferencias extremas basados en diferentes percentiles (0.900, 0.990 y 0.999). La diferencia de temperatura 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.

6.4 Picos extremos en días consecutivos

Este control se enfoca en la identificación de picos extremos en las series de temperaturas (mínima, máxima, media y de rocío). Se define un "pico" como un valor de temperatura muy diferente (mucho 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 temperaturas entre un día i y los días inmediatamente anterior y posterior:

$$\Delta T_1 = \left| T_i - T_{i-1} \right|, \text{ y} \tag{11}$$

$$\Delta T_2 = \left| T_{i+1} - T_i \right|, \tag{12}$$

donde se usa la notación genérica T para todas las temperaturas incluidas en la base de datos (mínima, máxima media y de rocío).

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 6.3. 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 base a un percentil empírico $perc^T$ (por ejemplo, el percentil 0.995) que se estima a partir de la distribución histórica de valores absolutos de diferencias observadas. Aquellos valores para los cuales *ambas* diferencias de temperatura ΔT_1 y ΔT_2 sean mayores que el umbral se identifican como sospechosos.

La implementación del control se ilustra en la Figura 17 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 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).

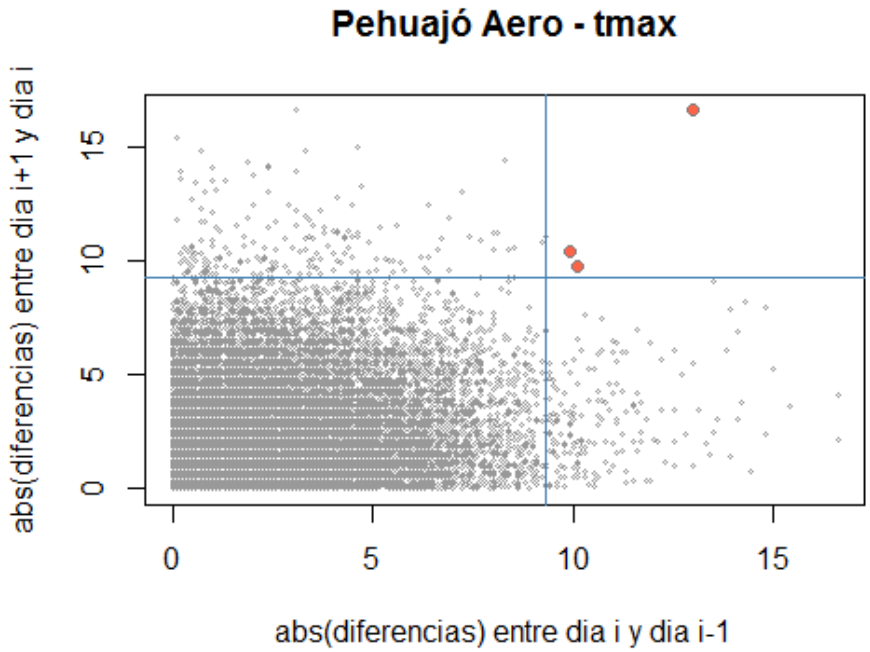


Figura 17. Diferencias absolutas de temperatura máxima entre un día y los días anterior (eje x) y siguiente (eje y) para Pehuajó, 1961-2012. En cada eje se indica el valor del umbral que define diferencias extremas (en este caso, el percentil 0.99). Los puntos para los cuales las dos diferencias de temperatura exceden los dos umbrales (marcados en rojo en el cuadrante superior derecho) se consideran "picos" extremos.

7 Descripción detallada de los controles de consistencia entre variables

Estos controles se basan en las relaciones teóricas que existen entre las distintas variables que forman parte de la base de datos del CRC-SAS. Este tipo de controles es ampliamente utilizado en distintos sistemas de control de calidad, dado que permite detectar inconsistencias o incoherencias entre variables incluidas en las bases de datos. Los diferentes controles que forman parte de esta familia se discuten en secciones subsiguientes. Sin embargo, es posible encontrar más discusiones en Meek y Hatfield (1994), Kunkel et al. (1998) y Estévez et al. (2011).

7.1 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 para el 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 temperatura mínima, media y máxima diarias. 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 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. 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 de la temperatura del aire 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: $\mathrm{Tmed}_i^* = (\mathrm{Tmax}_i + \mathrm{Tmin}_i)/2$. 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 pequeña tolerancia en los valores de temperatura calculados de diferentes maneras, en este control se identifican como sospechosos a los valores de temperatura media en los cuales la diferencia absoluta entre los dos valores de temperatura media sea superior al percentil 0.999 de todas las diferencias diarias.

Consistencia entre las temperaturas máximas y mínimas de días consecutivos. Dada la continuidad que generalmente presentan los valores de temperatura en días sucesivos, se pueden plantear los siguientes criterios para evaluar la consistencia de los registros de temperatura de 3 días consecutivos $i-1$, i y $i+1$:

$$\mathrm{Tmin}_{i-1} \le \mathrm{Tmax}_i \ge \mathrm{Tmin}_{i+1} \text{ , y} \qquad (14)$$

$$\mathrm{Tmax}_{i-1} \ge \mathrm{Tmin}_i \le \mathrm{Tmax}_{i+1} \ . \qquad (15)$$

Cada una de las relaciones anteriores se verifica en un control separado (controles CEV_03 y CEV_04). En los registros para los cuales no se cumple alguna de las dos relaciones, se marcan todas las temperaturas involucradas (incluyendo el día anterior y el siguiente) como sospechosas.

Consistencia entre temperatura media y temperatura de rocío. Por su definición, la temperatura de rocío Td es menor o igual a la temperatura del aire. Por lo tanto el promedio diario de la temperatura de rocío tiene que ser menor o igual al promedio diario de la temperatura del aire:

$$\mathrm{Td}_i \le \mathrm{Tmed}_i \ . \qquad (16)$$

[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.

En los registros en que la relación anterior no se cumple, se marcan ambas variables como sospechosas.

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

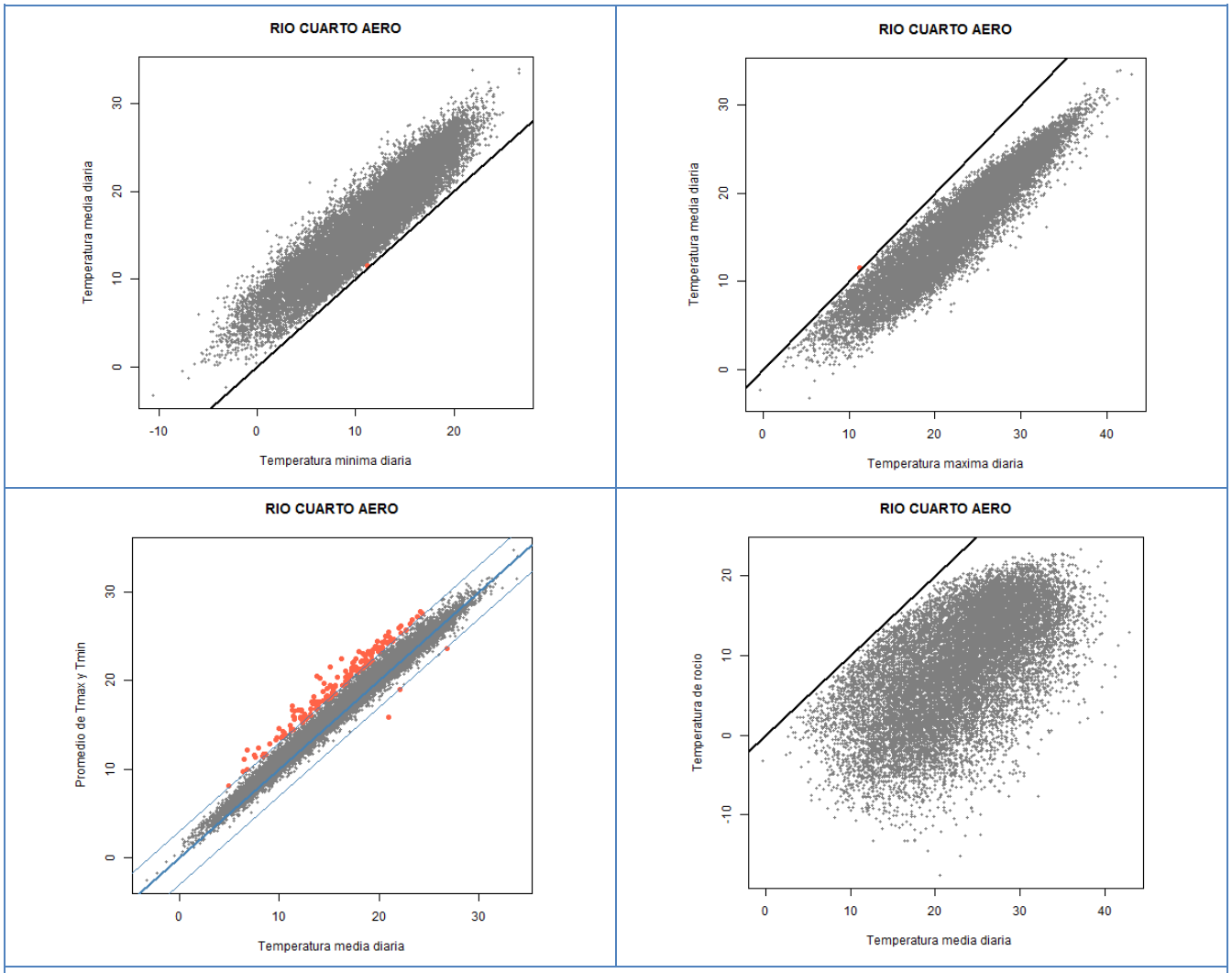


Figura 18. Visualización de los controles de consistencia entre temperaturas diarias. Arriba, izquierda: relación entre temperatura mínima (eje x) y media (eje y). La línea 1:1 se indica en negro; hay un punto (en rojo) identificado como sospechoso; aunque este punto está por encima de la línea 1:1 (o 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).

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]. Los diferentes controles de consistencia en temperaturas se ilustran en la Figura 18 con datos para Río Cuarto, Córdoba, Argentina.

7.2 Consistencia entre datos de presión atmosférica

Como la presión atmosférica es la fuerza que la columna de aire ejerce sobre un determinado lugar, 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 se identifican como sospechosos a los registros de ambas variables cuando la presión atmosférica al nivel de la estación Pres_{est} es mayor que la presión reducida al nivel del mar Pres_{nm}, o sea cuando no se cumple el siguiente criterio:

$$\mathrm{Pres}_{est} \leq \mathrm{Pres}_{nm} \ . \qquad (17)$$

7.3 Consistencia entre datos de viento

Los datos de viento en la base de datos del CRC-SAS incluyen tres variables: la (a) dirección media y (b) 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.

Para que la información de viento sea consistente tiene que cumplir las convenciones con las cuales fueron registradas las observaciones. Una de estas convenciones es que la dirección 0 indica calma, o sea que la velocidad es igual a 0 m s^{-1}. Por lo tanto, si un registro incluye velocidades mayores a 0, la dirección no puede ser 0 (el viento norte se indica con la dirección 36). Por lo tanto, los dos criterios siguientes deben cumplirse simultáneamente, o los datos serán marcados como sospechosos:

$$\mathrm{vmax.d} = 0 \wedge \mathrm{vmax.f} = 0 \ , \quad \text{y} \qquad (18)$$

$$\mathrm{vmax.d} \neq 0 \wedge \mathrm{vmax.f} \neq 0 \qquad (19)$$

Asimismo, el viento medio (o promedio diario de la velocidad del viento) debe ser menor o igual al 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:

$$\mathrm{vmed} \ \leq \ \mathrm{vmax.f} \ . \qquad (20)$$

7.4 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 a los días en los cuales se hayan observado precipitaciones (precipitación > 0.1 mm) pero el promedio de nubosidad es igual a 0 – indicando cielo despejado. 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 datos diarios de estas dos variables se miden en distintos tiempos. La nubosidad es el promedio de la nubosidad en las horas 6, 12, 18 y 0 UTC, mientras que la precipitación es el acumulado desde las 12 UTC del día *i* hasta las 12 UTC del día *i+1*. Entonces, por ejemplo, si la precipitación ocurre sólo durante la noche, y el cielo estuvo despejado (nubosidad 0) durante el día, los datos no parecen consistentes pero sin embargo son válidos.

7.5 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 octas de nubosidad o nub = 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 a los días en los cuales la nubosidad promedio del día analizado es igual a 8 octas (cielo completamente cubierto) y la heliofanía es mayor que 0. El dato erróneo puede ser la nubosidad o la heliofanía, por lo que ambas variables se marcan como sospechosas.

8 Descripción detallada de los controles de consistencia espacial

Todos los controles que se han descrito hasta ahora se han basado en series temporales para la estación meteorológica cuyos datos se estaban controlando. En la familia de controles de consistencia espacial, en cambio, se comparan los datos de la estación meteorológica siendo controlada – denominada “estación central” – con datos para estaciones geográficamente vecinas. Las estaciones vecinas son primero definidas en base a (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 geográfica intentan limitar las comparaciones a estaciones cuyos valores para las variables climáticas sean comparables con los observados en la estación central. Una desventaja de este tipo de controles es que diferencias grandes entre valores de estaciones vecinas puede deberse a causas como variaciones considerables de altitud en la región, o el pasaje de frentes atmosféricos (Hubbard et al., 2012; Kunkel et al., 2005). Las herramientas que se utilizan para el control de valores sospechosos en base a estaciones vecinas pueden ser utilizadas también para llenar datos faltantes (Hubbard et al., 2012).

8.1 Control de regresión espacial ponderada

Este control verifica que el valor de una variable meteorológica para una estación y un día determinado caiga dentro de un intervalo de confianza calculado a partir de ajustes estadísticos basados en datos de estaciones vecinas. Los intervalos de confianza se realizan en base a una serie de regresiones simples entre los valores de una variable en la estación central y los valores de esa variable para 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 utilizar 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, valores para un 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 como un *"spatial regression test"* o SRT por (Hubbard et al., 2012; Hubbard et al., 2005; Hubbard et al., 2007; Kunkel et al., 2005). Para ilustrar el uso del método, supongamos que estamos controlando el valor de temperatura máxima del aire en la estación meteorológica c (central) para el día 22 de agosto de 1962.

(1) El primer paso es extraer los valores de temperatura máxima para cada estación i de las N estaciones vecinas, y 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 en cada estación vecina i (variable independiente) y la temperatura máxima en la estación central (variable dependiente). Cada regresión produce (a) una estimación x'_i de la temperatura máxima en la estación c para el centro de la ventana temporal y (b) el error standard s_i de la regresión (o valor cuadrático medio, o *rms* por las siglas en inglés de *"root mean square").* Ambos valores están basados en la estación vecina i.

(3) Usando las regresiones para todas las estaciones vecinas, se calcula una estimación no sesgada x_c^* de la temperatura máxima en la estación central:

$$x_c^* = \sum_{i=1}^{N}\left(x_i' / s_i^2\right) \Big/ \sum_{i=1}^{N}\left(1/s_i^2\right) . \qquad (21)$$

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

$$N / s_c^{*2} = \sum_{i=1}^{N} 1/s_i^2 . \qquad (22)$$

Este método da mayor peso a estimaciones provenientes de estaciones vecinas que estén más asociadas estadísticamente 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 físicamente. 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 vecina (Hubbard et al., 2005).

(4) En base a 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 x_c, el valor observado para la estación central en el día central de la ventana temporal:

$$x_c^* - f\, s_c^{*2} \le x_c \le x_c^* + f\, s_c^{*2} , \qquad (23)$$

donde f es un factor de multiplicación. Si el valor observado x_c no cae dentro del intervalo de confianza, entonces debe considerarse sospechoso.

El desempeño de este control se ejemplifica en la Figura 19 usando datos de temperatura máxima en Pehuajó, Provincia de Buenos Aires, Argentina. La Figura 19 muestra series de temperatura máxima para Pehuajó y las seis estaciones más cercanas geográficamente. 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 para ese día en las estaciones vecinas.

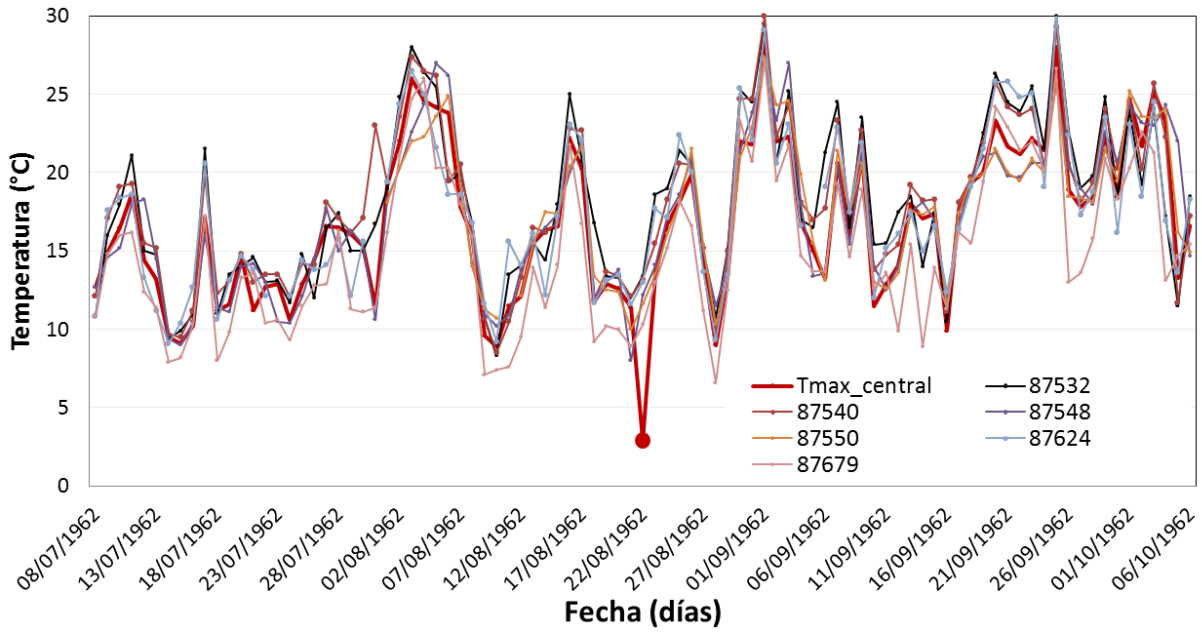


Figura 19. Series observadas de temperatura máxima para una estación central (Pehuajó, Buenos Aires, Argentina) y seis estaciones vecinas. Las series corresponden a una ventana temporal de 91 días centrada en el 22 de agosto de 1962; la ventana incluye el período entre el 8 de julio y el 6 de octubre de 1962. Puede verse que, en general, las trazas observadas de temperatura máxima en todas las estaciones caen dentro de una banda relativamente estrecha. Una excepción es la observación en Pehuajó para el 22 de agosto de 1962, que es sensiblemente más baja que los valores para ese día en las estaciones vecinas.

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 20. 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 2.02°C. Utilizando un factor de multiplicación de 3.5, el intervalo de confianza – calculado siguiendo la Ecuación (23) – para la observación estudiada se extiende desde 5.03°C a 19.17°C. El valor registrado es 2.9°C (indicado con un punto rojo en la Figura 20) claramente cae fuera del intervalo y, en consecuencia, debe considerarse sospechoso. Este valor se identifica como sospechoso. Una verificación posterior reveló que el dato correcto es 12.9°C – es decir, al digitalizar la información no se incluyó el número "1".

8.2 Control de regresión espacial basado en índice de concordancia

En este control, se utilizan coeficientes de regresión e índices de concordancia calculados a partir de los valores en una estación central (aquella que está siendo controlada) y un número de estaciones vecinas. El control está descrito en detalle en Durre et al. (2010) y es relativamente similar al control de regresión espacial utilizado por Hubbard et al. (2005) y discutido en la sección anterior (Sección 8.1). En el control usado por Hubbard et al. (2005), la ponderación de las diferentes estaciones vecinas se basa en el error estándar de las regresiones para

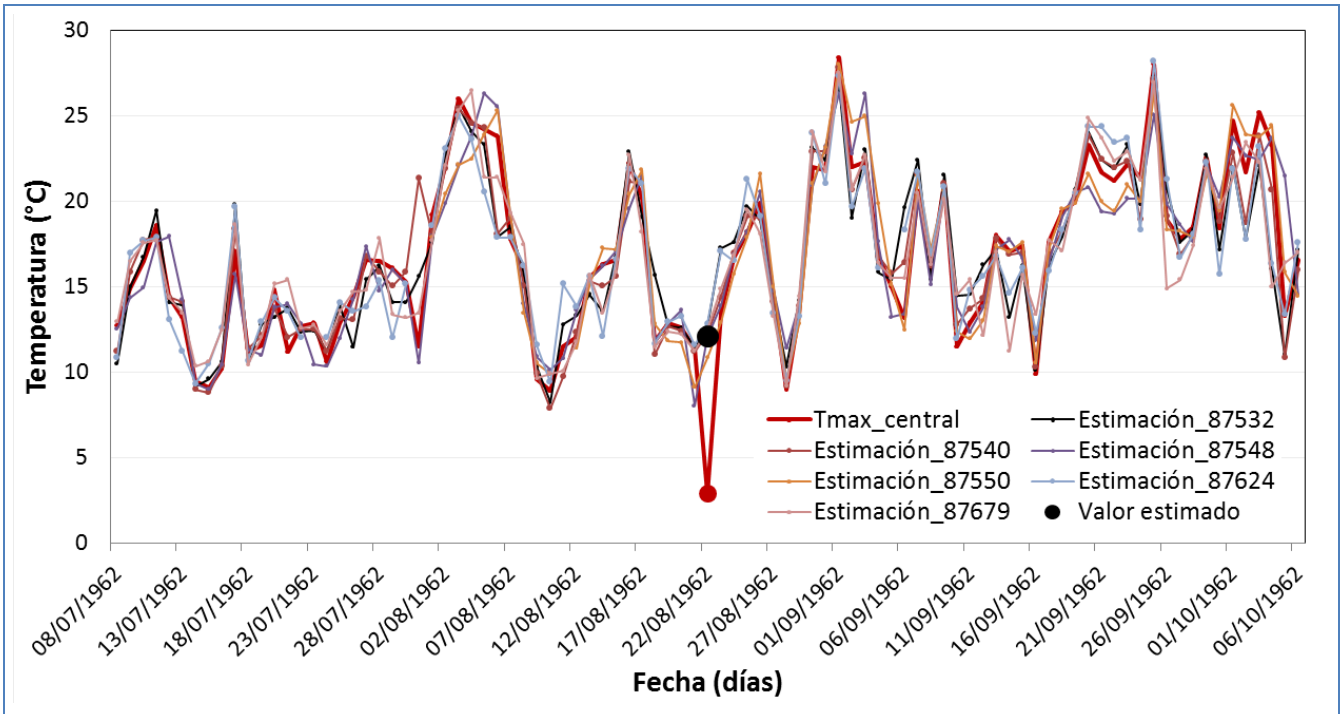


Figura 20. Series de temperatura máxima observada (línea roja gruesa) y estimadas a partir de regresiones con estaciones vecinas para Pehuajó, Buenos Aires, Argentina. La estimación ponderada del valor en la estación central para el 22 de agosto de 1962 (12.1°C) se indica con un punto negro. La temperatura máxima observada en Pehuajó (2.9°C) se indica con un punto rojo. Este valor se confirmó como erróneo.

cada vecina. En este caso, sin embargo, la ponderación se realiza mediante el "índice de concordancia" d propuesto por Legates y McCabe Jr. (1999). Los pasos involucrados en este control se describen en los párrafos siguientes.

(1) El primer paso es seleccionar – para cada estación vecina – datos para la variable analizada (por ejemplo, temperatura máxima) para una ventana de 3 días centrada en el día que se está controlando para la estación central. Por ejemplo, si se está analizando la temperatura máxima para el día t en la estación central c que tiene 4 estaciones vecinas, se reunirán 12 valores de temperaturas máximas (3 observaciones correspondientes a los días $t-1$, t, y $t+1$ para cada una de las 4 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) El segundo paso es calcular – para cada estación vecina – el valor absoluto de las diferencias entre (i) los 3 valores dentro de la ventana temporal y (ii) el valor para la estación central en el día t (centro de la ventana). Para cada estación vecina, se reemplaza el dato original para el día t 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 para el 22 de agosto corresponden en realidad al 21 de agosto, y se indican con un color verde. Este paso se repite para cada día t en la serie original de datos, y así se forma una "serie nueva" de datos.

Tabla 4. Temperaturas máximas en Pehuajó, Provincia de Buenos Aires, 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 como "estación central" y el día del centro de la ventana temporal (22 de agosto de 1962); la celda de la tabla correspondiente al valor siendo controlado está sombreada en azul. Las celdas en verde claro corresponden a los valores de los vecinos con menor diferencia absoluta respecto a la estación central (ver paso 2).

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

(3) Para cada mes/año (por ejemplo, “agosto de 1962”) en la “serie nueva” creada en el paso 2 se define una ventana temporal que se extiende 15 días antes y 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 usando los valores de la vecina como variable independiente 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. 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 en este caso.

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

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

donde m es el número depares de datos válidos dentro de la ventana, $x(i)$ e $y(i)$ son las observaciones en la estación vecina y la estación central para el día i de la ventana, e $\overline{y}$ denota un promedio sobre todas las observaciones dentro de la ventana temporal. Valores altos de d indican tanto una alta correlación como pequeñas diferencias absolutas entre x e y (Durre et al., 2010).

(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 concordancia para cada estación vecina. La estimación se calcula como

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

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 la ordenada al origen o intercepto y la pendiente de la regresión para la estación k , y $x'(i,k)$ es la observación para el vecino 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 en este control consiste en la determinación de valores sospechosos. Para hacer esta determinación, (Durre et al., 2010) definen como sospechosos a los valores que cumplen dos condiciones:

- El valor absoluto de la diferencia entre el valor en la estación central y la estimación de ese valor (que se denomina "residuo") debe ser > 8°C para el control de temperaturas; y
- El valor absoluto del "residuo estandarizado" debe ser > 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. En cambio, se define como sospechoso a todo valor cuyo residuo y residuo estandarizado sean ambos mayores que un cierto percentil límite (en este caso, 0.99).

Según Durre et al. (2010), este control incluye varias diferencias con respecto a la regresión espacial descrita en Hubbard y You (2005) con el objetivo de minimizar la cantidad de "falsas alarmas" (valores identificados como sospechosos que son realmente correctos). Primero, en lugar de usar la correlación o el error estándar de la regresión para ponderar la asociación con cada vecino se usa el índice de concordancia de Legates y McCabe Jr. (1999), que mide no sólo la covarianza entre vecino y estación central, sino también las diferencias absolutas entre ambas series. En consecuencia, la selección de estaciones vecinas con valores altos de d debería reducir el riesgo de 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, el uso de una ventana de 3 días reduce errores en la estimación que puedan ser causados por diferencias entre estaciones 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 reducen el riesgo de emitir muchas falsas alarmas cuando la desviación estándar de los residuos es pequeña.

La Figura 21 muestra una asociación bastante cercana entre los valores de temperatura máxima observados en Pehuajó (considerada como "estación central") y los valores estimados para esa estación a partir de 4 estaciones geográficamente vecinas. En la Figura 22 se continúa con el ejemplo presentado en secciones anteriores: la temperatura máxima para Pehuajó en agosto de 1962. En general, el panel superior de la figura se observa que los valores observados y los estimados a partir de estaciones vecinas son muy similares, excepto para el 26 de agosto – fecha para la cual el valor observado ya fue reportado como erróneo. En el panel inferior, se muestran residuos absolutos y estandarizados. Claramente el día con valor erróneo muestra residuos que superan ambos umbrales utilizados en el test.

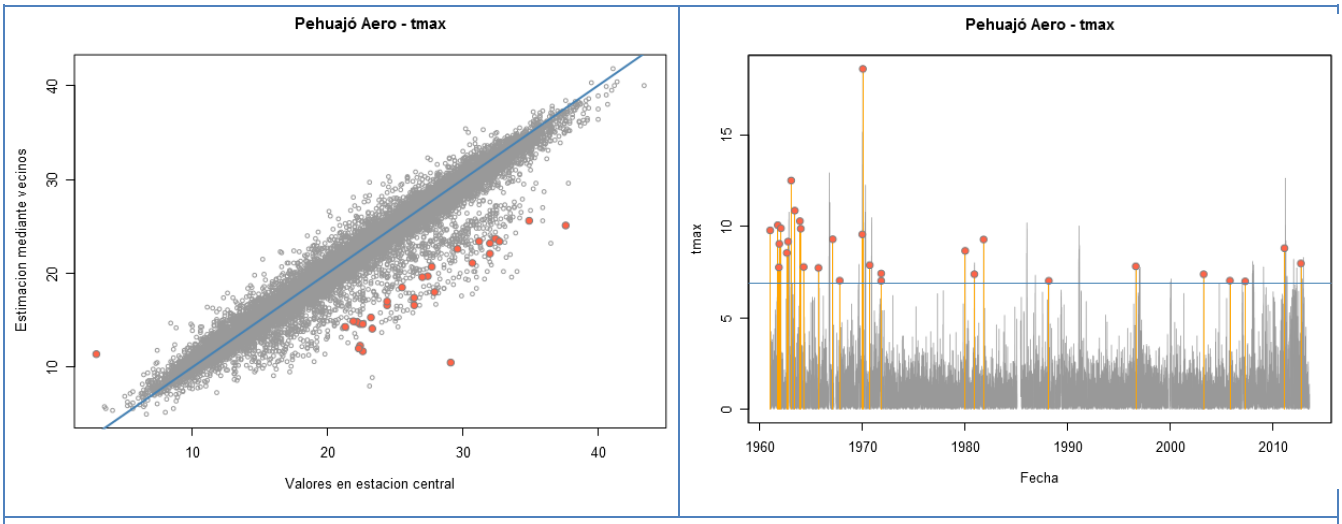
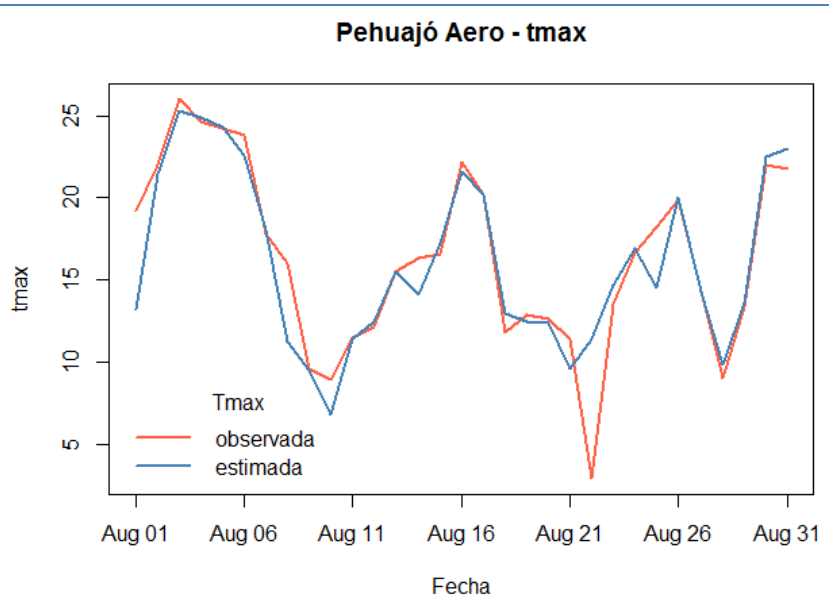


Figura 21. Izquierda: relación entre los valores observados de temperatura máxima en Pehuajó, Argentina y los valores estimados en base a 4 estaciones geográficamente vecinas. Los puntos identificados como sospechosos se muestran en rojo y la línea 1:1 en azul. Derecha: Serie temporal de residuos para la estación central. El umbral para la identificación de valores sospechosos es la línea horizontal azul.



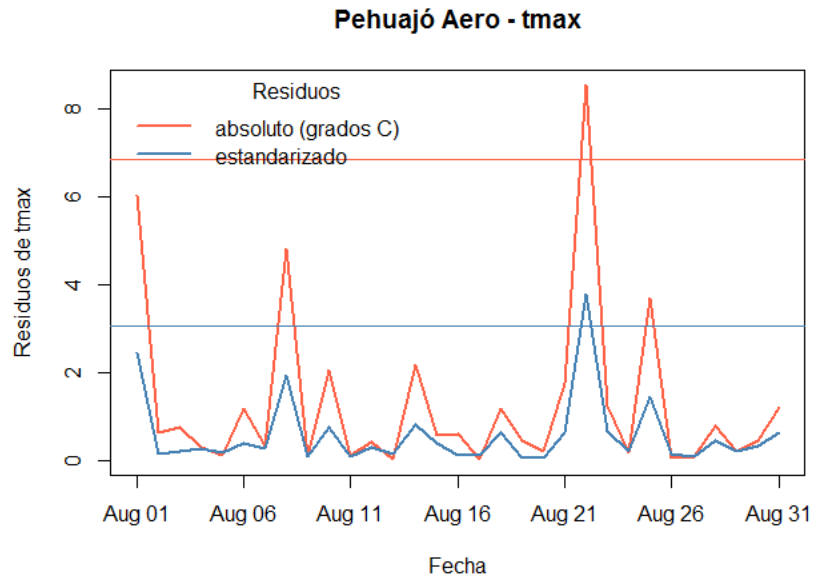


Figura 22. Panel superior: Valor observado de temperatura máxima para una estación central (Pehuajó, Buenos Aires, Argentina) y valor estimado a partir de 4 estaciones vecinas. Las series corresponden a agosto de 1962. Panel inferior: Series temporales de diferencias o residuos entre valores observados 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.

8.3 Control de corroboración espacial para temperaturas

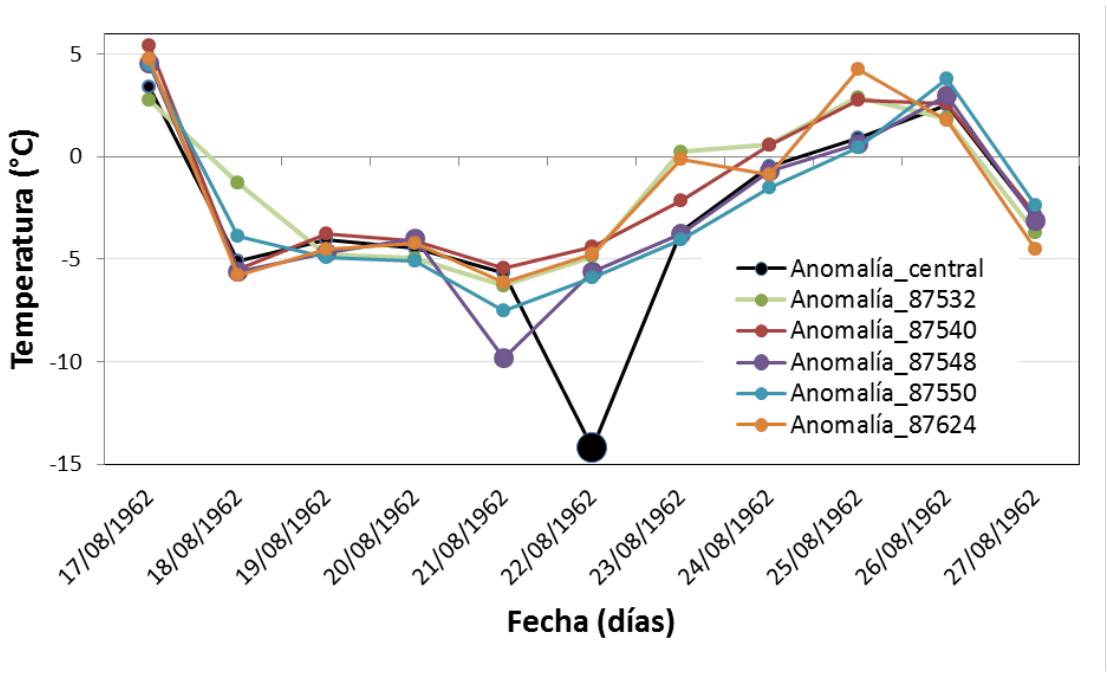
Este control se aplica a temperaturas diarias mínimas, máximas, medias y de rocío. El control – basado en el control de corroboración espacial descrito por Durre et al. (2010) – intenta determinar si un valor cae fuera de un rango de valores reportados en estaciones vecinas. Las estaciones vecinas se seleccionan puramente en base a su cercanía espacial –no en función de la correlación o concordancia con la estación central, como en controles anteriores. La principal ventaja de este control es su aplicabilidad en áreas donde la alta variabilidad espacial o la falta de datos completos impiden la estimación apropiada de una regresión entre estaciones (Durre et al., 2010).

El control 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 de un promedio para 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 dentro de una ventana temporal centrada en ese día del año. Por ejemplo, Durre et al. (2010) usan una ventana de 15 días de ancho: en este caso la media para el 8 de enero se calcula con los valores observados entre el 1 y el 15 de enero de todos los años disponibles para la estación. A continuación se construye una serie temporal de anomalías con respecto al valor promedio para cada día del año.

El segundo paso es seleccionar – para cada estación vecina – las anomalías de la variable analizada (por ejemplo, anomalías de temperatura máxima) para una ventana de 3 días centrada en el día que se está controlando para la estación central. Este paso es similar al usado en el control de regresión espacial basado en el índice de concordancia, descrito en la Sección 8.2. Continuando el ejemplo listado en esa sección, si se está analizando la temperatura máxima para el 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 poder realizar el control, se requiere un número mínimo de anomalías (por ejemplo, 9) para las estaciones vecinas.

El tercer paso es calcular las diferencias absolutas entre (i) las anomalías en las estaciones vecinas y (ii) la anomalía para 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 en la estación central, y por lo tanto el control falla. El umbral se define en términos de temperatura: por ejemplo, Durre et al. (2010) usan un umbral de 10°C. En este caso, 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 (por ejemplo, 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 estaciones central y 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 resumen, los controles de regresión y corroboración se complementan mutuamente (Durre et al., 2010). La Figura 23 ilustra el uso del control de corroboración espacial para temperatura máxima diaria en Pehuajó.



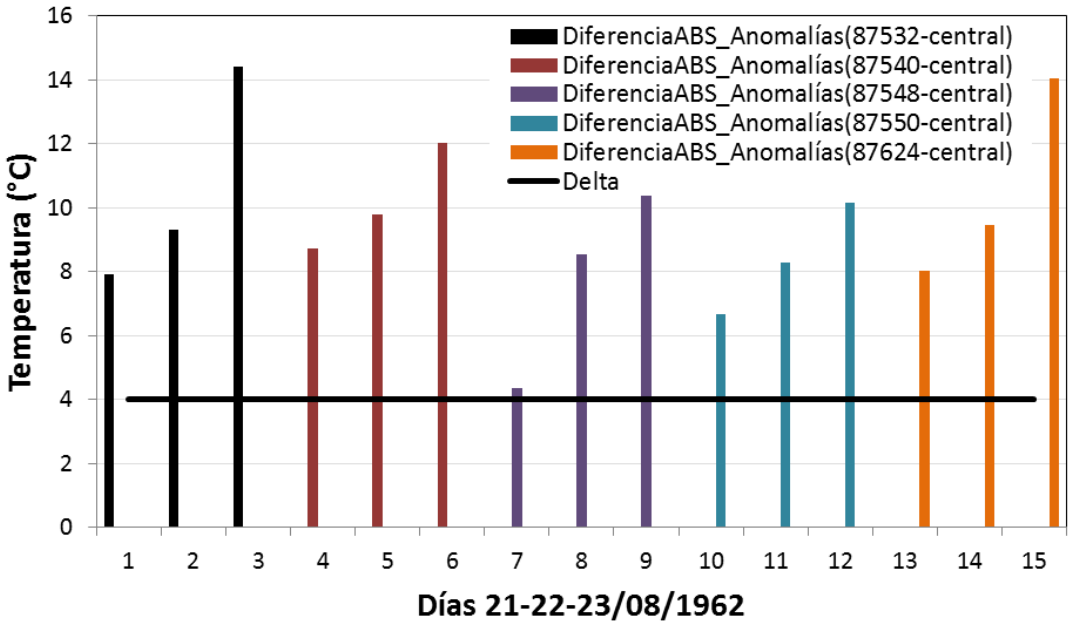


Figura 243. Panel superior: Valor observado de temperatura máxima para una estación central (Pehuajó, Buenos Aires, Argentina) y cinco estaciones vecinas (General Pico (87532), Trenque Lauquén (87540), Junín (87548), Nueve de Julio (87550) y Anguil (87624)). Las series corresponden al período entre el 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 para la estación central es sospechoso.

8.4 Control de corroboración espacial para precipitación

Este control – basado en el control de corroboración espacial de Durre et al. (2010) – se aplica solamente a las precipitaciones diarias. El control intenta determinar si un valor es muy diferente del rango de valores reportados en estaciones vecinas dentro de una ventana temporal de 3 días. 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* del rango, el valor central pasa el control y no se considera sospechoso. En cambio, si la lluvia para 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 el umbral, primero se calcula una cantidad llamada MATD, por las siglas de "Minimum Absolute Target–Neighbor Difference" (Durre et al., 2010). El cálculo se realiza primero en base a las diferencias entre el "valor central" (i.e., la lluvia de la estación central en el día t) 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, 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 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 basado en la Ecuación C1 de Durre et al. (2010):

$$U = -45.72 \ln\left(\mathrm{MATD}_{ranking}\right) + 269.24 \ . \qquad (26)$$

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 (26) – en este caso ≈ 269 mm.

9 Veredictos resultantes de los controles de calidad

El campo "estado" en la tabla de datos diarios en la base de datos registra si un registro determinado (es decir, la combinación única de una estación meteorológica y una fecha) ha pasado o no por algún control de calidad y, si es así, qué resultados se han obtenido del control de calidad para ese registro.

- Cuando un registro no ha pasado por ningún control de calidad, el campo "estado" en la base de datos tiene el valor ***"pendiente"*** para ese registro.

Luego de aplicar los controles de calidad a las series de datos diarios de cada estación meteorológica, el campo "estado" para cada registro puede tomar los valores *"validado", "dudoso" o "faltante"*, que se definen a continuación:

- El estado ***"validado"*** indica que *todas* las variables en el registro analizado superaron *todos* los controles aplicados.
- El estado ***"dudoso"*** indica que al menos una de las variables en el registro analizado no ha pasado *al menos uno* de los controles aplicados.
- El estado ***"faltante"*** indica que no hay datos en ninguna de las variables en el registro analizado.

9.1 Verificación manual de datos "sospechosos"

Los datos sospechosos serán verificados manualmente por los responsables de cada base de datos nacional o institucional. Luego de la verificación manual, a cada dato *"sospechoso"* se le asigna uno de los siguientes códigos:

- ***"Ratificado"*** cuando la verificación manual confirma que el valor de una variable en un registro dudoso es correcto pese a haber fallado al menos uno de los controles. Por ejemplo, al realizar la verificación se encuentra que se trata de un valor extremo, observado correctamente en la estación meteorológica. Pese a ser un valor sospechoso, pudo constatarse que este valor ocurrió realmente (o, por lo menos, que este valor se registró en los registros originales).
- "***Corregido"***, cuando el valor que no pasó al menos un control (a) se confirma como erróneo durante la verificación manual y (b) puede ser corregido (por ejemplo, cotejando las libretas meteorológicas originales).Un ejemplo de esta situación puede ser cuando existen errores en la digitalización del dato original, esto puede corregirse fácilmente recurriendo a la libreta meteorológica donde se encuentra el dato observado original registrado en la estación meteorológica.
- ***"Eliminado"*** cuando el valor que no superó al menos un control (a) se confirma como erróneo durante la verificación manual y (b) NO puede ser corregido (por ejemplo, esta incorrecto en la libreta meteorológica o se extravió la misma); el valor "eliminado" es excluido de los datos controlados (se lista como "faltante").Una situación posible de dato rechazado se presenta cuando es detectado como sospechoso y al recurrir al dato original se encuentra que la medición no fue realizada como dictan las normas de observación.
- ***"No corregible"*** se utiliza cuando el valor no puede ser *ratificado*, *corregido* ni *rechazado*, porque no se tienen las herramientas para hacerlo, el dato que continua siendo sospechoso pero no puede identificarse dentro de las categorías anteriores. En este caso se conserva el dato original, pese a ser sospechoso. Un ejemplo posible se presenta cuando se registra una precipitación extraordinaria que no puede ser ratificada con la libreta meteorológica, por ausencia de la misma o por falta de observaciones, pero mediante otros medios, noticias por ejemplo, se informan las consecuencias de las lluvias, por lo

tanto no puede eliminarse ni ratificarse, dado que hay información de que este dato pudo haberse registrado realmente.

El resultado de cada control es binario (o sea, puede solamente tomar los valores TRUE o FALSE, correspondiendo a valores potencialmente “válidos” o “sospechosos”, respectivamente). Una etapa subsiguiente de verificación manual (documento sobre flujo de controles) confirma si los valores sospechosos son realmente incorrectos. En consecuencia, con los resultados de los controles de calidad y de la verificación manual, se puede explorar la performance de un control construyendo una matriz de contingencia de 2 filas y 2 columnas. Un ejemplo de análisis de este tipo se detalla en el Reporte Técnico CRC-SAS-2014-002.

Apéndice A

A1. Configuración de los controles de calidad para la base de datos del CRC-SAS

Los parámetros estadísticos de cada control fueron ajustados para optimizar la tasa de falsos errores (datos correctos identificados como sospechosos), buscando una eficiente relación entre el número de posibles errores y el tiempo que se utiliza en la inspección manual de cada uno de ellos. Siempre buscando que los errores existentes en la base de datos sean detectados y que los datos etiquetados como sospechosos sean analizados por un operador de manera eficiente en un periodo realmente disponible para tal fin en las instituciones de la región. (Ver Reporte Técnico CRC-SAS-2014-002).

```
# ARCHIVO DE CONFIGURACION PARA CONTROLES DE CALIDAD DE DATOS METEOROLOGICOS
# (PARAMETROS DE CONTROL DE CALIDAD)
# ------------------------------------------------------------------------------

# ------------------------------------------------------------------------------
# Parametros relacionados con la seleccion de estaciones meteorologicas vecinas
# ------------------------------------------------------------------------------

vecinos  :

  # Maxima distancia geografica (en km) entre estaciones para ser consideradas
vecinas
  max_distancia  :  300

  # Maxima diferencia de elevacion (absoluta, en metros) aceptada entre
estaciones vecinas
  max_diferencia_elevacion  : 100
# ------------------------------------------------------------------------------


# ------------------------------------------------------------------------------
# Parametros (limites inferiores y superiores) para controles de calidad
GENERALES
# ------------------------------------------------------------------------------

gen03  :

  # Umbral de precipitacion diaria que define un "dia lluvioso".
  umbral.dia.lluvioso  :  0.1  # Si llueve MAS de 0.1 mm en un dia, es un "dia
lluvioso"

# ------------------------------------------------------------------------------


# ------------------------------------------------------------------------------
# Parametros (limites inferiores y superiores) para controles de calidad de
RANGO FIJO
# ------------------------------------------------------------------------------
```

```
# NOTA: Los limites estan incluidos DENTRO del intervalo de valores aceptables

# Limites inferior y superior de temperatura maxima - unidad: grados C
tmax.inf  :  -39.0
tmax.sup  :  49.0

# Limites inferior y superior de temperatura minima - unidad: grados C
tmin.inf  :  -39.0
tmin.sup  :  49.0

# Limites inferior y superior de temperatura media diaria - unidad: grados C
tmed.inf  :  -39.0
tmed.sup  :  49.0

# Limites inferior y superior de temperatura de rocio - unidad: grados C
td.inf  :  -39.0
td.sup :  49.0

# Limites inferior y superior de precipitacion acumulada diaria - unidad: mm
dia^-1
prcp.inf  :  0
prcp.sup  :  300

# Limites inferior y superior de humedad relativa media diaria - unidad:
porcentaje
hr.inf  : 0
hr.sup  : 100

# Limites inferior y superior de heliofania u horas de sol - unidad: horas
decimales
helio.inf  :  0
helio.sup  :  18

# Limites inferior y superior de nubosidad media diaria - unidad: octas
nub.inf  :  0
nub.sup  :  9

# Limites inferior y superior de direccion de viento maximo diario - unidad:
decenas de grados
vmax.d.inf  :  0
vmax.d.sup  :  36

# Limites inferior y superior de velocidad de viento maximo diario - unidad: m
segundo^-1
vmax.f.inf  :  0
vmax.f.sup  :  62  # equivalente a 120 nudos

# Limites inferior y superior de velocidad de viento promedio diario - unidad:
m segundo^-1
vmed.inf  :  0
vmed.sup  :  26    # equivalente a 50 nudos

# Limites inferior y superior de presion atmosferica en la estacion - unidad:
hPa
```

```
pres.est.inf  :  530
pres.est.sup  :  1060


# Limites inferior y superior de presion atmosferico a nivel del mar - unidad:
hPa
# Estos limites dependen de la altura de la estacion sobre el nivel del mar.

# 1. Para alturas de estacion <= 800 m
pres.mar.inf.1  :  930
pres.mar.sup.1  :  1060

# 2. Para alturas de estacion > 800 m y <= 2300 m
pres.mar.inf.2  :  1000
pres.mar.sup.2  :  1650

# 3. Para alturas de estacion > 2300 m y <= 3700 m
pres.mar.inf.3  :  1600
pres.mar.sup.3  :  3200

# 4. Para alturas de estacion > 3700 m
pres.mar.inf.4  :  3200
pres.mar.sup.4  :  6300
# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Parametros (limites inferiores y superiores) para controles de calidad de
RANGO VARIABLE
# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control RV01, basado en ajuste de ciclo estacional y percentiles de residuos
# ------------------------------------------------------------------------

rv01:

  min.n  :   730            # Numero minimo de dias para poder ajustar ciclo

  tmax.inf  :  0.001        # TMAX Limite inferior (percentil)
  tmax.sup  :  0.999        # TMAX Limite superior (percentil)

  tmin.inf  :  0.001        # TMIN Limite inferior (percentil)
  tmin.sup  :  0.999        # TMIN Limite superior (percentil)

  tmed.inf  :  0.001        # TMED Limite inferior (percentil)
  tmed.sup  :  0.999        # TMED Limite superior (percentil)

  td.inf  :  0.001          # TD Limite inferior (percentil)
  td.sup  :  0.999          # TD Limite superior (percentil)

  hr.inf  :  0.001          # HR Limite inferior (percentil)
  hr.sup  :  0.999          # HR Limite superior (percentil
```

```
  helio.inf  :  0.001       # Heliofania Limite inferior (percentil)
  helio.sup  :  0.999       # Heliofania Limite superior (percentil)

  pres.est.inf  :  0.001    # Presion en la estacion Limite inferior(percentil)
  pres.est.sup  :  0.999    # Presion en la estacion Limite superior(percentil)

  pres.mar.inf  :  0.001    # Presion a nivel del mar Limite inferior
(percentil)
  pres.mar.sup  :  0.999     # Presion a nivel del mar Limite superior
                                (percentil)

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
Control RV02, basado en calculo robusto de media y desv. standard para ventanas
de 3 y 5 dias
# ------------------------------------------------------------------------

rv02:

  # Minimo numero de registros necesarios
  min.n  :  730

  # Ancho de ventana (dias) para estimacion de media y desvio estandar. Valores
posibles: 3 o 5 dias
  ancho.ventana  :  5

  # TMAX Numero de desviaciones estandard para identificar "sospechosos"
  tmax.z  :  3

  # TMIN Numero de desviaciones estandard para identificar "sospechosos"
  tmin.z  :  3

  # TMED Numero de desviaciones estandard para identificar "sospechosos"
  tmed.z  :  3

  # TD Numero de desviaciones estandard para identificar "sospechosos"
  td.z  :  3

  # HR Numero de desviaciones estandard para identificar "sospechosos"
  hr.z  :  3

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control RV03, basado en calculo de mediana y pseudo desv. standard para
ventanas de 3 y 5 dias
# ------------------------------------------------------------------------

rv03:
  # Minimo numero de registros necesarios
  min.n  :   730
```

```
  # Ancho ventana (dias) para estimar mediana y pseudo desv standard. Valores
posibles: 3 o 5 dias
  ancho.ventana  :  5

  # TMAX Numero de desviaciones estandard para identificar "sospechosos"
  tmax.z  :  3

  # TMIN Numero de desviaciones estandard para identificar "sospechosos"
  tmin.z  :  3

  # TMED Numero de desviaciones estandard para identificar "sospechosos"
  tmed.z  :  3

  # TD Numero de desviaciones estandard para identificar "sospechosos"
  td.z  :  3

  # HR Numero de desviaciones estandard para identificar "sospechosos"
  hr.z  :  3

  # PRES_EST Numero de desviaciones estandard para identificar "sospechosos"
  pres.est.z  :  3

  # PRES_MAR Numero de desviaciones estandard para identificar "sospechosos"
  pres.mar.z  :  3

# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Control RV04, solo para heliofania.
# En este control no se ajusta un ciclo estacional, sino que se usa
# el largo maximo del dia (calculado con la latitud de la estacion y dia del
ano)
# como limite superior para valores de heliofania.

# Este control no tiene parametros especificados

# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Control RV05, solo para precipitacion.
# Basado en estimacion de percentil 75 y rango intercuartil de totales de
preciptacion por mes
# (usando solamente valores para dias lluviosos).
# ------------------------------------------------------------------------

rv05:

  # Minimo numero de registros necesarios
  min.n  :   730

  # Umbral de precipitacion diaria que define un "dia lluvioso".
```

```
  umbral.dia.lluvioso  :  0.1  # Si llueve MAS de 0.1 mm en un dia, es un "dia
lluvioso"

  # Factor que multiplica al rango intercuartil, y que sumado al percentil 75
  # define umbral de lluvias sospechosas
  factor.control  :  4

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control RV06, solo para precipitacion. Basado en ajuste de gamma y estimacion
de percentiles extremos
# ------------------------------------------------------------------------

rv06:

  # Umbral de precipitacion diaria que define un "dia lluvioso".
  umbral.dia.lluvioso  :  0.1  # Si llueve MAS de 0.1 mm en un dia, es un "dia
lluvioso"

  # Numero minimo de valores != 0 requeridos para poder ajustar una gamma
  min.N  :  30

  # Percentil que define umbral de lluvias diarias sospechosas
  percentil  :  0.99
# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control RV07, solo para amplitud termica.
# Basado en ajuste de gamma y estimacion de percentiles extremos
# ------------------------------------------------------------------------

rv07:

  # Numero minimo de valores != 0 requeridos para poder ajustar una gamma
  min.n  :  10

  # Numero de desviaciones estandar que definen amplitudes termicas sospechosas
  n.SD  :  3

# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Parametros para controles de calidad de CONTINUIDAD TEMPORAL
# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CT01, que detecta valores repetidos mas de N veces consecutivas
# ------------------------------------------------------------------------

ct01:

  # Maximo numero aceptado de valores repetidos consecutivos
```

```
  max.repetidos  :  3

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CT02 para PRECIPITACION, que detecta secuencias muy largas de dias
secos
# ------------------------------------------------------------------------

ct02:

  # Umbral de precipitacion diaria que define un "dia lluvioso".
  umbral.dia.lluvioso  :  0.1  # Si llueve MAS de 0.1 mm en un dia, es un "dia
lluvioso"

  # Maximo numero aceptado de valores repetidos consecutivos (para este control
usar 0)
  max.repetidos  :  0

  # Percentil de largo de secuencias de dias secos que define largos
sospechosos
  percentil  :  0.999

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CT03, que detecta SALTOS (diferencias entre dias consecutivos) muy
pronunciados
# ------------------------------------------------------------------------

ct03:

  # Percentil usado como limite para saltos sospechosos (para todas las
variables)
  percentil  :  0.995

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CT04, que detecta PICOS (diferencias entre dias previos y
posteriores) muy pronunciados
# ------------------------------------------------------------------------

ct04:

  # Percentil usado como limite para saltos sospechosos
  percentil  :  0.95

# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Parametros para controles de calidad de CONSISTENCIA ENTRE VARIABLES
# ------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------
# Control CEV02 de consistencia entre valores de tmed y promedio de tmax y tmin
# ------------------------------------------------------------------------

cev02:

  # Percentil de tolerancia de las diferencias entre Tmed y (Tmax + Tmin)/2
  perc.tol.tmed2  :  0.99

# ------------------------------------------------------------------------


# ------------------------------------------------------------------------
# Parametros para controles de calidad de CONSISTENCIA ESPACIAL
# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CES01 de regresion espacial entre estaciones vecinas (Hubbard)
# ------------------------------------------------------------------------

ces01:

  # Numero de estaciones vecinas a retener (en orden descendiente de
correlacion)
  n.vecinos.a.usar  : 10

  # Ancho de la ventana temporal (en dias) usada para regresiones con
estaciones vecinas
  ancho.ventana  :  91

  # Numero minimo de datos (dias) para cada estacion vecina dentro de una
ventana temporal
  n.datos.min  :  30

  # Numero minimo de estaciones vecinas en cada ventana para realizar control
de consistencia espacial
  n.vecinos.min  :  2

  # Correlacion (Pearson r) minima necesaria para usar una estacion vecina
  corr.min  :  0.80

  # Factor utilizado para construir intervalos de confianza
  factor.regr.esp  :  2

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CES02 de regresion espacial entre estaciones vecinas (Durre et al.
2010)
# ------------------------------------------------------------------------

ces02:
```

```
  # Numero de estaciones vecinas a retener (en orden descendiente de
correlacion)
  n.vecinos.a.usar  : 10

  # Minimo numero de estaciones vecinas en ventana temporal
  min.n.vecinos  :  2

  # Numero minimo requerido de pares de valores para estacion central y vecina
  # dentro de ventana temporal
  min.n.pares  :  30        # Numero minimo

  # Correlacion (Pearson r) minima necesaria para usar una estacion vecina
  min.corr  :  0.80

  # Percentil limite para residuos en unidades geofisicas
  perc.lim.resid  :  0.975

  # Percentil limite para residuos estandarizados
  perc.lim.std.resid  : 0.975

# ------------------------------------------------------------------------

# ------------------------------------------------------------------------
# Control CES03 de corroboracion espacial entre estaciones vecinas, para
TEMPERATURA (Durre et al. 2010)
# ------------------------------------------------------------------------

ces03:

  # Numero de estaciones vecinas a retener (en orden descendiente de
correlacion)
  n.vecinos.a.usar  :  5

  # Ancho de ventana temporal usada para calculo de la media para cada dia del
anio
  vent.calc.estadisticas :  21

  # Ancho de la ventana temporal (en dias) usada para corroboracion espacial
  ancho.ventana  :  3

  # Numero requerido de anomalias para vecinos para que se realice el control
de corroboracion
  min.anoms  :  9

  # Maxima diferencia tolerada (en grados C) entre anomalia central y TODAS las
anomalias vecinas
  delta  :  2

# ------------------------------------------------------------------------
```

Agradecimientos

Las actividades detalladas en este reporte fueron financiadas por el Banco Interamericano de Desarrollo (BID) a través del contrato C0121-13 con la Universidad de Miami como parte del proyecto *"Hydro-climate Services in La Plata River Basin."* Otros fondos fueron provistos por el Instituto Inter-Americano para el Estudio del Cambio Global (IAI) mediante el proyecto CRN-035, *"Towards usable climate science: informing sustainable decisions and provision of climate services to the agriculture and water sectors of southeastern South America."* El IAI está financiado por la Fundación Nacional de Ciencias de los Estados Unidos de Norteamérica (NSF) a través del grant GEO-1128040. Finalmente, uno de los autores (GP) fue apoyado económicamente por la Fundación Nacional de Ciencias de los Estados Unidos de Norteamérica a través del proyecto 1049109 del programa *"Decadal and Regional Climate Prediction using Earth System Models (EaSM)".*

Referencias

Aizpuru, J. y 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. y 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. y 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. y 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. y 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. y 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. y Valero, F., 2001. Quality Control and Homogeneity of Precipitation Data in the Southwest of Europe. Journal of Climate, 14(5): 964-978.

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

High-level taskforce for the Global Framework for Climate Services, 2011. Climate knowledge for action: a global framework for climate services - empowering the vulnerable, World Meteorological Organization, Geneva, Switzerland.

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

Hubbard, K., You, J. y 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. y 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. y 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. y 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., Easterling, D.R., Hubbard, K., Redmond, K., Andsager, K., Kruk, M.C. y Spinar, M.L., 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.

Legates, D.R. y McCabe Jr., G.J., 1999. Evaluating the use of ‘‘goodness-of-fit’’ measures in hydrologic and

hydroclimatic model evaluation. Water Resources Research, 35: 233-241.
Meek, D.W. y 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. y 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.