

# TRABAJO FIN DE MÁSTER

## RECONSTRUCCIÓN CLIMÁTICA INSTRUMENTAL A ESCALA DIARIA. MODELIZACIÓN Y CARTOGRAFÍA DE ALTA RESOLUCIÓN: APLICACIÓN AL PARQUE NATURAL DEL MONCAYO

*Autor: Roberto Serrano Notivoli*

**Directores: Dr. Martín de Luis Arrillaga  
Dr. Miguel Ángel Saz Sánchez**

**Máster Universitario en  
Tecnologías de la información geográfica para la ordenación del  
territorio: sistemas de información geográfica y teledetección**

Septiembre de 2013



**Universidad  
Zaragoza**

**Departamento de Geografía  
y Ordenación del Territorio**



## Resumen

*La disponibilidad de una base de datos meteorológica de alta calidad y resolución es esencial en toda investigación climática. Actualmente, las fuentes de datos oficiales ofrecen series de datos incompletas, y en muchos casos con errores e inhomogeneidades que no se ajustan a los requerimientos de calidad que exige el estudio científico. En este trabajo, se presenta una metodología de control de calidad y relleno de lagunas de información en series climáticas de temperatura a escala diaria, programada en lenguaje R. Además se ofrecen los resultados en forma de cartografía de alta resolución, todo ello para el área de estudio correspondiente al Parque Natural del Moncayo por su interés ecológico y por presentar una notable diversidad topográfica que determina la presencia de diferentes ambientes climáticos.*

**Palabras Clave:** *Climatología, sistemas de información geográfica, R-project, temperaturas, cartografía, Moncayo, Aragón, España.*

## Abstract

*Availability of a high quality and high resolution meteorological database is essential in every climatic research. Nowadays, official data sources offer incomplete series, and in most cases with errors and inhomogeneities that are not fitted to quality requirements which are required by scientific research. In this dissertation, is presented a methodology to control the quality of temperature series and filling information gaps on them, programming the process on R language. Also, results are offered on high resolution cartographic format. All this is presented to Moncayo Natural Area because its ecological interest and because it presents an unusual topographic diversity which determines the presence of different climatic environments.*

**Key Words:** *Climatology, geographic information systems, R-project, temperatures, cartography, Moncayo, Aragón, Spain*

## Índice

1.	Introducción.....	pág. 4
1.1.	El sistema climático y el papel de las temperaturas .....	pág. 5
1.2.	Antecedentes y justificación.....	pág. 8
2.	Objetivos.....	pág. 10
3.	Material y métodos .....	pág. 12
3.1.	Área de estudio .....	pág. 13
3.2.	Datos utilizados .....	pág. 15
3.3.	Control de calidad.....	pág. 16
3.3.1.	Coherencia interna de las temperaturas .....	pág. 18
3.3.2.	Cálculo de las anomalías .....	pág. 18
3.3.3.	Homogeneización de las series climáticas .....	pág. 23
3.3.4.	Relleno de lagunas de información .....	pág. 26
3.4.	Creación de cartografías de alta resolución a escala diaria .....	pág. 28
3.4.1.	Geoestadística y modelos de predicción espacial.....	pág. 29
3.4.2.	Modelización en GRID a escala temporal diaria.....	pág. 33
3.5.	Creación de una base de datos en NetCDF .....	pág. 35
3.6.	Programación y paralelización en lenguaje R .....	pág. 37
4.	Resultados.....	pág. 40
4.1.	Homogeneización y relleno de lagunas de los observatorios del Parque Natural del Moncayo y áreas limítrofes .....	pág. 41
4.1.1.	Control de calidad de las series de temperatura .....	pág. 41
4.1.2.	Relleno de lagunas y reconstrucción de las series.....	pág. 42
4.2.	Creación de cartografías de alta resolución diaria.....	pág. 48
4.3.	Creación de una base de datos climática en NetCDF.....	pág. 54
5.	Conclusiones.....	pág. 55
6.	Referencias .....	pág. 60

Anexo I. Estaciones utilizadas.....	pág. 65
Anexo II. Funciones programadas en R .....	pág. 68
Anexo II. Síntesis de cartografías.....	pág. 77

## Índice de figuras

1. Visión esquemática de los componentes del sistema climático .....	pág. 5
2. Ciclo diario de la temperatura. ....	pág. 7
3. Localización del área de estudio.....	pág. 13
4. Número de observatorios con dato de temperatura .....	pág. 15
5. Localización de los observatorios utilizados. ....	pág. 16
6. Métodos de ajuste de los datos para un día concreto a los valores de altitud.....	pág. 21
7. Proceso de cálculo de los valores Z.....	pág. 22
8. Esquema de segmentación de una serie.....	pág. 25
9. Proceso de cálculo de los predichos para relleno .....	pág. 28
10. Esquema de la predicción espacial en 1-D y 2-D.....	pág. 32
11. Mallas de GRID calculadas para distintas resoluciones .....	pág. 34
12. Representación gráfica del modo de almacenamiento de datos en formato NetCDF .....	pág. 36
13. Representación gráfica del tiempo de procesado .....	pág. 39
14. Resultados del relleno de datos en 4 estaciones .....	pág. 42
15. Ejemplo de segmentación de una serie climática .....	pág. 43
16. Comparativa Observados-Predichos para todas las estaciones .....	pág. 45
17. Comparativa Observados-Predichos en dato diario para tres estaciones .....	pág. 47
18. Comparativa entre la correlación Observados-Predichos, y la altitud de las estaciones trabajadas .....	pág. 48
19. Distribución de puntos Training y Test .....	pág. 51
20. Ejemplo de cartografía de temperaturas máximas y mínimas para tres días concretos. ...	pág. 53

## Índice de tablas

1. Disponibilidad inicial de datos de temperatura en las 66 estaciones seleccionadas.....	pág. 41
2. Test estadísticos aplicados a los diferentes métodos de creación de cartografías .....	pág. 44
3. Valores de error extraídos de los test estadísticos para cada tipo de interpolación.....	pág. 52
4. Observatorios utilizados .....	pág. 66

*Este trabajo de investigación no hubiera sido posible sin la construcción del modelo estadístico inicial del Dr. Martín de Luis. Agradezco igualmente el apoyo en análisis espacial al Dr. Miguel Ángel Saz, así como en los aspectos teóricos y en la corrección al Dr. José María Cuadrat. Han sido imprescindibles los datos aportados por el Parque Natural del Moncayo a través del Dr. Luis Alberto Longares. También se agradecen los comentarios, aportaciones, y críticas al método propuesto, por parte del Dr. José Carlos González Hidalgo, el Dr. Santiago Beguería, y el Dr. Josep Raventós. También ha sido vital el apoyo moral de los doctorandos Klemen Novak, Ernesto Tejedor, y Edurne Martínez.*

# **1. INTRODUCCIÓN**

## 1.1 El sistema climático y el papel de las temperaturas

La climatología es una de las ciencias que mayor avance ha tenido en las últimas cuatro décadas. La necesidad de anteponerse a la ocurrencia de fenómenos meteorológicos adversos ha hecho y hace a los científicos indagar en el comportamiento del clima en el pasado, y para ello cada día aparecen nuevos métodos de análisis sobre volúmenes de datos cada vez mayores. El estudio de este comportamiento pasado proporciona en muchas ocasiones la oportunidad de hacerlo extensible hacia el futuro, y para ello es necesario comprender el funcionamiento del complejo sistema del que forman parte todos los factores y elementos del clima.

El clima de la Tierra constituye un amplio sistema físico dominado por intercambios energéticos. Estos intercambios mantienen un equilibrio global que, no obstante, se ve matizado por la dinámica interna del sistema que a lo largo del tiempo puede tener situaciones de exceso o de defecto que determinan las condiciones atmosféricas en cada momento y lugar de la superficie.

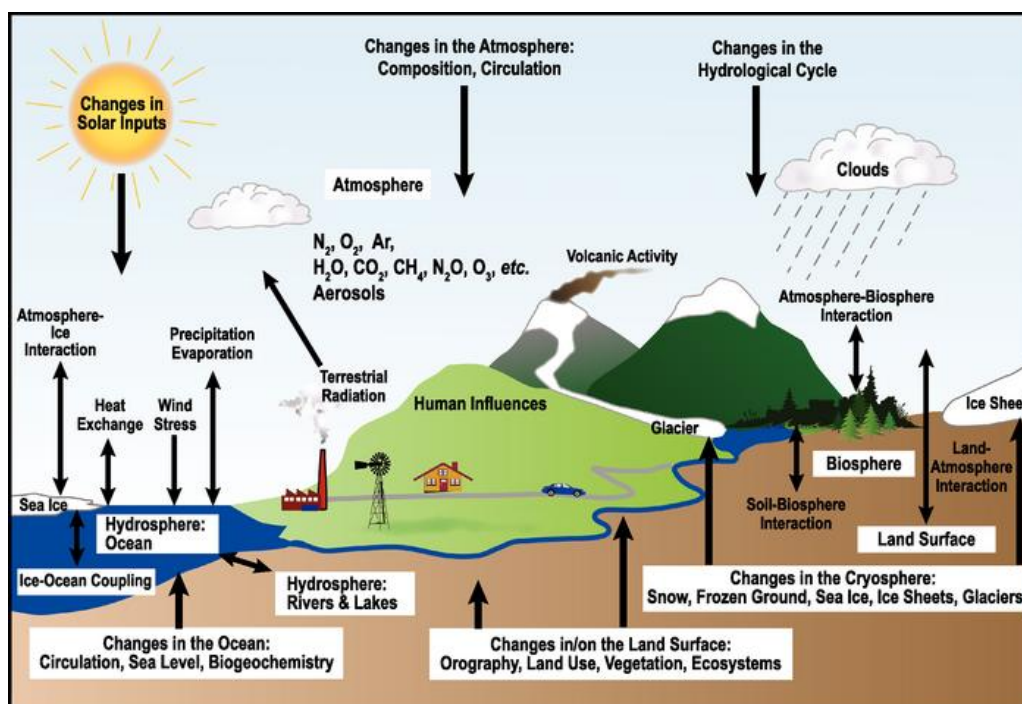


Figura 1. Visión esquemática de los componentes del sistema climático, sus procesos, e interacciones. (Solomon *et al.*, 2007)

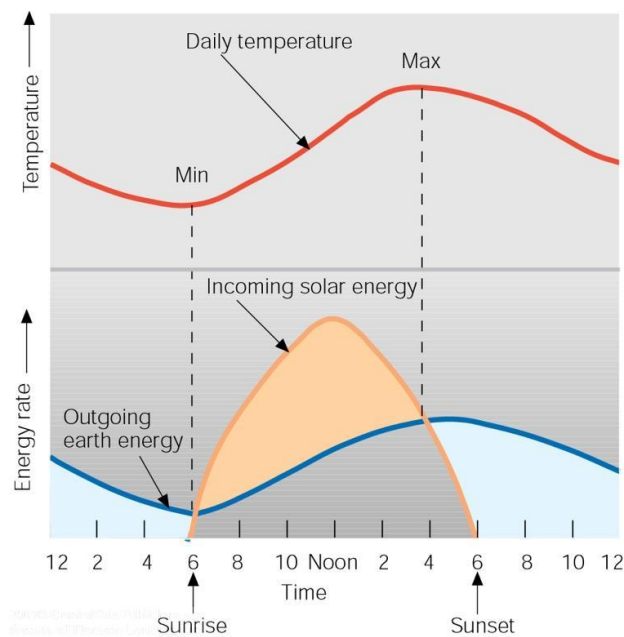
Como consecuencia de todos los procesos y mecanismos que influyen en el sistema climático, se pueden obtener los principales factores reguladores del comportamiento de la temperatura en un punto cualquiera de la superficie terrestre. Según Cuadrat y Pita (2000) estos factores son los siguientes:

- A. La *latitud*, que determina la radiación incidente en el límite superior de la atmósfera a lo largo de las distintas épocas del año.
- B. La *transparencia atmosférica*, que determina el balance de radiación del lugar al condicionar tanto las pérdidas de energía por absorción y reflexión difusa como las retenciones atmosféricas de la radiación infrarroja emitida por la Tierra.
- C. La *naturaleza de la superficie*, que completa el balance al condicionar las pérdidas por reflexión y por irradiación desde la superficie terrestre. Siendo cada tipo de superficie la que impone su impronta sobre las temperaturas (superficie marina, superficie nevada, superficies forestales...)
- D. Las *circulaciones atmosférica y oceánica*, que cumplen el papel de agentes transmisores de calor (por advección).
- E. La *altura sobre el nivel del mar*, cuyo estudio determina que mientras mayor sea ésta, menores serán las temperaturas registradas.
- F. La *topografía*, que determina las diferencias térmicas en función de las orientaciones, así como su papel de efecto barrera para las masas de aire que determinan en gran medida los movimientos verticales de éste.

Todos estos factores determinan la temperatura de un lugar en un momento dado, aunque a lo largo de un periodo de tiempo, la combinación de todos ellos, cada uno con sus variantes, producen variaciones que hacen que el clima de un lugar no sea uniforme, sino que responda a unas oscilaciones más o menos predecibles.

Las variaciones del clima se deben, por una parte, a las fluctuaciones sistemáticas producidas por los movimientos de rotación y traslación de la tierra, que en las zonas subtropicales y templadas espacialmente generan cambios predecibles en los regímenes de temperatura y precipitación que varían de una región a otra, son las denominadas “estaciones”. Sin embargo, existen otro tipo de variaciones que en un principio no responden a fluctuaciones más o menos conocidas, sino que en ellas intervienen efectos más locales (variación en el intercambio de energía, retroalimentaciones regionales, etc.). Los ritmos de llegada de radiación solar a la superficie terrestre, que determinan el

carácter térmico de ésta, se traducen en la existencia de dos ciclos básicos para la temperatura (Cuadrat y Pita, 2000): un *ciclo anual*, que refleja el comportamiento cíclico que registra la radiación solar en su incidencia sobre la superficie terrestre, y un *ciclo diario*, que refleja la alternancia entre días y noches para los distintos lugares del planeta. El presente trabajo está orientado al estudio de la temperatura diaria, y es necesario conocer que estas temperaturas experimentan un ciclo diario en el cual los valores mínimos se registran poco después del amanecer y los máximos unas horas más tarde de que pase el Sol por su cenit sobre el lugar.



**Figura 2.** Ciclo diario de la temperatura. Por término medio el mínimo diario se produce poco después del amanecer y el máximo unas horas después del mediodía.

(<http://clem.mscd.edu/~wagnerri/radtemp.htm>)

La combinación de ambos ciclos (anual y diario) configura los regímenes térmicos de los distintos lugares del planeta.

Con el estudio de los datos climáticos lo que se obtiene son generalizaciones de situaciones atmosféricas genéricas que afectan a determinadas regiones en momentos concretos. Para abordar la interpretación de esas generalizaciones es necesario contar con un largo número de observaciones del tiempo atmosférico concreto de cada momento, y a partir de aquí, observar tendencias en la evolución de los mismos.

Esta generalización del tiempo atmosférico resulta obligada para la caracterización de los climas porque es la única manera de retener la gran variabilidad del tiempo atmosférico diario. Si eliminamos las variaciones estrictamente periódicas que



experimentan las magnitudes climáticas (especialmente el ciclo diurno y el anual), éstas adoptan un comportamiento variable pero estacionario, es decir, fluctúan aleatoriamente en torno a un valor central (Cuadrat y Pita, 2000).

Un número escaso de observaciones meteorológicas sobre un lugar no es capaz de describir el clima de esa región, ya que la alta variabilidad que se da entre diferentes meses, años, o incluso décadas, puede ser de tal magnitud que el resultado final sea una caracterización climática equivocada. Es por eso que la Organización Meteorológica Mundial (WMO) recomienda un periodo de datos observados sobre el terreno (obtenidos con instrumental calibrado) de al menos 30 años para abordar estudios climáticos. El gran problema de la ciencia climática es la obtención de series continuas de datos sobre una región de un periodo tan largo. Esto se soluciona mediante la estadística aplicada a los datos meteorológicos, que es la base de la climatología. Cada serie vendrá definida por algún parámetro de tendencia central (normalmente la media o la mediana), que expresaría el valor en torno al cual fluctuarán todos los demás, y por algún parámetro de dispersión (desviación típica, recorrido semi-intercuartílico, etc.), expresivo de la magnitud de esas fluctuaciones).

El clima de un lugar sería, pues, la generalización estadística de su comportamiento atmosférico, considerado éste como estable, como estacionario, aunque enormemente variable de unos momentos a otros (Cuadrat y Pita, 2000).

## **1.2 Antecedentes y justificación**

El análisis de esta generalización exige disponer de información meteorológica continua y homogénea que posibiliten la correcta caracterización del clima, lo cual obliga a realizar un proceso riguroso de control de calidad y homogeneización de los datos. Los trabajos desarrollados de control de calidad y relleno de información climática de datos mensuales son abundantes (Brunetti et al., 2006; González-Hidalgo et al., 2011; Mitchell & Jones, 2005; Aguilar et al., 1999), aunque los trabajos sobre dato diario también han sido tratados (Cano & Gutierrez, 2004; Herrera, 2011 y 2012; El-Kenawy, 2012). Metodológicamente, ambas escalas no pueden ser tratadas de la misma manera, ya que climáticamente el comportamiento no es igual. El dato diario introduce mucha más incertidumbre y variabilidad, con un porcentaje mucho mayor de aparición de datos extremos, que a escala mensual podrían considerarse datos anómalos.

Para el presente trabajo, se ha considerado el periodo de estudio 1950-2012 porque, abarcando un amplio escenario de variabilidad climática temporal, es el que mayor número de datos disponibles tiene en relación a las bases de datos manejadas. A pesar de todo, la cantidad de datos de temperatura especialmente desde 1950 hasta los 1970 es escasa, pero como el interés final del proyecto reside en obtener series continuas de datos para lugares donde nunca ha habido un observatorio meteorológico, o para aquellos observatorios donde no se ha recogido dato en un determinado periodo de tiempo, la metodología propuesta intenta resolver estos problemas con los datos disponibles.

La importancia de obtener un dato climático a escala diaria puede ser escasa si lo que se necesita es conocer la tendencia a largo plazo en series muy largas de información (como las manejadas), para esto sería suficiente con el dato mensual. Sin embargo, existen otras aplicaciones que sí lo requieren, y que se están llevando a cabo en el Macizo del Moncayo, como la dendrocronología, que se encarga de estudiar la respuesta del crecimiento de los árboles a la variabilidad del clima. Esta respuesta es tan sensible que la variación diaria del clima en diferentes altitudes puede modificarla. Además, a partir de una serie de datos diarios de calidad y homogeneizados, se pueden extraer valores mensuales que sí son adecuados para observar tendencias.

La obtención de resultados en forma de cartografía de alta resolución implica que previamente ha habido que corregir todas las series climáticas de los observatorios, y repetir el proceso de relleno de información para cada uno de los píxeles que compondrán el mapa final diario del área de estudio. Una cartografía de estas características permite comprobar los cambios que se dan en micro-ambientes generados en el Macizo del Moncayo. La diferencia de altura es tan alta, que una gran resolución aportará información sobre los límites de distribución en altura de las heladas, o los fenómenos de inversión térmica, tan frecuentes en los inviernos del Moncayo, información muy valiosa para estudios dendrocronológicos y ecológicos de la vegetación.

## **2. OBJETIVOS**

El objetivo principal del trabajo es **obtener una cartografía de alta resolución y espacialmente continua del Parque Natural del Moncayo en la que se representen las temperaturas máxima y mínima diaria**. Para ello es necesario crear un modelo de distribución de las temperaturas que obtenga como resultado una malla regular.

Tal y como señala Cuadrat (2000), un modelo es una representación simplificada de la realidad; en consecuencia, un modelo climático es una representación simplificada de los procesos que gobiernan el clima, o lo que es lo mismo, del funcionamiento del sistema climático. Su objetivo es simular el funcionamiento de este sistema a fin de comprender los procesos que lo gobiernan y de predecir los efectos que sobre el clima pudieran derivarse de cambios en estos procesos.

En este caso, se toman como factores que influyen en la distribución espacial de las temperaturas aquellos que son inherentes al dato climático como tal: latitud, longitud, y elevación. Éstos componen los metadatos de un valor climático, y por tanto el modelo propuesto construye el *output* final a partir de los parámetros geográficos más básicos.

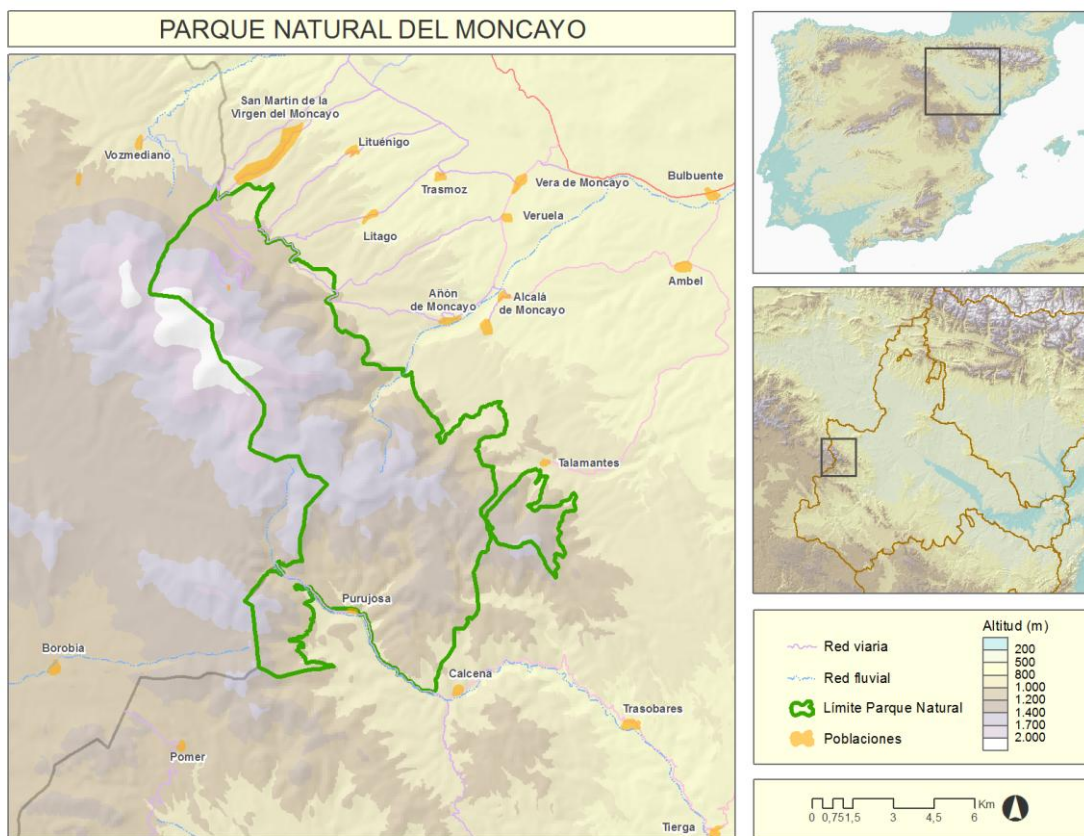
Para conseguir el objetivo principal es necesario desarrollar otros objetivos secundarios:

- I. Reconstrucción de los observatorios del P.N. del Moncayo y áreas limítrofes
  - a. Control de calidad de las series de temperatura
    - i. Cálculo de anomalías
    - ii. Homogeneización de las series
  - b. Relleno de lagunas de información y reconstrucción de las series
- II. Creación de cartografías de alta resolución a escala diaria
  - a. Modelización del dato climático en GRID a escala temporal diaria
- III. Creación de una base de datos en NetCDF de la información climática

## **3. MATERIAL Y MÉTODOS**

### 3.1 Área de estudio

El área de estudio que aquí se presenta, se ubica en el extremo centro-occidental de la provincia de Zaragoza (Fig. 3), repartiéndose por un total de 9 términos municipales diferentes (Tarazona, San Martín de la Virgen del Moncayo, Lituénigo, Litago, Trasmoz, Añón de Moncayo, Talamantes, Calcena y Purujosa). Se trata de un territorio inscrito dentro de la figura de Espacio Natural Protegido de la Red Natural Aragón (Ley 12/2004, de 29 de diciembre), respondiendo a criterios relacionados con la conservación de la naturaleza, promoción y desarrollo socioeconómico y lugar de esparcimiento y disfrute público, siendo su objetivo principal conservar la biodiversidad.



**Figura 3.** Localización del área de estudio

En la actualidad su denominación es la de Parque Natural del Moncayo, pero desde su creación, esta ha sido objeto de cambios importantes tanto en su catalogación, como en su extensión y régimen de protección. A pesar de no ser el primer espacio protegido declarado en territorio español, su creación llega en fechas muy tempranas, al

ser declarado en 1927 "Sitio Natural de Interés Nacional", para más tarde, en 1978 convertirse en "Parque Natural de la Dehesa del Moncayo" y posteriormente en 1998 alcanzar la denominación definitiva de "Parque Natural del Moncayo".

Estas declaraciones han traído consigo además la ampliación en la superficie protegida, pasando de las iniciales 1.500 has. del denominado "Monte del Moncayo" a las 9.848 has. que alcanzaron protección en 1998, para alcanzar las actuales 11.144 has., que hoy conforman este Parque Natural, tras su ampliación en 2007 a petición de los ayuntamientos de Talamantes y Purujosa. Además, en el confluyen otras figuras de protección como la de Lugar de Importancia Comunitaria LIC LIC ES2430028 Sierra del Moncayo, o Zona de Especial Protección para las Aves ZEPA 0000297 Sierra del Moncayo-Los Fayos-Sierra de Armas, habiendo sido propuestos varios Lugares de Interés Geológico como: los Circos glaciares de San Miguel y San Gaudioso (Tarazona) y Morca (Añón), Cueva Hermosa (Calcena), Peñas Albas (Calcena), Cueva de los Rincones (Purujosa) y el Barranco de Valcongosto (Purujosa).

En relación con las figuras de protección, la conservación de este espacio se ha visto reforzada con la aprobación de varias herramientas relacionadas con la gestión, encaminadas a la conservación de los valores que promovieron su inclusión en la Red Natural de Aragón. Así en el año 1998, se aprueba la primera de ellas, tras un intenso proceso de elaboración que concluye con el primer Plan de Ordenación de los Recursos Naturales para el espacio, que servirá de base para el diseño y aprobación del Plan Rector de Uso y Gestión (PRUG) del Parque en 2002, en vigor y prorrogado desde 2007 hasta el momento actual, a la espera de la aprobación de la revisión del nuevo PRUG que debe continuar las principales líneas de gestión del espacio protegido. En este sentido los resultados del trabajo que aquí se muestran, pretender aportar información útil de cara a la gestión, contribuyendo a incrementar el conocimiento del espacio y sus dinámicas desde el punto de vista de la cubierta vegetal.

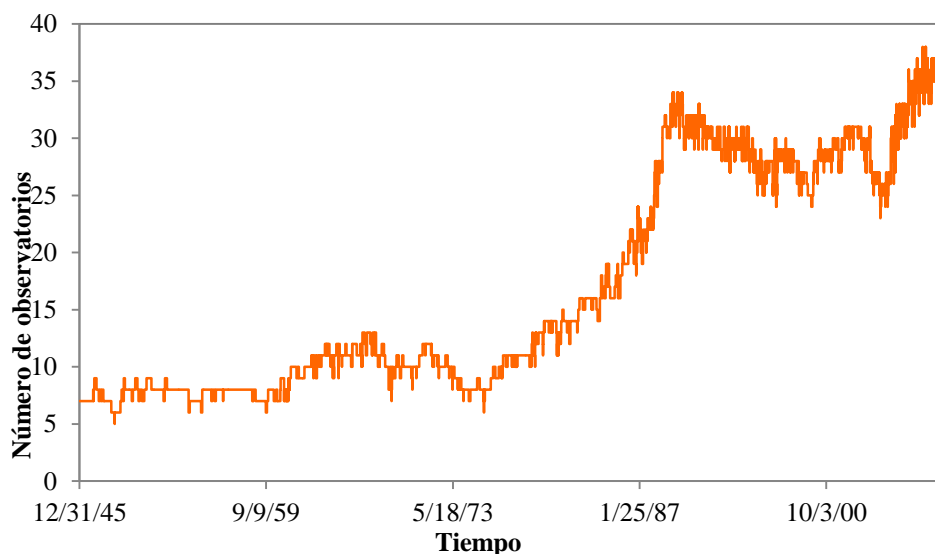
Bajo este contexto administrativo se engloba un amplio territorio con carácter geográficamente singular, de rasgos diferenciadores del resto de relieves, morfologías y paisajes circundantes. Desde su especial configuración topográfica, que le otorga un carácter de macizo montañoso aislado entre los ambientes esteparios del centro de la Depresión del Ebro y los más continentalizados de la meseta castellana en su sector oriental, hasta su identidad como territorio frontera entre ecosistemas contrastados y refugio natural de elementos bióticos propios de ambientes eurosiberianos, frecuentes en cordilleras alejadas como la cantábrica o los pirineos. Por otra parte, su orientación,

estructura geológica y composición litológica, le han dotado a lo largo del tiempo de rasgos favorecedores para la génesis y funcionamiento de procesos glaciares y periglaciares, que han dejado su impronta a través de morfologías fácilmente reconocibles, del mismo modo que su gradiente altitudinal ha marcado sus rasgos climáticos a lo largo del tiempo y hoy en día es el principal factor de la distribución de la vegetación en altura, una vez que la influencia del hombre en el manejo de la cubierta vegetal va dejando paso a la dinámica natural. Será el primer aspecto, el de los rasgos climáticos el objeto de estudio del presente trabajo, en especial de la temperatura y su distribución en la superficie del Parque Natural.

### 3.2 Datos utilizados

Para el presente estudio, se han utilizado un total de 66 observatorios con registro de temperatura (relación en Anexo I). De todos ellos, 59 forman parte de la red de la Agencia Estatal de Meteorología (AEMET), 4 corresponden a la Confederación Hidrográfica del Ebro (CHE), y los 3 restantes pertenecen al Parque Natural del Moncayo y están gestionadas por el Departamento de Geografía y O.T. de la Universidad de Zaragoza (Agramonte, Calcena, y Majada Baja).

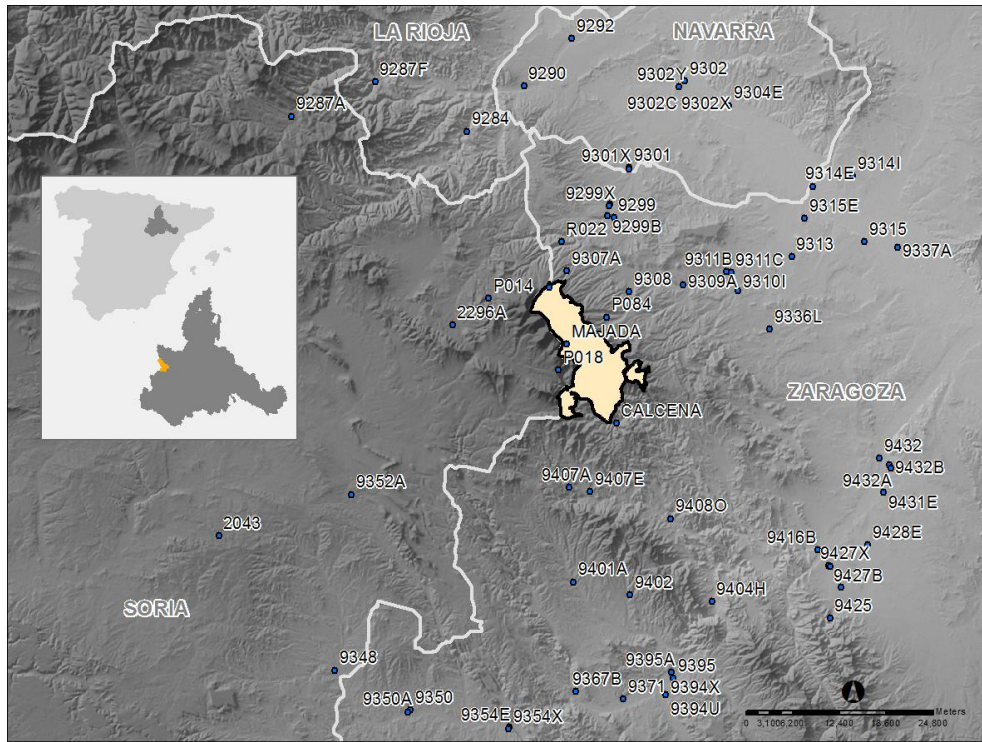
En el caso de AEMET, tiene las series más largas, con algunas estaciones con dato disponibles desde el año 1950. Las estaciones pertenecientes a la red de la CHE tan solo tienen datos disponibles desde el año 2003, y las tres estaciones del PN del Moncayo comenzaron a registrar datos en el año 2010



..  
**Figura 4.** Número de observatorios con dato de temperatura para cada día en el periodo 1950-2012



Los observatorios disponibles con dato hasta los años 80 son menores que los de precipitación, aunque más continuos en el tiempo. La disponibilidad muchos observatorios en cada día es clave para asegurar la calidad del dato final, no obstante, con un mínimo de 10-15 observatorios disponibles para todo el periodo, es más que suficiente para realizar un proceso de relleno fiable.



**Figura 5.** Localización de los observatorios utilizados. En el centro, el Parque Natural del Moncayo

La selección de observatorios se ha realizado en función de la localización del Parque Natural del Moncayo, con el objetivo de crear una cartografía de alta resolución del espacio natural. Esta cartografía permitirá complementar a otros estudios ecológicos del PN aportando series climatológicas homogeneizadas y de calidad.

### 3.3 Control de calidad

El control de calidad es un paso fundamental para eliminar los datos incorrectos y comprobar la consistencia de los mismos (Feng *et al.*, 2004). Para obtener como resultado unas series climatológicas fiables es necesario que los datos pasen por un estricto control de calidad. Tanto las series extraídas del Banco de Datos Climatológico Nacional de AEMET, como las provenientes de la CHE, o los datos obtenidos de las

estaciones del PN del Moncayo, producen bases de datos en estado bruto, carentes de cualquier tipo de filtrado, y susceptibles de contener todo tipo de errores.

Algunos tipos de errores asociados a la toma de datos pueden ser:

- Codificación errónea de decimales
- Duplicación de registros
- Asociación del valor a un momento distinto al que le corresponde
- Mal funcionamiento del termómetro
- Instrumental mal calibrado
- Cambios en la localización del observatorio

Todos estos posibles errores generan datos aislados, o series temporales continuas que no concuerdan con la variabilidad propia del ambiente en el que está inmerso el observatorio, y que solo son posibles de identificar comparándolos con los datos de las estaciones del entorno más inmediato que, previsiblemente, mantendrán la misma variabilidad temporal.

Al trabajar con datos a escala temporal diaria, la variabilidad es mucho más manifiesta que con variables mensuales o anuales. La aparición de valores extremos que pueden ser confundidos con datos anómalos (outliers) es mucho más frecuente, por tanto es imprescindible adoptar una metodología que permita identificar este tipo de anomalías correctamente.

Las metodologías utilizadas para el control de calidad en datos mensuales consideran la comparación de la serie candidata con las de referencia en toda su dimensión temporal, es decir, desde el inicio hasta el fin de cada una de las series (p.ej.: Esteban *et al.*, 2009; Herrera, 2011 y 2012; El-Kenawy, 2012; entre otros). **El método propuesto en este trabajo compara, para cada registro diario, las circunstancias que se dieron en todos los observatorios del entorno que condicionaron el valor de la variable en cada uno de ellos**, esto es, se hace un control, día por día, de cómo la topografía, la latitud y la longitud, influyeron en el dato de temperatura y precipitación de cada observatorio y en qué medida. Así, se estima un valor promediado entre todas las estaciones vecinas que servirá de referencia para observar las anomalías que puede haber entre ésta (serie referencia) y la estación objetivo a corregir (serie candidata).

### 3.3.1 Coherencia interna de las temperaturas

Un primer análisis del control de calidad pasa por analizar la coherencia interna de las temperaturas. Al tratar de manera separada las temperaturas máximas y mínimas, pueden aparecer incoherencias entre ambas, cuando realmente deben guardar una proporción de superioridad una (máximas) sobre la otra (mínimas). Antes de comenzar con el proceso de eliminación de las anomalías de ambas bases de datos, se comparan una con otra para comprobar que las máximas son superiores a las mínimas en todos los días considerados y para todos los observatorios. En el caso de un día en el que esta declaración no se cumpla, se eliminan ambos datos, ya que no es posible verificar de manera fiable cuál de los dos es cierto, o si ambos son falsos.

### 3.3.2 Cálculo de las anomalías

Como consecuencia de los tipos de errores enumerados en el apartado 4.1, pueden aparecer datos anómalos en las bases de temperatura. Al estudiar las series climáticas en comparación a las de sus vecinos, se pueden determinar los días concretos en los que el valor de la variable climática de un observatorio tiene una varianza muy superior a la supuesta por la propia variabilidad local del clima. El objetivo es conseguir una serie de valores Z para cada observatorio, que representan la varianza extrema del valor original respecto a la modelización de los valores predichos de sus vecinos.

El proceso sigue los siguientes pasos hasta obtener una serie de anomalías:

i) Búsqueda de vecinos: En este caso, se trabaja con todos los datos disponibles, lo ideal es que existan al menos cinco observaciones de estaciones vecinas para cada día.

ii) Estandarización de las series vecinas

ii.a) Búsqueda de días con registro comunes entre la serie candidata y los vecinos

ii.b) Media móvil: Para cada día, promedio de un mes de datos (15 días anteriores, y 15 posteriores). Se generan dos tablas, una con los datos observados de la serie candidata para cada vecino (1), y otra con los datos de los vecinos para cada uno de ellos (2)

$$MMCAND_i = \frac{CNCAND_{i-15} + \dots + CNCAND_{i+15}}{n} \quad (1)$$

$$MMVECI_i = \frac{CNVECI_{i-15} + \dots + CNVECI_{i+15}}{n} \quad (2)$$

Donde  $CNCAND_i$  es el dato de la candidata en el día que tiene registro en común con su vecino;  $CNVECI_i$  es el dato que tiene en común cada vecino con la serie candidata en el día  $i$ ; y  $n$  es el número de días comprendidos en el cálculo de la media.

ii.c) Desviación Estándar: del mismo periodo que las medias móviles calculadas en el paso anterior:

$$SDCAND_i = \sqrt{\frac{(CNCAND_{i-\dots} + MMCAND_i)^2}{N}} \quad (1)$$

$$SDVECI_i = \sqrt{\frac{(CNVECI_{i-\dots} + MMVECI_i)^2}{N}} \quad (2)$$

ii.d) Estandarización: normalización en función de la media y la desviación estándar, y reescalamiento por la desviación estándar y la media de la serie candidata para ese periodo.

$$Est_i = \left( \frac{CNVECI_i - MMVECI_i}{\sigma_i} \right) \sigma_{cand} + \bar{x}_{cand}$$

Donde  $Est_i$  es el valor estandarizado para el día concreto;  $x_i$  es el dato observado en el día  $i$ ;  $\bar{x}_i$  es la media móvil de un mes de los valores observados de cada vecino;  $\sigma_i$  es la desviación estándar del conjunto de valores de un mes de cada vecino;  $\sigma_{cand}$  es la desviación estándar del conjunto de valores de un mes de la serie candidata; y  $\bar{x}_{cand}$  es la media móvil de un mes de los valores observados de la serie candidata

iii) Cálculo de valores predichos de los parámetros (Predichos Parciales): los valores estandarizados previamente se ajustan a los parámetros de Altitud, Latitud, y Longitud mediante el modelo de regresión que mejor se ajuste. El ajuste se hace

mediante polinomios desde orden 1 (ajuste lineal) hasta 4. El que mejor  $r^2$  produce es el método elegido para ese parámetro en concreto y ese día.

El dato predicho mediante una **regresión lineal** se obtiene con el cálculo de un solo coeficiente:

$$Y_i = \beta_0 + \beta_1 X_i$$

Donde  $Y_i$  es la variable dependiente (temperatura o precipitación) en un día concreto;  $\beta_0$  es el intercepto de la pendiente, función de la regresión lineal; y  $\beta_1$  es el coeficiente para la variable independiente  $X_i$  que representa la altitud, longitud, o latitud, de ese observatorio en ese día (es constante para cada observatorio).

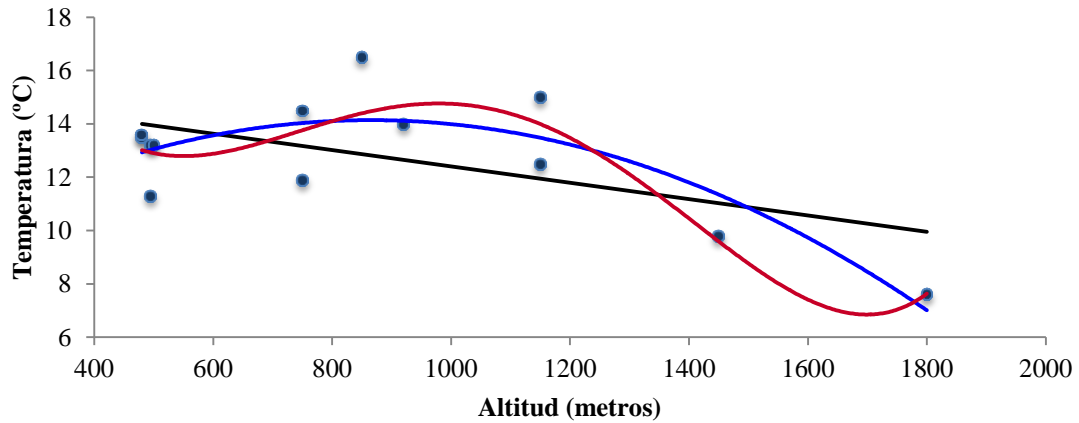
La predicción de mediante una **regresión polinomial** de orden 2 (1), 3 (2), y 4 (3) se obtiene con el cálculo de coeficientes de número variable en función del orden de polinomio elegido:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \quad (1)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \quad (2)$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 \quad (3)$$

Si el número de observaciones es abundante, el ajuste será mucho mejor, ya que la interpolación del valor predicho se hará sobre el cálculo de un modelo con más opciones posibles. Es por ello que para realizar la regresión polinomial de máximo orden, han de existir al menos 5 observaciones.



**Figura 6.** Diferentes métodos de ajuste de los datos de temperatura máxima para un día concreto a los valores de altitud. Desde regresión lineal (negro) ( $r^2 = 0.31252$ ), hasta modelos de regresión con polinomios de orden 4 (rojo) ( $r^2 = 0.75481$ ), pasando por una regresión con orden de polinomio 2 (azul) ( $r^2 = 0.6823$ ).

iv) Estandarización de los Predichos Parciales: se estandarizan los valores obtenidos previamente, para normalizarlos y que sean comparables.

v) Promedio de los valores Predichos Parciales y estandarización de los mismos. Esto genera unos valores predichos finales, que representan la **serie de referencia**. El promedio de los Predichos Parciales es una ponderación en función de sus valores de  $r^2$ .

$$PFinal_i = \frac{(PParcial_{i alt} r^2_{i alt}) + (PParcial_{i lon} r^2_{i lon}) + (PParcial_{i lat} r^2_{i lat})}{r^2_{i alt} + r^2_{i lon} + r^2_{i lat}}$$

Donde  $PFinal_i$  es el valor Predicho Final para un día;  $PParcial_{i alt}$  es el Predicho Parcial de la Altitud para ese día;  $r^2_{i alt}$  es el valor de  $r^2$  para el modelo de predicción de la altitud de ese día; ;  $PParcial_{i lon}$  es el Predicho Parcial de la Longitud;  $r^2_{i lon}$  es el valor de  $r^2$  para el modelo de predicción de la longitud;  $PParcial_{i lat}$  es el Predicho Parcial de la Latitud; y  $r^2_{i lat}$  es el valor de  $r^2$  para el modelo de predicción de la latitud.

vi) Cálculo de diferencias: diferencia entre la serie de referencia y la serie candidata

$$Dif_i = PFinal_i - Obs_i$$

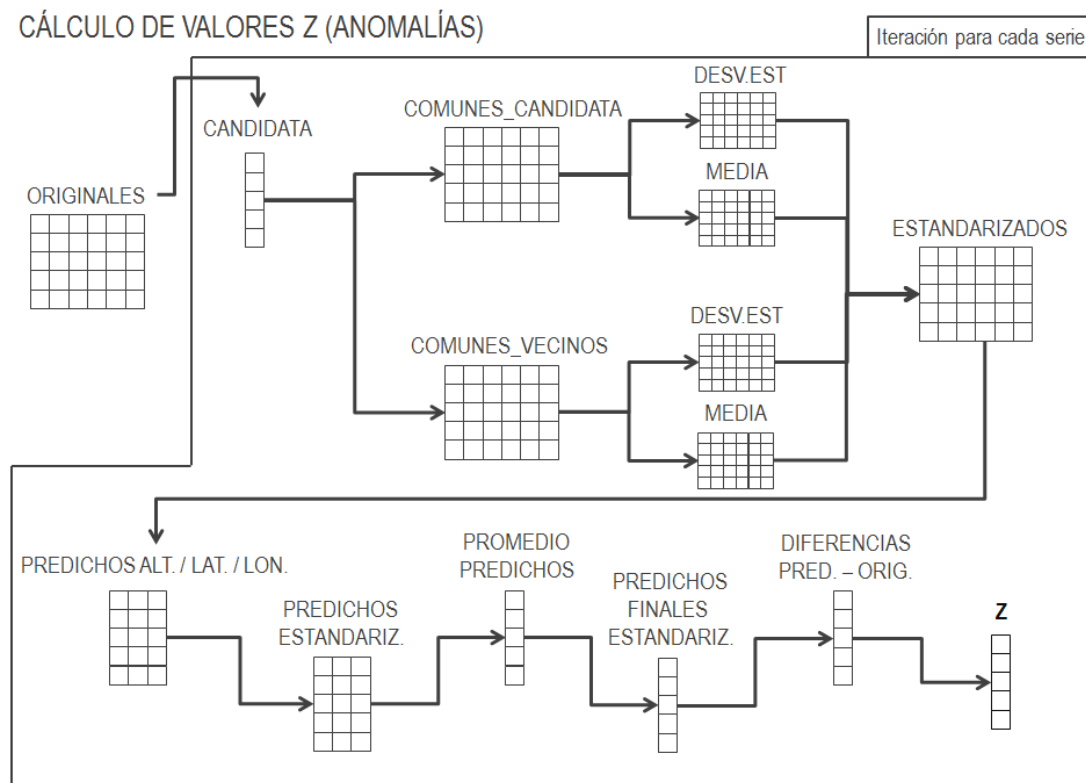
Donde  $Dif_i$  es la diferencia entre el valor observado de la serie candidata en el día  $i$ , y el valor observado de la serie de referencia en el mismo día; y  $Obs_i$  es el valor de la variable observada para ese día.

vii) Cálculo de los valores Z: cociente entre la serie de diferencias calculadas previamente y la desviación estándar de esa misma serie.

$$Z_i = \frac{Dif_i}{\sigma_{Dif}}$$

Donde  $\sigma_{Dif}$  es la desviación estándar de toda la serie de diferencias.

Esta serie de valores Z representa una serie de anomalías que se interpreta como el número de desviaciones estándar que la serie original se aleja de la serie de referencia estandarizada a partir de los datos de todos los observatorios vecinos, e individualizada para cada uno de los días de la serie candidata.



**Figura 7.** Proceso de cálculo de los valores Z.

Cuando se ha realizado el primer cálculo de valores Z, se eliminan, de los valores originales, aquellas Z cuyo valor es superior a 3,5 (en valor absoluto). Estos valores indican los días en los que el valor original está 3,5 desviaciones estándar por encima o por debajo respecto al predicho a través de la modelización de sus vecinos, por eso se eliminan de la base de datos original, sobre la que se vuelve a repetir la iteración.

Después de las iteraciones necesarias, en los valores Z finales ya no existirán días con desviación estándar superior a 3,5. En ese momento se vuelve a repetir el proceso una última vez sobre la matriz original depurada, pero esta vez modificando el cálculo para que se calculen valores predichos para todos los días en los que originalmente existía dato. Así pues volverán a aparecer valores Z, de los que se eliminarán de la matriz de datos original aquellos que superen el umbral de 5.

Con este proceso se obtiene una matriz de datos limpia de cualquier anomalía posible, considerando siempre la variable climática como datos estadísticos en comparación con sus vecinos más próximos, y otorgando robustez a la base de datos inicial.

### ***3.3.3 Homogeneización de las series climáticas***

Cualquier cambio en las variables climáticas que no responda a una variación producida por la propia dinámica del clima se considera una inhomogeneidad. Las inhomogeneidades producen saltos continuos en el tiempo de los datos climáticos, y la complejidad para detectar estos saltos o puntos de ruptura reside en la dificultad para diferenciar los cambios producidos por causas ajenas a la variabilidad climática.

En los estudios sobre variabilidad y cambio climático, tan importante como el dato es el metadato, ya que este último proporciona información valiosa sobre posibles comportamientos no relacionados con circunstancias climáticas (Aguilar *et al.*, 2003). La ausencia de metadatos en la mayoría de los observatorios no permite conocer las características históricas de los mismos (cambios de ubicación, instalación de infraestructuras cercanas, cambios en la instrumentación, etc...) que han podido introducir un factor en la ecuación climática que esté modificando los datos aportados por el instrumental instalado. Por esta razón es necesario adoptar un método o aplicar un algoritmo adecuado para detectar estos cambios.



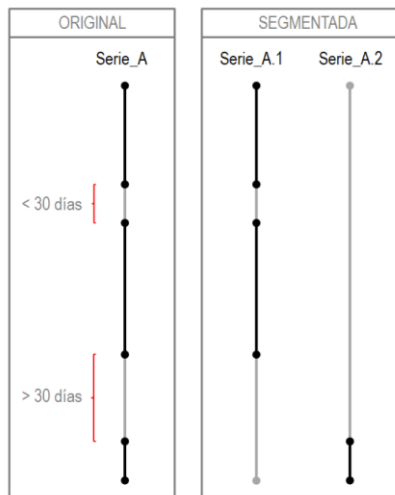
La homogeneización en series climáticas se ha trabajado históricamente sobre dato mensual (Guijarro, 2008; Stepanek, 2003) y anual (García-Barrón y Pita, 2001). El trabajo de Venema *et al.* (2012) resume y compara las técnicas más utilizadas habitualmente y recomienda el uso de la herramienta HOME (HOME, 2011), un paquete estadístico para lenguaje R que a través de una sencilla interfaz es capaz de homogeneizar series mensuales.

La homogeneización de series diarias se trabaja mucho menos en investigación climática debido a la compleja variabilidad del clima a esta escala temporal, de hecho, la mayor parte de los trabajos cubren esta parte del proceso con una homogeneización del dato mensual o anual, y aplicando los parámetros obtenidos a las series de escala diaria (El-Kenawy *et al.*, 2013; El-Kenawy, 2012, Mestre *et al.*, 2011). Se asume que, si es necesario un ajuste en el dato diario, el método más simple de ajuste reside en la interpolación de los coeficientes de ajuste mensuales (Vincent *et al.*, 2002).

Para el presente trabajo, **se ha utilizado otra manera de tratar la homogeneización de series climáticas, la segmentación.**

Las series originales de datos no son continuas, de hecho tienen *huecos* que responden a días en los que el observatorio no registró nada, o que han sido eliminados por considerarse anomalías ajenas a la respuesta digital de los instrumentos de medida.

Si estos *huecos* han sido superiores a 30 días, la serie se fragmenta en ese punto y pasa a considerarse como otra nueva. La razón de esta fragmentación es que, al tratar los datos de una misma serie como periodos continuos independientes, se elimina una posible contaminación de las medias y desviaciones estándar en el caso de que ambos periodos fuesen inhomogéneos (es decir, que tuviesen una señal climática diferente que respondiese a efectos ajenos a la propia evolución del clima).



**Figura 8.** Esquema de segmentación de una serie. Los periodos con dato (negro) separados por periodos sin dato (gris) de más de 30 días se fragmentan para crear una serie nueva.

El resultado de la fragmentación provoca que las 66 series originales de temperatura se conviertan en 360 en el caso de las máximas, y 363 en el caso de las mínimas.

No obstante, con el fin de evaluar la homogeneidad de estos tramos segmentados, se ha utilizado el Standard Normal Homogeneity Test, SNHT (Alexandersson, 1986), aplicado a cada fragmento. Tal y como explican Keiser y Griffiths (1996), el SNHT incorpora una secuencia de diferencias de temperaturas o ratios de precipitación definidos como  $q_i$  entre los valores de la estación test candidata y algún tipo de valor medio, denominado valor de referencia, procedente de las estaciones del entorno:

$$q_i = tsta_i - refsta_i, (1 \leq i \leq n)$$

donde  $tsta_i$  son los valores de la estación candidata; y  $refsta_i$  son los valores de referencia. Estas series de diferencias se estandarizan:

$$z_i = [q_i - \bar{q}]/\sigma, (1 \leq i \leq n)$$

donde  $\bar{q}$  es la media de  $q_i$ , las series de diferencias; y  $\sigma$  es desviación estándar de las series de diferencias completas. Esta nueva serie estandarizada cero de valor medio,

y uno como desviación estándar. Se calculan entonces dos series de medias  $\bar{z}_1$  y  $\bar{z}_2$ , que son simplemente las medias antes y después de cada valor en las series estandarizadas. Estas dos series se combinan en una serie de valores  $T_v$ :

$$T_v = [v\bar{z}_1 + (n - v)\bar{z}_2], (1 \leq i \leq n)$$

donde de nuevo  $v$  varía y  $n$  es el número total de valores en la serie estandarizada. El valor del test estadístico es el valor máximo en las series  $T_v$  e indica dónde está la inhomogeneidad (o break) más larga.

### ***3.3.4 Relleno de lagunas de información***

Normalmente, las series de datos climáticos brutos, incorporan datos vacíos, bien por un error o mal funcionamiento durante la toma de datos, o bien porque el periodo de datos considerado no es cubierto totalmente por la serie considerada.

Con el fin de completar una serie homogénea y continua, es necesario rellenar estos huecos o lagunas de información, con valores lo más reales y fiables posibles. A pesar de que ante la ausencia de dato, cualquier relleno de ese hueco va a ser una interpretación de la realidad en ese momento, si la información climática de observatorios cercanos es completa y de calidad, es posible realizar un relleno adecuado mediante la influencia de las condiciones geográficas (altitud, latitud y longitud) en los datos observados. Las variables geográficas son las únicas que pueden incluirse como parámetros en la ecuación final de predicción, ya que son invariables en el tiempo. No obstante, para periodos de tiempo en los que se disponga de información detallada de otros parámetros como usos del suelo, vegetación, radiación solar, estructura forestal, etc., sería más que adecuado introducirlos como parámetros de predicción. Para el presente trabajo se ha seleccionado como periodo de estudio desde 1950 hasta 2012, por lo que el relleno de las variables climáticas se hará sobre todos los días que lo comprenden considerando, en cada caso, los datos observados existentes en las estaciones del entorno, y los tres parámetros geográficos ya nombrados.

Para rellenar los huecos de información, la secuencia de comandos a ejecutar es prácticamente la misma que para calcular anomalías. El proceso de relleno se compone

de 4 etapas bien diferenciadas y todas ellas necesarias para obtener una serie final de valores robusta estadísticamente a la par que lo más cercana a la realidad posible:

1. *Estandarización de vecinos:*

Se realiza a partir de los datos de los vecinos depurados mediante la eliminación de las anomalías calculadas previamente.

2. *Relleno de todas a periodo común:*

Para crear una serie de referencia que sirva de primer relleno se seleccionan como vecinos aquellas estaciones que para el periodo con dato en la candidata, compartan al menos 30 días observados. Esto asegura un relleno con homogeneidad temporal de al menos un mes para dato diario rellenado.

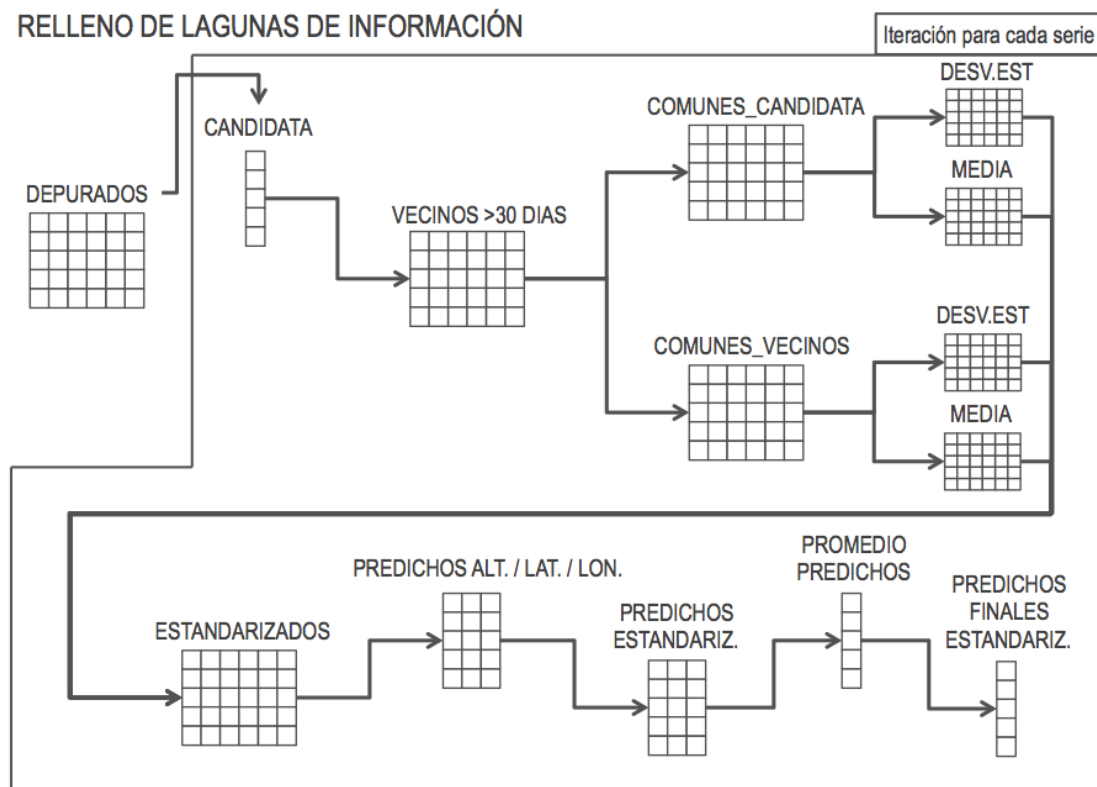
3. *Siguiente iteración, utilizar datos generados en la anterior para seguir con el relleno:*

Después del primer relleno se genera una nueva matriz de datos que incorporarán los datos originales depurados, y los predichos por un primer relleno. En sucesivas iteraciones, esta matriz se utilizará como series de vecinos, y volverán a rellenar todos los huecos de la serie candidata original.

Este proceso se repite tantas veces como sea necesario hasta que estén rellenadas todas las series en su periodo completo. El método asegura un relleno total en apenas 2 o 3 iteraciones, en función de la completitud de los datos originales.

4. *Relleno final:*

El último relleno se realiza sobre la matriz de datos en la que ya todas las estaciones están rellenadas. A partir de aquí solamente es necesario hacer una modelización de todos los vecinos para cada candidata, y para cada uno de los días. El resultado final es un conjunto de series completas desde el inicio hasta el final sin un solo hueco, que mantiene la variabilidad de las series originales, matizada por la situaciones locales tanto espacial como temporalmente.



**Figura 9.** Proceso de cálculo de los predichos para relleno.

### 3.4 Creación de cartografías de alta resolución a escala diaria

La influencia que el clima tiene sobre muchos de los procesos bióticos y abióticos que se desarrollan en la superficie terrestre, su decisiva participación en la agricultura, su relación con actividades humanas tan importantes desde el punto de vista económico como el turismo e incluso su incidencia sobre la salud humana, entre otros aspectos, han hecho que cada vez más y desde más ámbitos, se estén demandando cartografías que representen con precisión la variabilidad espacial de las variables climáticas (Saz *et al*, 2010).

El fundamento de la **Predicción Espacial** o de la **Interpolación Espacial** es predecir los valores de la variable objetivo a lo largo de todo el área de estudio, cuyos resultados se presentan habitualmente en forma de imágenes o mapas. Hay que matizar que existe una pequeña diferencia entre ambos conceptos, ya que la *predicción* puede implicar interpolación y extrapolación. La predicción fuera del rango práctico (cuando la predicción excede la varianza global) se denomina extrapolación. En otras palabras, la extrapolación es la predicción en aquellas ubicaciones donde no hay suficientes

evidencias estadísticas como para que las predicciones sean significativas (Hengl, 2009).

### ***3.4.1 Geoestadística y modelos de predicción espacial***

Hay una diferencia importante entre la geoestadística y la cartografía convencional de variables ambientales: la primera de ellas está basada en la aplicación de técnicas estadísticas cuantitativas. La cartografía geoestadística se basa en el uso de algoritmos semi-automatizados, a pesar de que pueda parecer que la predicción espacial se haga exclusivamente con programas informáticos, los analistas son capaces de elegir usar modelos no lineales, considerar ubicaciones espaciales o no, transformar o usar los datos originales, considerar efectos de multicolinealidad o no, etc. Se trata pues de un sistema experto de decisión. En resumen, **la cartografía geoestadística se define como una producción analítica de mapas utilizando observaciones de campo, parámetros definitorios, y un software informático que calcula los valores en las ubicaciones requeridas del área de estudio** (Hengl, 2009).

El gran problema de la representación de la información obtenida en ubicaciones puntuales es que la gran mayoría de ella está concentrada en las zonas de mayor población o las más accesibles, dejando de lado zonas elevadas, de gran densidad de vegetación, o sin aprovechamiento humano, cuando realmente son las que mayor interés tienen para comprender el funcionamiento de los ecosistemas naturales. En el caso de variables climáticas existen un gran número de métodos de interpolación, desde los que se limitan a asignar a cada punto el valor de la observación más próxima a el (p.ej. vecino más cercano), hasta los que consideran explícitamente el estado sinóptico de la atmósfera en el proceso de interpolación (p.ej. interpolación condicionada, ver Hewitson and Crane, 2005) (Herrera, 2011). El comportamiento de estos métodos depende principalmente de las escalas espacial y temporal, y la densidad de la red de estaciones considerada (Hofstra *et al.*, 2008).

Los diferentes **tipos de interpoladores** se pueden clasificar en (Saz *et al.*, 2010; Vicente *et al.*, 2003):

### Locales:

No utilizan información auxiliar y únicamente se sirven de los datos climáticos de los observatorios existentes en la base de datos. Captan bien la variabilidad de alta frecuencia, es decir, las variaciones del clima debidas a factores de carácter más local.

- *Inverso de la Distancia (IDW)*: estima los valores a partir de un promedio ponderado con los datos de los puntos de muestreo existentes en el conjunto de interpolación seleccionado, teniendo más peso los puntos más cercanos al existir una ponderación en función de la distancia
- *Splines*: obtiene los datos de los observatorios a un radio determinado. Es un interpolador exacto porque las funciones se adaptan a los valores climáticos en los puntos muestreados. Permiten introducir parámetros de suavizado o tensión, pero los valores predichos siempre coinciden con los observados en el observatorio.
- *Funciones Radiales (Radial Basis Functions)*: es similar a si se ajustase una membrana de goma a lo largo de los puntos observados, minimizando el total de la curvatura de la superficie. Según la función de las cinco elegidas, el ajuste a los valores observados será diferente.

### Geoestadísticos:

Asumen que la variación espacial de una variable climática continua es demasiado irregular para ser modelada por una función matemática, y su variación espacial podría ser mejor predicha por una superficie probabilística. La ventaja frente a otros métodos es que tienen en cuenta la autocorrelación espacial de los datos.

- *Kriging*: estiman el valor que un determinado atributo tomaría en un punto a partir de los valores observados en los puntos vecinos, teniendo en cuenta unos factores de ponderación determinados. Esta ponderación no sólo tiene en cuenta la distancia, sino que también toma en consideración la relación espacial general que en todo el área de estudio se establece entre los valores de los diferentes pares de puntos situados a distancias similares, integrando de esta forma una consideración espacial más amplia sobre la variación del atributo que la que asumen los métodos de interpolación local. Hay diferentes tipos de kriging, y varios trabajos los describen en detalle (Isaaks & Strivastava, 1989; Goovaerts, 1997; Borrough & McDonnell, 1998).

### Globales:

Usan todos los observatorios. Crean modelos de dependencia entre los datos climáticos y otras variables independientes. Son inexactos porque los valores predichos no coinciden con los datos climáticos medidos sobre el terreno (Borrough & McDonnell, 1998).

- *Superficies de tendencia*: Se modeliza la variación espacial de una variable que se realiza de forma continua a partir de la obtención de superficies matemáticas suavizadas, mediante el ajuste de las observaciones a ecuaciones polinomiales de distinto orden, de tal forma que los valores de aquellos puntos en los que no existe información puedan ser calculados a partir de la latitud y longitud
- *Modelos de regresión múltiple*: En el caso de las variables climáticas, la elevación es la variable independiente que un mayor porcentaje de varianza suele explicar. Junto a ella, es habitual incluir como variables explicativas la latitud, la longitud y la distancia al mar.

#### Mixtos:

Combinan las ventajas de la aplicación de los métodos globales y su capacidad para mostrar las variaciones a escala regional del atributo cartografiado, con los métodos locales y sus ventajas a la hora de captar las variaciones debidas a factores locales. Se basan en la corrección de los residuales (diferencia entre observados y predichos) mediante la suma de la interpolación de éstos a los valores predichos iniciales.

- *Modelos de regresión múltiple con corrección de residuales*: Se basan en los modelos de regresión múltiple simples, aunque en este caso la diferencia entre los observados y los predichos por éstos, se interpola normalmente con un método local y se suma al modelo inicial para corregir así los valores originales en aquellos lugares donde hubo observación.

En un modelo de **predicción espacial**, si se considera que la variabilidad de las variables ambientales a estudiar (en este caso las temperaturas) está determinada por un conjunto finito de observaciones y éstas siguen exactamente alguna ley física conocida, si el algoritmo para desarrollar esta ley se conoce al detalle, entonces los valores en las ubicaciones objetivo podrán predecirse de manera exacta (Hengl, 2009). No obstante, en la realidad estas condiciones nunca se dan, por lo que la creación de modelos de



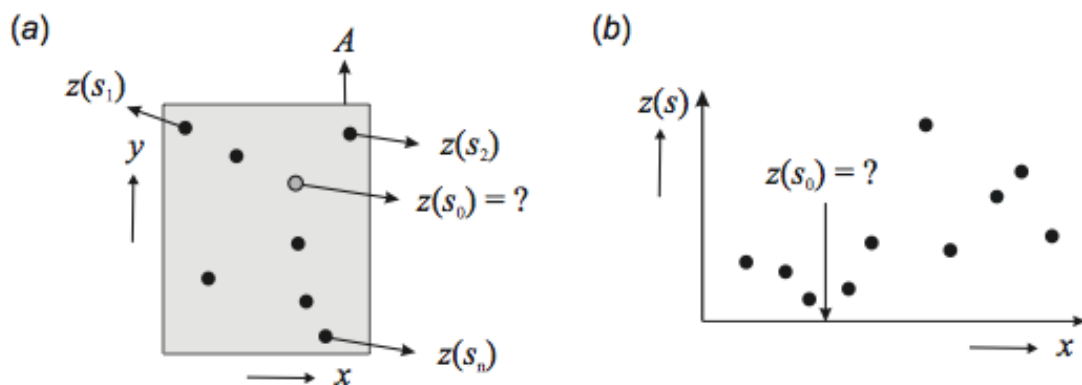
distribución se realiza para aproximar de la mejor manera posible los valores finales a los observados inicialmente.

De otra manera, siguiendo el desarrollo matemático de Hengl (2009) se puede demostrar la anterior afirmación: Dado un grupo de observaciones de una variable  $Z$ , expresada como  $z(s_1), z(s_2), \dots, z(s_n)$ , donde  $s_1 = (x_i, y_i)$  es la ubicación y  $x_i$  y  $y_i$  son las coordenadas en el espacio geográfico, y  $n$  es el número de observaciones. El dominio geográfico del área de estudio puede ser expresado como  $\mathbb{A}$ .

Asumiendo que las muestras son representativas, no-preferentes, y consistentes, los valores objetivo en las nuevas ubicaciones  $s_0$  pueden derivarse utilizando un **modelo de predicción espacial**. En términos estadísticos, un modelo de predicción espacial dibuja realizaciones (o el más probable o equiprobable conjunto de realizaciones), del elemento de interés dada una lista de inputs (u observaciones iniciales):

$$\hat{z}(s_0) = E\{Z|z(s_i), q_k(s_0), \gamma(h), s \in \mathbb{A}\}$$

Donde  $z(s_i)$  es el mconjunto de datos con las localizaciones iniciales;  $\gamma(h)$  es el modelo de covarianza que define la estructura de autocorrelación espacial, y  $q_k(s_0)$  es la lista de predictores determinísticos, también conocidos como *covariables* o variables explicatorias, las cuales requieren estar disponibles para todas las ubicaciones  $\mathbb{A}$ . En otras palabras, un modelo de predicción espacial comprende una lista de procedimientos para generar predicciones de valores en unas localizaciones determinadas, dados los datos de calibración (los iniciales), y el dominio espacial de interés (el área de estudio).



**Figura 10.** La predicción espacial es un proceso de estimación del valor (cuantitativo) en aquellas ubicaciones no visitadas dentro del área cubierta por las observaciones existentes: (a) esquema en un espacio 2D, (b) valores de una variable en un espacio unidimensional.

### ***3.4.2 Modelización del dato climático en GRID a escala temporal diaria***

A pesar de que la creación de una colección de cartografías ingente como es este caso *a priori* no aporta facilidad en el tratamiento de los datos, es necesaria para observar las diferencias espaciales de distribución de las variables climáticas a gran escala.

El método elegido para interpolar los datos climáticos observados a los puntos de GRID creados en el interior del área de estudio es una combinación de algunos de los tipos de interpoladores ya existentes con características comunes a éstos:

- Es un método local, ya que considera los observatorios cercanos delimitados por un radio de acción determinado (los más cercanos al área de interpolación)
- Es inexacto, ya que no mantiene los valores observados de los observatorios originales, sino que realiza una modelización de la influencia de los diferentes parámetros geográficos (latitud, longitud y altitud) que han tenido en la variable climática ese día concreto.
- Tiene en consideración la autocorrelación espacial porque la modelización pondera únicamente los observatorios cercanos, que siempre tendrán mayor peso en la determinación de la variable climática.
- Integra variables independientes y las pondera en función de su peso en la determinación de los valores observados.

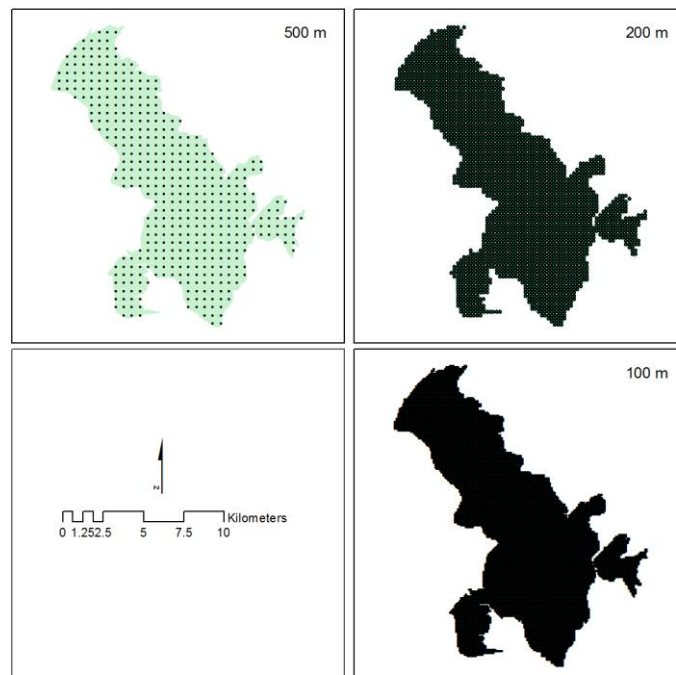
El proceso de cálculo se realiza para cada punto de GRID (para cada pixel) y consta de dos fases:

1. Para cada día y cada punto, calcula las ecuaciones de regresión polinomial de orden 1 (lineal) a 4 utilizando como variables independientes los parámetros

geográficos de cada par de coordenadas, y determina (en función de su valor de  $r^2$ ) cuál será utilizada para calcular el valor predicho de ese punto<sup>1</sup>.

2. Aplica a cada punto, en cada día, la ecuación correspondiente para obtener un valor predicho. Este predicho no puede ser estandarizado, ya que no existe ningún valor observado en el punto de destino al que ajustarse.

Un GRID es un conjunto de pares de coordenadas equidistantes entre sí que se ubican en una determinada superficie, y cuya conversión en píxeles de lado igual a la equidistancia entre los puntos da lugar a un raster. En el presente trabajo se ha creado, con el software QuantumGIS una malla regular de puntos de 100 metros sobre el área de estudio, lo que produce un total de 11.179 pares de coordenadas que se ubican en el interior del P.N. del Moncayo.



**Figura 11.** Mallas de GRID calculadas para distintas resoluciones en el interior del PN del Moncayo, se eligió la más fina (100 metros).

Una vez calculados los valores predichos para un día, y para todos los puntos del GRID, se pueden tratar de dos formas:

1. Se acumulan en una matriz multidimensional para su posterior consulta (Apartado 3.5)

---

<sup>1</sup> En este caso, se escoge una única función polinomial para todo el área de estudio, ya que los vecinos son los mismos para todos los puntos. Sin embargo, esta función sí varía para cada día del periodo considerado.

2. Se escriben a un fichero ASCII integrando la información espacial.

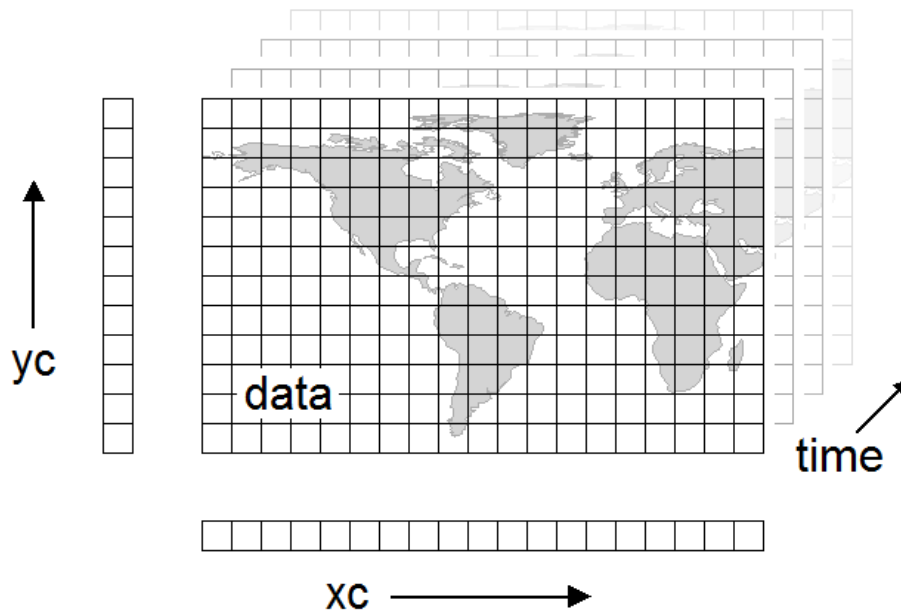
Con la primera opción es posible disponer de una base de datos consultable en tres dimensiones e incluso representable en Sistemas de Información Geográfica y otro tipo de software capaces de leer este tipo de archivos (p.ej. librerías GDAL en cualquier software OpenSource; Panoply (Schmunk)). Con la segunda opción se puede generar una cartografía en prácticamente cualquier SIG de distribución libre o propietaria. El archivo ASCII ha de tener el siguiente formato (donde xxx es numérico):

```
<NCOLS xxx>
<NROWS xxx>
<XLLCENTER xxx | XLLCORNER xxx>
<YLLCENTER xxx | YLLCORNER xxx>
<CELLSIZE xxx>
{NODATA_VALUE xxx}
row 1
row 2
.
.
row n
```

### 3.5 Creación de una base de datos en NetCDF de la información climática

Considerando el periodo completo de datos desde 1950 hasta 2012 en valores diarios, creando una cartografía por cada día, el resultado es de 23.011 grids por cada variable (temperatura máxima y mínima), lo que produce una **colección final de 46.022 matrices de datos** que, obviamente, no son representables de una manera legible en ningún formato ni equipo informático. Es por ello que el método más adecuado para almacenar toda esta información tiene que contener una estructura que permita la consulta en las dimensiones espacial y temporal al mismo tiempo.

Este modo de almacenamiento es NetCDF, un formato que almacena, de forma tridimensional (longitud, latitud, tiempo) variables continuas en una matriz orientada, es decir, superpone cada capa ráster para guardarla en un único archivo.



**Figura 12.** Representación gráfica del modo de almacenamiento de datos en formato NetCDF. *North American Regional Climate Change Assessment Program* (<http://www.narccap.ucar.edu/users/user-meeting-08/handout/handout.html>)

Este formato es muy popular en las bases de datos climáticas, siendo usado por el CRU (*Climate Research Unit*) o el NOAA (*National Oceanic and Atmospheric Administration*) entre otros. Su ventaja reside en que permite hacer consultas personalizadas, ya que a través de la referencia de las 3 dimensiones (fila, columna, hoja) es más fácil extraer información de una zona en concreto para un periodo determinado sin necesidad de mostrar toda la información al mismo tiempo.

Para crear un archivo NetCDF en lenguaje R, se utiliza la librería `netCDF` (<http://cran.r-project.org/web/packages/netcdf/>). Una vez que se crean la matrices de información sobre la variable climática a representar en toda la dimensión temporal, se crea un objeto `netcdf`:

```
create.ncdf( filename, vars )
```

`filename`: nombre del archivo netCDF file a crear.

`vars`: puede ser un objeto de clase `var.ncdf` describiendo la variable a crear, o un vector con los objetos a crear.

A continuación se introduce la variable:

```
put.var.ncdf( nc, varid, vals, start=NA, count=NA )
```

nc: objeto de clase `ncdf` (como retorno de cualquier función `open.ncdf()` o función `create.ncdf()`), indicando el archivo en el que escribir.

varid: variable a escribir los datos. puede ser una cadena de texto con el nombre de la variable, un objeto de clase `var.ncdf`, o el campo "id" de un objeto `var.ncdf`.

vals: valores a escribir.

start: vector de índices indicando dónde empezar a escribir los valores (empezando en 1). La longitud de este vector tiene que ser igual a las dimensiones de la variable. El orden es X-Y-Z-T (p.ej., la dimensión temporal es la última). Si no se especifica, se empieza a escribir en el inicio del archivo (1,1,1,...).

count: vector de enteros indicando el número de valores a escribir a lo largo de la dimensión (el orden es X-Y-Z-T). la longitud de este vector debe ser igual al número de dimensiones que tiene la variable. Si no se especifica y la variable no tiene dimensión ilimitada, se escribe toda la variable. SI la variable tiene una dimensión ilimitada, tiene que ser especificada. Un caso especial, el valor "-1" indica que todas las entradas a lo largo de la dimensión deberían ser escritas.

Finalmente se cierra el archivo:

```
close.ncdf( con )
```

con: Objeto de clase `ncdf` (como retorno de cualquier función `open.ncdf()` o función `create.ncdf()`), indicando qué archivo se leerá.

### 3.6 Programación y paralelización en lenguaje R

Todos los cálculos realizados sobre los datos para el control de calidad, el relleno, y la creación del GRID han sido programadas en lenguaje R. Este lenguaje de programación es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de (R Development Core Team, 2000):

- almacenamiento y manipulación efectiva de datos,

- operadores para cálculo sobre variables indexadas (Arrays), en particular matrices,
- una amplia, coherente e integrada colección de herramientas para análisis de datos,
- posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y
- un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

A través de la creación de funciones, los datos iniciales en formato de hoja de cálculo (filas y columnas) se van tratando para obtener, en pasos sucesivos, los diferentes productos necesarios en cada momento.

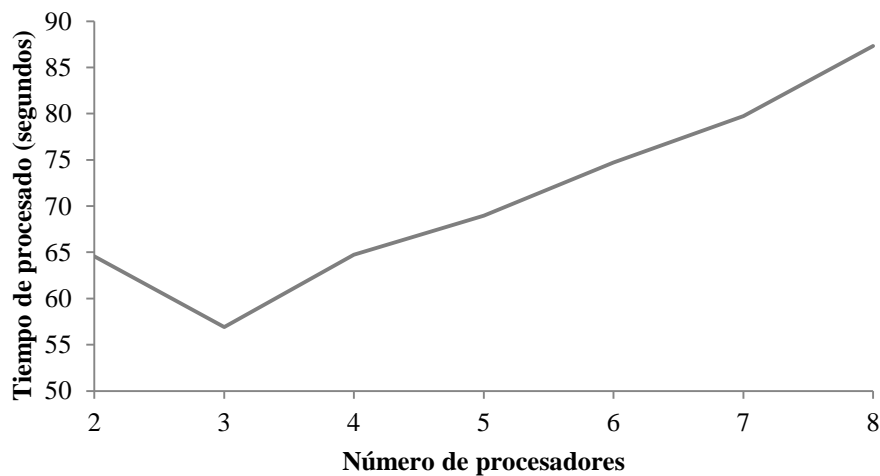
Como el volumen inicial de datos ya es de por sí de gran tamaño, los productos intermedios van creciendo y ralentizando la velocidad de cálculo, por esta razón se ha elegido trabajar con una **programación paralela**. Este método permite, a través de una serie de pequeños cambios en la escritura del código, aprovechar todos los recursos disponibles del hardware. La filosofía de la computación paralela reside en dividir el trabajo asignado por la programación en tantas partes como núcleos de procesamiento de que dispone el hardware con el fin de optimizar el tiempo de procesamiento.

Existen varias librerías o paquetes en el repositorio de R que son capaces de generar este tipo de trabajo. En este caso se ha optado por la librería *snowfall* (<http://cran.r-project.org/web/packages/snowfall/snowfall.pdf>) por su facilidad de uso y por los buenos resultados que genera. En cualquier caso, el proceso de programación paralela ha seguido los siguientes pasos:

- 1) Inicializar un clúster con el número de procesadores deseados/disponibles
- 2) Ejecutar el código
- 3) Detener el clúster

A pesar de las evidentes ventajas de este tipo de programación, en ocasiones es mejor utilizar la programación clásica o usar un número menor de procesadores que los disponibles en el hardware, ya que el tiempo dedicado a dividir el proceso primario en

otros secundarios, incluyendo el tiempo de lectura/escritura puede ser mucho mayor que si se ejecuta el código sin paralelizar.



**Figura 13.** Representación gráfica del tiempo de procesamiento de la función de relleno para la validación de los datos observados (Apdo. 4.1.2). A mayor número de procesadores asignados, disminuye la eficiencia de la programación.

En el ejemplo representado en el gráfico anterior, cuando se incrementa la disponibilidad de procesadores para una función concreta que no requiere demasiado esfuerzo de hardware, es menos eficiente cuando se le asignan más de 3 procesadores. Así pues, para cada función ha sido necesario hacer un estudio de optimización con el fin de obtener los mismos resultados en un periodo de tiempo abaricable.

Toda esta explicación no es trivial, ya que el tiempo de procesamiento ha sido clave en este proyecto, el resultado final del relleno de temperaturas máximas y mínimas ha producido más de 15 millones de datos (más de 23.000 registros para cada uno de los 700 observatorios fragmentados). Sin una programación planificada y optimizada de forma eficiente, el estudio hubiera sido inabordable.



## **4. RESULTADOS**

Los resultados del trabajo pretenden responder a los objetivos planteados al inicio. La respuesta al objetivo principal respecto de la obtención de una cartografía de alta resolución y espacialmente continua del Parque Natural del Moncayo en la que se representen las temperaturas máxima y mínima diaria, se responderá al final de las conclusiones, pues su comprobación depende necesariamente de la consecución de los objetivos secundarios, que se centraban en:

- I. Reconstrucción de los observatorios del P.N. del Moncayo y áreas limítrofes
  - a. Control de calidad de las series de temperatura
    - i. Cálculo de anomalías
    - ii. Homogeneización de las series
  - b. Relleno de lagunas de información y reconstrucción de las series
- II. Creación de cartografías de alta resolución a escala diaria
  - a. Modelización del dato climático en GRID a escala temporal diaria
- III. Creación de una base de datos en NetCDF de la información climática

## **4.1 Homogeneización y relleno de lagunas de los observatorios del P.N. del Moncayo y áreas limítrofes**

### ***4.1.1 Control de calidad de las series de temperatura***

Se ha trabajado con un número inicial de 66 observatorios de temperatura y sobre un total de 23.011 días desde el 1 de enero de 1950 hasta el 31 de diciembre de 2012.

La completitud de los datos iniciales respecto al periodo total de estudio se resume en la siguiente tabla:

**Tabla 1.** Disponibilidad inicial de datos de temperatura en las 66 estaciones seleccionadas

	Datos respecto al total
<b>Media</b>	25.7 %
<b>Estación más completa</b>	98.1 %
<b>Estación más incompleta</b>	0.13 %

El cálculo y posterior eliminación de datos anómalos ha extraído de la base de datos original aproximadamente entre un 5% y un 7% del total de los datos disponibles

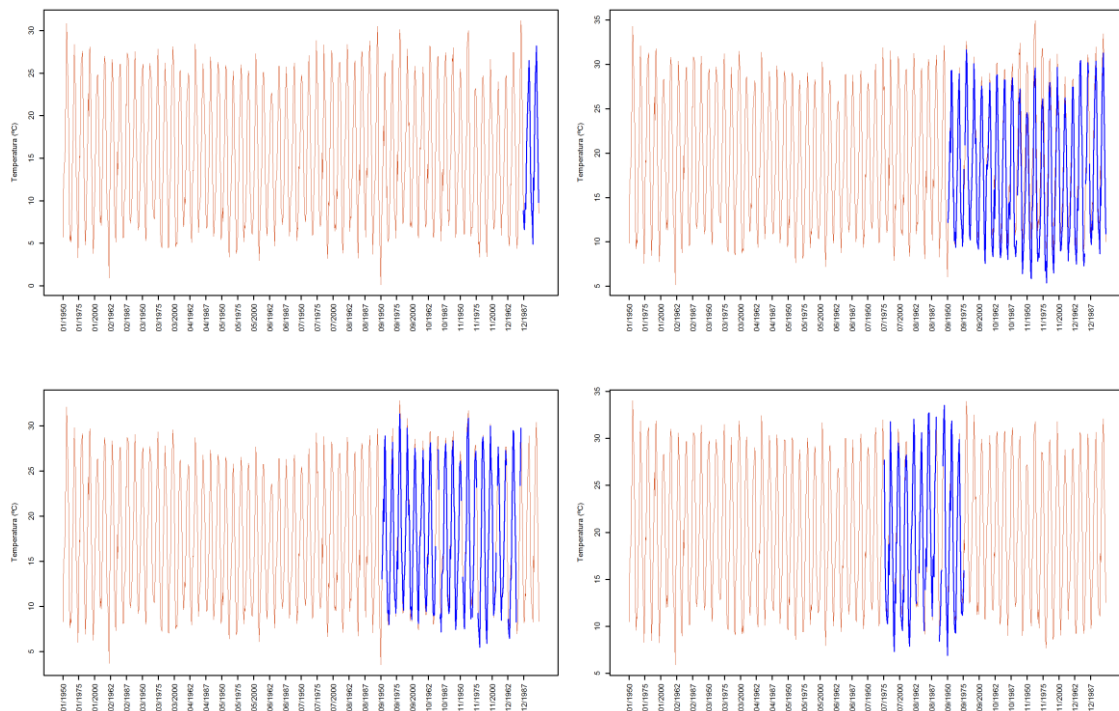
en cada serie, siendo prácticamente las mismas cifras tanto en temperaturas máximas como en mínimas.

Tras el cálculo de las anomalías se ha procedido a la homogeneización de las series mediante segmentación, de las que se han obtenido, a partir de las 66 series iniciales para las dos variables, 360 series segmentadas en el caso de las máximas y 363 series segmentadas en las mínimas. De todas estas series, 15 en el caso de las máximas y 17 en las mínimas se han eliminado porque el periodo segmentado ha sido inferior a 30 días, por lo tanto no han podido utilizarse para ser homogeneizadas ni estandarizadas.

Una vez aplicado el test SNHT no se han detectado periodos inhomogéneos en las series segmentadas, lo que confirma la idoneidad de la fragmentación de las series.

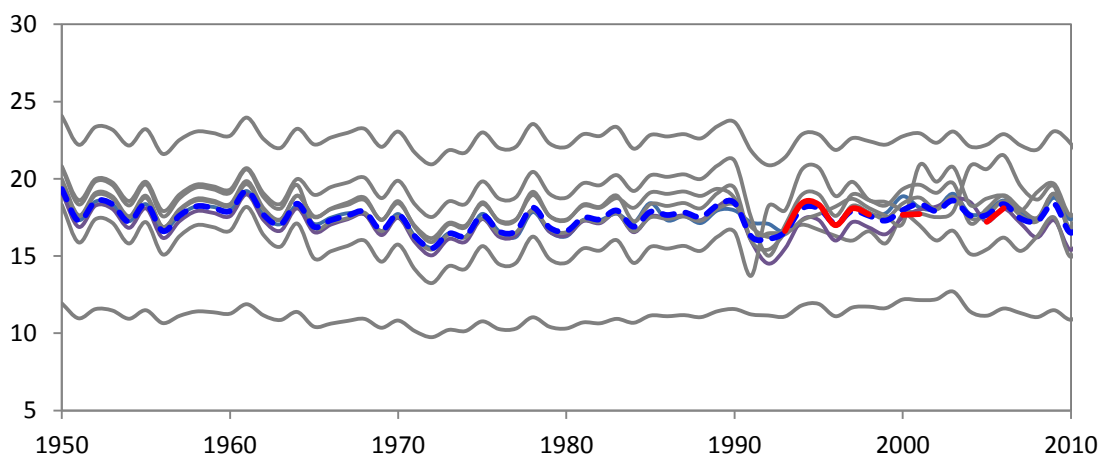
#### ***4.1.2 Relleno de lagunas de información y reconstrucción de las series***

El relleno de lagunas de información aplicado a las 345 series de temperatura máximas y a las 346 de temperatura mínima se ha llevado a cabo de manera satisfactoria mediante 2 iteraciones en ambos casos. El relleno final, producto de la modelización de todos los vecinos rellenados ha completado los 23.011 días de datos para todas las series.



**Figura 14.** Resultados del relleno de datos en 4 estaciones (medias mensuales). De arriba abajo y de izquierda a derecha: Agramonte; Maidevera (Embalse); San Martín del Moncayo; y Tarazona (Ayuntamiento). En azul el dato original, y en rojo el dato producto del relleno.

El producto final del proceso de homogeneización y relleno de temperaturas produce una base de datos de mayor tamaño que la inicial. Como consecuencia del proceso de segmentación, de una sola serie se generan varias con diferente media y desviación típica, en definitiva, con datos diferentes. A pesar de que la ubicación del observatorio es la misma, y por lo tanto sus parámetros de control (altitud, latitud, y longitud) también, la inclusión de diferentes vecinos en el proceso de relleno para cada día, produce que el dato final sea diferente en todos o casi todos los casos. Algunas de las series originales se han segmentado en más de diez partes, lo que produce una cantidad extraordinaria de datos que dan mucho más juego en estudios climáticos posteriores. Mientras que lo más adecuado para trabajar toda la serie sería hacer un promedio de todos los fragmentos para cada día, cuando se trata de trabajar con periodos de tiempo concretos es recomendable seleccionar o dar prioridad a través de algún tipo de ponderación a aquel fragmento que contenga mayor número de datos originales, así se pierde menos información original, y la no existente se complementa con el relleno estadístico realizado.



**Figura 15.** Una misma serie (San Martín del Moncayo (*medias anuales*)) se fragmenta en el proceso en varios segmentos (9 en este caso). En rojo los datos originales, en azul el promedio de los segmentos rellenos, en gris las segmentaciones por separado.

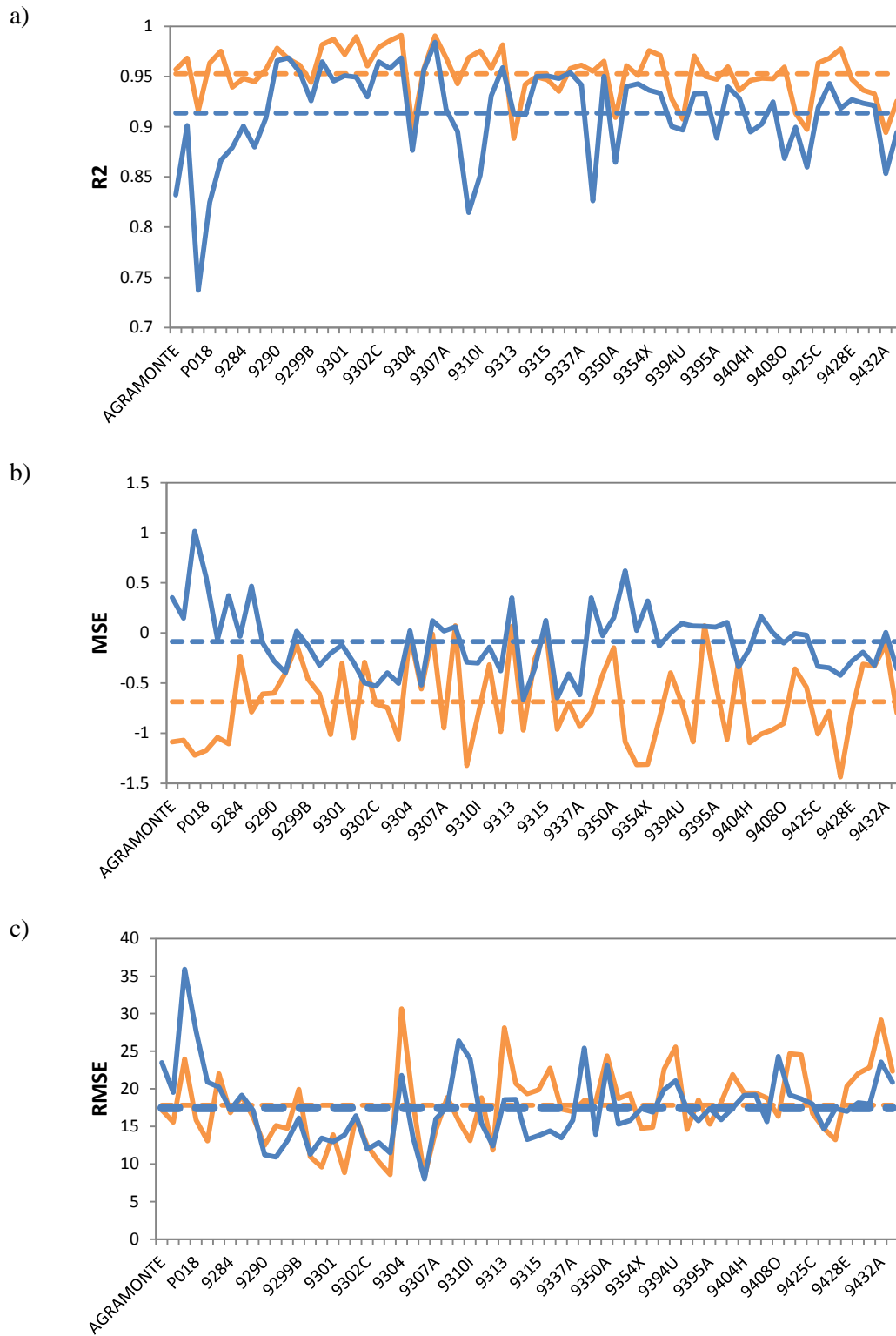
Respecto a la calidad del relleno de los datos, únicamente se puede comprobar comparando los datos observados con los predichos para ese mismo día en esa misma serie. En el actual trabajo se ha preferido conservar los datos originales, es decir, solamente se han calculado predichos para aquellos días en los que no existía una observación originalmente. Sin embargo, como método de control, se han calculado los predichos para todos los valores observados iniciales con el fin de asegurar la calidad del dato final.

Se ha calculado para cada método, por un lado el Error Absoluto Medio (MAE), que es una medida del error natural y no ambigua (Willmott & Matura, 2006) y que muestra los errores en las mismas unidades que la propia variable climática. Y por otro lado el Error Cuadrático Medio (RMSE), que es una medida usada habitualmente que determina la desviación respecto a los valores observados.

**Tabla 2.** Test estadísticos aplicados a los diferentes métodos de creación de cartografías.  $n$  es el número de observaciones;  $P_i$  es el valor predicho de un día concreto; y  $O_i$  es el valor observado en el mismo día.

Test	Formulación
Error Absoluto Medio (MAE)	$MAE = n^{-1} \sum_{i=1}^n  P_i - O_i $
Error Cuadrático Medio (RMSE)	$RMSE = \left( n^{-1} \sum_{i=1}^n (P_i - O_i)^2 \right)^{0,5}$

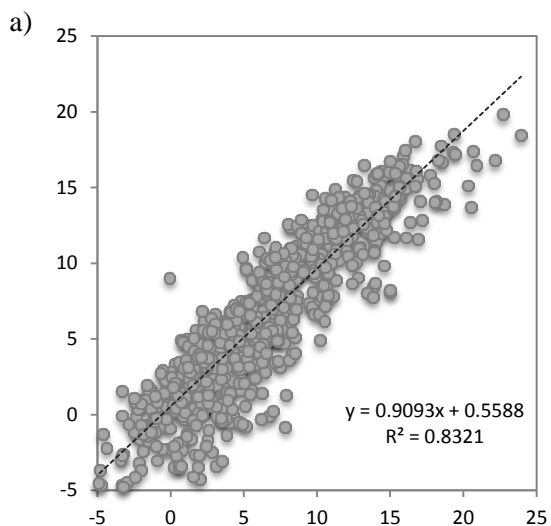
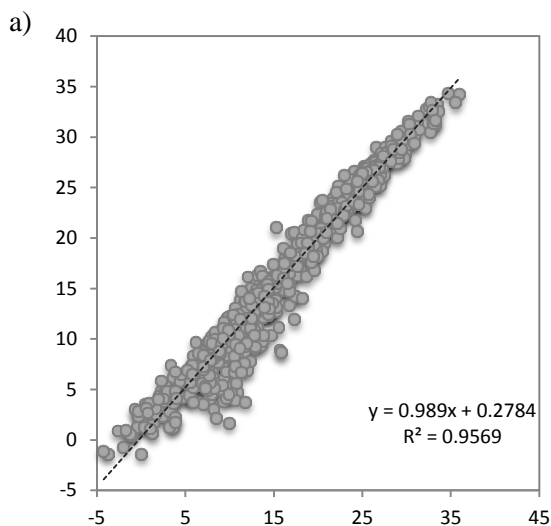
El ajuste de los valores predichos a los observados cuando se trabaja sobre el dato diario (del que se han modelizado todos los predichos) es algo inferior a la situación que representan las medias mensuales. Esta diferencia viene dada por el aumento de la variabilidad del clima con el aumento de la escala temporal. De esta manera, comparando por pares todos los observados iniciales con sus respectivos predichos, se obtiene un valor de  $r^2$  medio de 0.9529 para las máximas, y de 0.9137 para las mínimas, un ajuste bastante bueno teniendo en cuenta que se trata de dato diario.

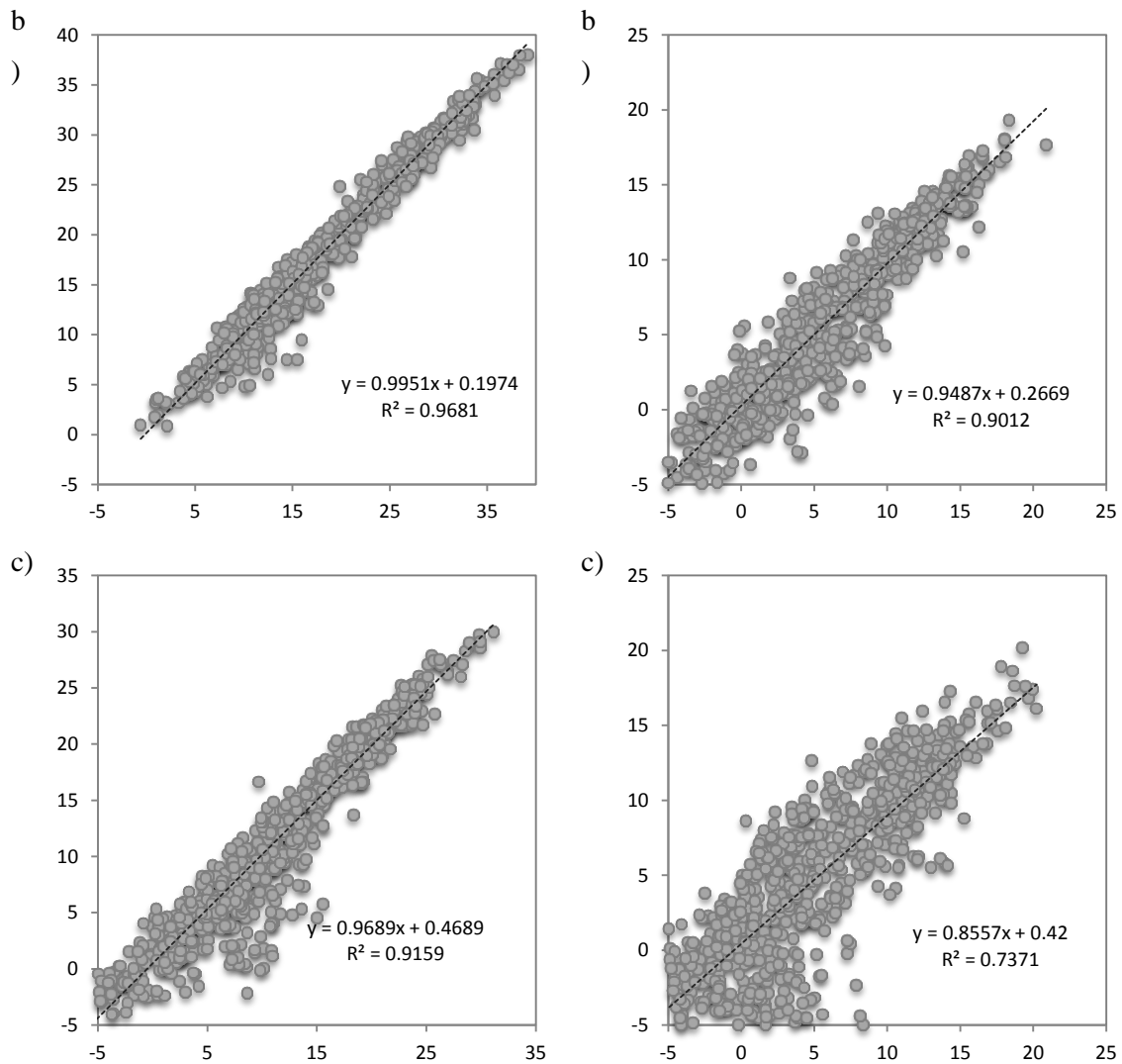


**Figura 16.** Comparativa Observados-Predichos en dato diario para todas las estaciones de temperatura máxima (naranja) y mínima (azul). a) Valores de  $r^2$  en correlación de Pearson; b) Valores de Error Medio Absoluto; c) Valores de Error Cuadrático Medio

Tras el análisis visual de los resultados globales para todas las estaciones, es importante destacar la importancia que tres de ellas tienen para los objetivos del trabajo. Los observatorios de Agramonte, Calcena, y Majada Baja llevan funcionando en el PN del Moncayo desde 2010. Tienen interés porque son los más cercanos al límite del Parque Natural, y por tanto son los más importantes para la modelización climática de la zona.

Los resultados de la modelización muestran unos resultados realmente buenos para las temperaturas máximas, y no tanto para las mínimas, especialmente en Majada Baja, donde el ajuste entre observados y predichos tiene apenas un coeficiente de correlación de 0.74. Esta estación se ubica a 1.798 metros de altitud, y esta zona es proclive a sufrir fenómenos de inversión térmica que no son bien detectados por el modelo, de hecho el problema seguramente esté en la dependencia de las temperaturas máximas sobre otros factores no considerados en el modelo, es decir, que la distribución de las mínimas no tiene que ver exclusivamente, o al menos no en gran medida, de los factores geográficos altitud, latitud y longitud, sino que es probable que influyan otros como las diferencias de presión entre zonas y días, la nubosidad, etc., parámetros que, por ausencia de datos disponibles, no se han introducido en esta primera versión del modelo. Las otras dos estaciones no tienen los coeficientes de ajuste de las temperaturas máximas, pero se pueden considerar buenos con un 0.83 para Agramonte (1090 metros) y con 0.90 para Calcena (811 metros).

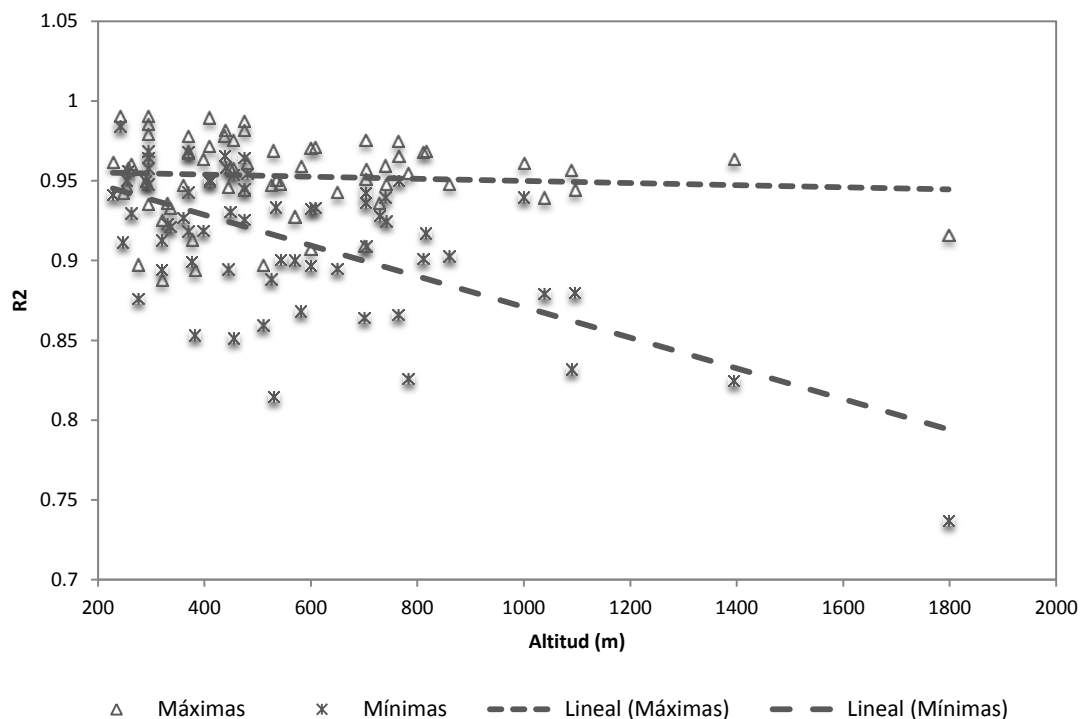




**Figura 17.** Comparativa Observados-Predichos en dato diario para las estaciones de temperatura máxima y mínima. a) Agramonte; b) Calcena; c) Majada Baja. Col. izquierda: Temp. máxima; Col. Derecha: Temp. mínima

Observando el ajuste de los datos de estas tres estaciones en el interior del PN del Moncayo, se observa que las temperaturas mínimas tienen más problemas para ser modelizadas correctamente que las máximas. En el siguiente gráfico se muestra la relación entre el ajuste por correlación entre predichos y observados, y la altitud. Como ya se ha comentado, es probable que se deba a la falta de parámetros explicativos de la distribución de esta variable sobre el territorio.





**Figura 18.** Comparativa entre la correlación Observados-Predichos, y la altitud de las estaciones trabajadas

Mientras en las temperaturas máximas el ajuste de los valores modelizados con los observados es más o menos constante en todas las altitudes, en las temperaturas mínimas hay una clara tendencia regresiva a medida que se aumenta el valor de altura. A pesar de esta tendencia, las peores correlaciones muestra valores por encima de 0.7, un valor muy bueno para tratarse de dato diario. En futuros trabajos, será necesario observar detenidamente esta tendencia y buscar si la causa de la misma se debe a la propia naturaleza de los datos, o si hay alguna parte del proceso de cálculo de los predichos en la que es necesario introducir algún factor corrector para obtener las temperaturas mínimas finales bien ajustadas.

## 4.2 Creación de cartografías de alta resolución a escala diaria

A través del método elegido se han creado los GRIDs de información climática a partir de los observatorios iniciales rellenos con el proceso explicado en el Apartado 3.4.2., con el fin de comprobar el funcionamiento del método se han realizado, para los observatorios de un mismo día, cartografías con diferentes métodos de interpolación.

Para el presente trabajo se han realizado diferentes cartografías para el mismo día mediante los siguientes métodos con sus respectivos parámetros

a) *Inverso de la Distancia (IDW)*

- Power = 2
- Maximum neighbours = 15
- Minimum neighbours = 10

b) *Funciones radiales*

- Kernel function = Completely Regularized Spline
- Kernel parameter = 0.0438085
- Maximum neighbours = 15
- Minimum neighbours = 10

c) *Kriging Simple*

- Optimize model = None
- Variable = Semivariogram
- Model Nugget = Enabled
- Model #1 = Type Stable

d) *Kriging Ordinario*

- Transformation Type = Normal Score
- Decluster before transformation = False
- Order of Trend Removal = None
- Optimize model = None
- Variable = Covariance
- Model Nugget = Enabled
- Model #1 = Type Stable

e) *Superficies de tendencia de orden 2*

- Polynomial Order = 2

f) *Superficies de tendencia de orden 4*

- Polynomial Order = 2

g) *Modelo de regresión múltiple con altitud, latitud y longitud sin corrección de residuales*

- Función de interpolación con los coeficientes obtenidos de la regresión múltiple:

(máximas)

$$-101.1832089 + ("altitud" * -0.064125371) - ("longitud" * 0.0000775066012126244) + ("latitud" * 0.0000794292140565995)$$

(mínimas)

$$-2838.177961 + ("elev" * 0.030612615) + ("longitud" * 0.000942506) + ("latitud" * 0.000489821)$$

*h) Modelo de regresión múltiple con corrección por residuales*

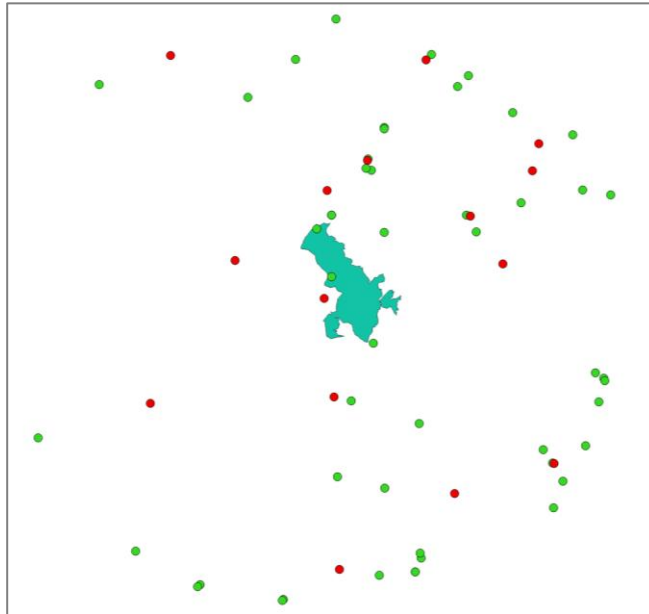
- A los resultados del método anterior se le suma la interpolación mediante IDW de sus residuales (Observados - Predichos)

*i) Método de interpolación desarrollado en este trabajo*

Este método se explica detalladamente en el apartado 3.4.2.

Estos métodos se han aplicado, como prueba de control, para el día 5 de febrero de 2011 tanto a las temperaturas máximas como a las mínimas.

Con el fin de obtener una validación de las cartografías generadas mediante cualquiera de los nueve métodos de interpolación aplicados, se ha utilizado una reserva del 20% de los observatorios iniciales (Test), con el 80% restante (Training) se han realizado las cartografías mostradas de temperatura máxima y mínima. Esta reserva de puntos Test se ha escogido de manera aleatoria sobre los observatorios de los que existía dato observado para no contaminar el resultado final de la evaluación del error.



**Figura 19.** Distribución de puntos Training utilizados para interpolar (verde) y Test utilizados para validar (rojo).

Los valores de error obtenidos demuestran que el método se encuentra dentro de los parámetros aceptables para realizar la cartografía, así que se ha aplicado a los datos rellenados.

El método utilizado muestra un error cuadrático medio de 1.973, mientras que los mejores resultados arrojan valores de 1.518 (Reg. sin residuales) y 1.623 (Splines 2), aunque en general se llevan bastante poco. En el caso de las máximas, los valores son un poco más altos, siendo el mejor de 1.602 (Reg. sin residuales), el valor de RMSE para el método utilizado es de 2.184. No obstante, es necesario destacar que estos valores se han obtenido con la comprobación de un día concreto, por lo que estos valores variarán en función del día que se tome como referencia, ya que la temperatura no se distribuye espacialmente con el mismo patrón todos los días. Es muy probable que en aquellos días en los que la distribución de la variable climática no responda exclusivamente a una relación lineal con las variables utilizadas, los valores de error disminuyan considerablemente respecto al resto de los métodos.

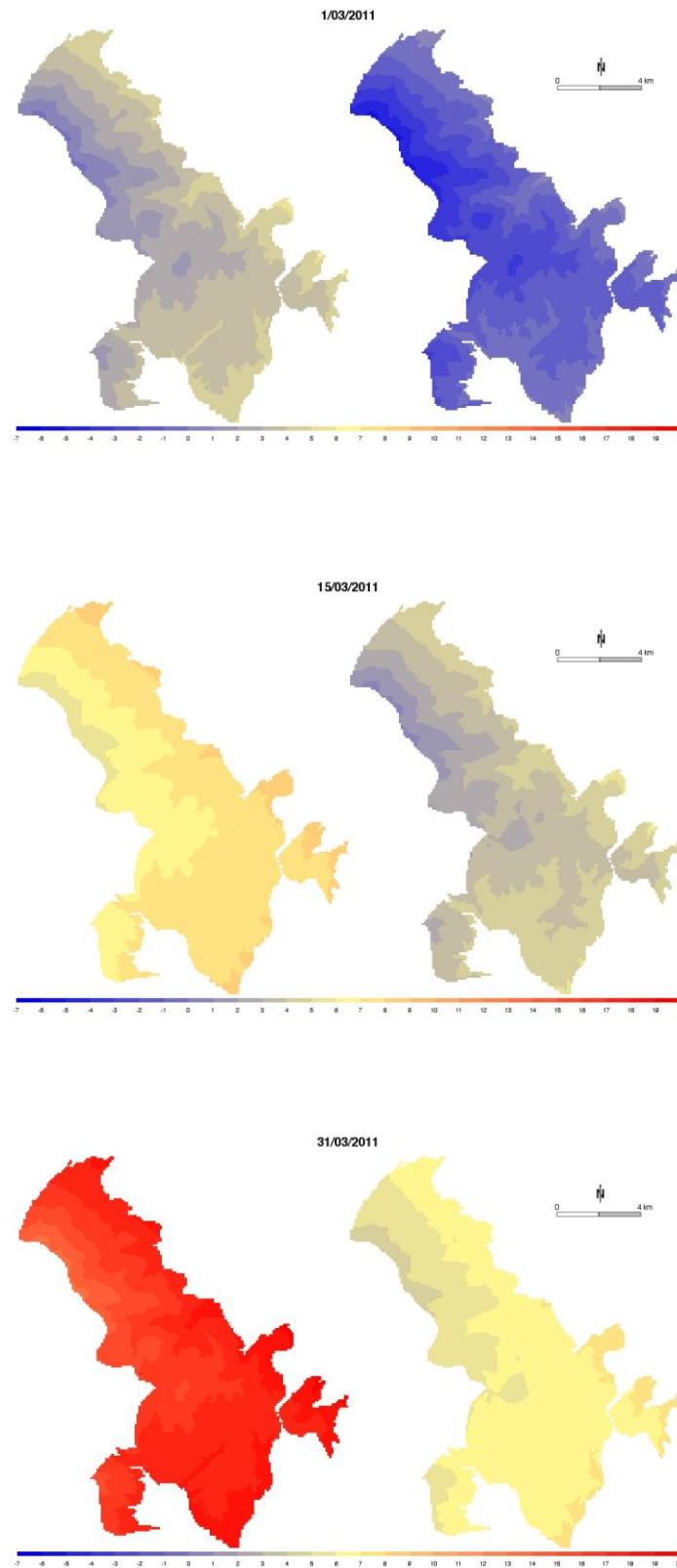
**Tabla 3.** Valores de error extraídos de los test estadísticos para cada tipo de interpolación.

Tipo de interpolación		Temp. máx.		Temp. mín.	
		MAE	RMSE	MAE	RMSE
a	IDW	0.633	2.057	1.165	2.110
b	RBF	0.605	1.705	1.019	1.951
c	Krig. Simple	1.081	2.162	0.874	2.281
d	Krig. Ordinario	0.477	1.672	1.007	1.878
e	Supf. Tend. Orden 2	0.270	1.623	0.584	1.958
f	Supf. Tend. Orden 4	-0.552	3.796	0.136	2.492
g	Regresión múltiple Sin Residuales	0.268	1.518	0.473	1.602
h	Regresión múltiple Con Residuales	0.594	1.965	0.958	2.011
i	Método Propio	0.339	1.973	1.307	2.184

Mediante el modelo creado para este trabajo, se han creado grids de temperatura máxima y mínima diferenciados para cada uno de los días, para el área que ocupa el P.N. del Moncayo. Se ha creado una malla con un total de 29.704 puntos de separación 100 metros, de los cuales se han rellenado 11.179 (los ocupados por el área de estudio) y el resto con un número que representa el dato vacío, todo ello para cada uno de los 23.011 días. El total de datos generados para todo el periodo para las dos variables ha sido de más de 1.360 millones de datos.

A continuación se presentan tres ejemplos de cartografías calculadas para tres días diferentes en un mismo mes, que representan un día de invierno (1 de marzo), la transición hacia la primavera (15 de marzo), y el inicio de la misma (31 de marzo). Se puede observar como la variabilidad entre las temperaturas máximas y las mínimas aumenta con la llegada de los días más cálidos, mientras que la diferencia entre las máximas y las mínimas es mucho menor durante la época invernal.

Además, se ve que en los días con mayor oscilación térmica el patrón de distribución espacial es muy diferente entre las temperaturas máximas y mínimas, mientras que contra menor oscilación térmica, el dibujo de la distribución es más similar. Este hecho se debe a que la variabilidad en altura en el área de estudio es muy grande, y eso se refleja en los contrastes de las temperaturas diarias, aumentando en altura cuanto mayor sea el contraste diario entre la máxima y la mínima.



**Figura 20.** Ejemplo de cartografía de temperaturas máximas (izquierda) y mínimas (derecha) para tres días concretos.

### **4.3 Creación de una base de datos en NetCDF de la información climática**

Tal y como se especifica en la metodología en el apartado 3.5, se han creado dos archivos en formato NetCDF (.nc) que contienen respectivamente la información sobre la temperatura máxima y la temperatura mínima. Cada uno de ellos contiene exactamente el mismo número de datos, el que corresponde al relleno de información en cada uno de los píxeles del área que ocupa el P.N. del Moncayo, y con una profundidad de 23.011 días, los correspondientes al periodo desde el 1 de enero de 1950 hasta el 31 de diciembre de 2012.

Este formato hace mucho más manejable la consulta, pues es relativamente fácil mediante lenguaje R extraer la información en cualquiera de sus dimensiones, así como representarlo de manera espacial, e incluso mediante visores *nc*, en una transición temporal. El inconveniente quizá es que el peso de cada archivo es de casi 2.5Gb, con lo que el manejo puede ser incómodo en función del hardware disponible.

## **5. CONCLUSIONES**



El método propuesto da respuesta a la dificultad del tratamiento estadístico del dato diario en climatología. A partir de un análisis exhaustivo a escala espacial, y no temporal, es posible desde detectar las anomalías, hasta rellenar adecuadamente los datos faltantes en las series incompletas.

En comparación a otros trabajos de escala diaria, Herrera (2011) acepta como validación un valor de R de 0.67 ( $R^2=0.45$ ) para dato diario en su base de datos de temperatura Spain02, Cano (2004) no ofrece ningún valor de correlación o de error medio para todo el periodo estudiado, y El-Kenawy (2013) solamente da valor de R para dos estaciones, 0.92 y 0.93 ( $R^2=0.85$  y  $0.86$  respectivamente), además sugiere que su método es bueno porque al comparar las series reconstruidas con las que no se han usado del entorno se parecen, en este caso asume una homogeneidad espacial del clima, ya que para reconstruir sus series finales ha elegido previamente aquellas cercanas que más se le parecen. En este trabajo se han obtenido valores medios de  $R^2$  de 0.95 para las temperaturas máximas, y de 0.91 para las mínimas, lo que indica que funciona mejor que otros métodos con similares objetivos.

A diferencia de otros métodos de tratamiento estadístico de series climáticas, es importante destacar que el tratado aquí se ha **desarrollado para crear bases de datos climáticas**, y que el tiempo de procesado de todos los datos es muy grande, ya que las sucesivas iteraciones consumen recursos en grandes cantidades, quizá para obtener tendencias en largos periodos de tiempo, sería más eficiente trabajar con dato mensual y extraer cartografías mediante algún método de interpolación ya implementado en un SIG, en el que, por la naturaleza de los datos, el error medio será lo suficientemente bajo como para aceptar el resultado.

Se han encontrado coeficientes de determinación más bajos en estaciones a mayor altura (p.ej.: Majada Baja, 1.800 metros). El ajuste de los datos en estas estaciones es menor, ya que el polinomio utilizado para interpolar el valor final está construido en muchas ocasiones con pocos datos (cada día cambia el número de vecinos) y todos ellos a menor altitud de estas estaciones candidatas a gran altitud. Es necesario revisar el método en esta parte del proceso y modificar el paso en el que el modelo construido sobre la altitud elige curvas de ajuste que producen datos fuera de la realidad. Para solucionar este inconveniente de manera provisional, se ha asignado en todos los días el modelo de orden 1 (regresión lineal), ya que predice de manera menos irreal el valor

final en altura para aquellas ubicaciones susceptibles de sufrir *overfitting* en su predicción (aquellas que están fuera del rango de los *input*).

Por otra parte, la estandarización de fragmentos que no contienen un año completo de observaciones, corre el riesgo de estar sesgada por los propios datos, ya que al estandarizar por ejemplo solamente con datos de verano o de invierno, modifica la variabilidad de la serie de referencia que rellena. Una posible corrección, a falta de realizar pruebas concluyentes, puede ser la de calcular una serie inicial de referencia a través de la modelización de los datos brutos, y extraerles su ciclo anual (media de cada día para todos los años), que después se añadiría a la serie de anomalías calculadas.

El **método cartográfico** utilizado puede proporcionar mapas de temperatura diaria con ajustes muy altos en términos de valores y resolución espacial. Programado en lenguaje R, el método es eficiente y automatizado, lo que facilita la tediosa tarea de cartografiar uno a uno cada día del periodo completo. A diferencia de otros métodos de interpolación de temperatura (Joly, 2003), la utilización de los metadatos inherentes a los valores climáticos supone una facilidad añadida, ya que no es necesario modelizar ningún otro parámetro geográfico, con lo que la cartografía es completamente automática.

En lo que se refiere a los resultados prácticos, la aplicación está limitada a un ejemplo (un día concreto), no obstante la bibliografía (Joly, 2003) muestra que la distribución de la temperatura diaria está relacionada con micro-factores locales, con lo que el ajuste estadístico de las condiciones diarias para que se dé determinada temperatura que se propone en el modelo, es más que adecuado. Además, tal y como señala Pita (2007), el ambicioso objetivo de la simulación del funcionamiento del sistema climático a escala planetaria plantea automáticamente la necesidad de nuevas fuentes de información hasta ahora no reclamadas o no disponibles. En la mayor parte de los casos se trata de información relativa a variables no consideradas en el ámbito de la climatología, tales como la cobertura vegetal, la temperatura y salinidad de los océanos, la extensión de la cubierta de hielo, el contenido de humedad del suelo, etc. Para ellas, la teledetección constituirá la principal fuente de información, alimentada por nuevos sensores cada vez más aptos para ofrecer información muy diversa.

Por otro lado, el resultado de la Figura 20 muestra tres cartografías correspondientes a la temperatura máxima y mínima de tres días concretos para el P.N. del Moncayo. El modelo utilizado aplica como ya se ha dicho, para todos los días, una regresión lineal a la hora de calcular los valores predichos dependientes de la altitud, a diferencia de los parámetros longitud y latitud, cuyo modelo varía en órdenes de polinomio de 1 a 4, en función del mejor valor de su  $R^2$ . El efecto producido tiene que ver con los datos disponibles y el modelo de aplicación, es decir, que la predicción de variables climáticas en localizaciones espaciales fuera del rango de los valores observados (p.ej. El rango de observaciones es de 600 a 1800 metros de altitud, y el punto en el que se quiere predecir el dato está a 2.200 m) mantiene una incertidumbre que no puede solucionar el modelo de polinomios, ya que produce valores finales alejados de la coherencia espacial prevista como consecuencia de la extensión de la curva del orden de polinomio correspondiente hacia valores extremos.

El **procedimiento de modelización sistemática** se presenta como un método muy útil para definir potenciales ecológicos en áreas de alta diversidad topográfica, pues es capaz de reflejar cambios en la temperatura en pequeñas zonas de alta rugosidad orográfica.

El control de calidad de las series de temperatura es el proceso básico para obtener series de referencia de calidad. La aplicación del mismo señala que entre un 5 y un 7% de los datos originales son susceptibles de ser eliminados por tratarse de datos anómalos, aunque estas cifras varían en función de la longitud de las series observadas y la calidad de las mismas. La detección y eliminación de anomalías se realiza únicamente teniendo en cuenta la influencia de las variables geográficas en la temperatura de cada día, por tanto el método es el más transparente de cuantos se han publicado hasta la fecha, ya que no considera ni la estructura temporal de la serie, ni el ajuste de los valores de inhomogeneidad detectados a nivel mensual, factores que habitualmente modifican las series candidatas para obtener una señal climática similar a la de sus series de referencia, cuando no tiene que ser necesariamente de esta manera.

Las series resultantes del relleno tienen todos los días del periodo con un dato modelizado que mantiene la variabilidad y la media de la serie original, aunque se da en algún caso, especialmente en las temperaturas mínimas y en estaciones a mayor altura, que el ajuste no es tan perfecto como en las temperaturas máximas, hecho que puede deberse a una estandarización sobre periodos cortos, inferiores a un año, que no reflejan

la variabilidad climática total que representa ese punto en el espacio. Este procedimiento será corregido en versiones posteriores del método.

La automatización de la cartografía mediante el mismo proceso de relleno que el utilizado para los observatorios individuales se presenta como una solución adaptada para cada localización espacial, con la ventaja de que el mapa final no es producto de un ajuste estadístico del área de estudio global (como hacen la mayoría de los métodos de interpolación), sino de los metadatos geográficos individuales de cada par de coordenadas que se quiere rellenar.

Actualmente, el modelo de interpolación utilizado está en revisión, intentando mejorar en cada momento la predicción, especialmente para localizaciones en altura, fuera del rango de las altitudes disponibles en los valores observados.

*Este estudio se ha realizado en el marco de los proyectos de investigación CGL2012-31668, y CGL2011-28255, ambos financiados por el Ministerio de Economía y Competitividad y cofinanciados con fondos FEDER; y el programa de Grupos de Investigación Consolidados “Grupo Clima, Agua y Cambio Global” del Gobierno de Aragón y Fondo Social Europeo.*

## **6. REFERENCIAS**

- Aguilar, E. et al (1999) "Control de calidad y proceso de homogeneización de series térmicas catalanas". En RASO, J.M. y MARTIN-VIDE, J.(Eds): *La climatología española en los albores del siglo XXI*. Publicaciones de la Asociación Española de Climatología (AEC), Serie A, nº 1 pp15-23.
- Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., y Wieringa, J. (2003): "Guidance on metadata and homogenization", *WMO-TD* No. 1186, (WCDMP-No. 53), pp. 51.
- Alexandersson, H. (1986): "A homogeneity test applied to precipitation data". *Journal of Climatology*. Vol. 6, pp. 661-675.
- Álvarez, E., Espejo, F., Cortés, F.J., Lafragüeta, C, Serrano-Notivoli, R. (2011): *Caracterización sinóptica de los procesos convectivos en el interior del Nordeste peninsular*. Agencia Estatal de Meteorología, <http://www.aemet.es/es/divulgacion/publicaciones/detalles/sinoptica>
- Brunetti M, Maugeri M, Monti F, Nanni T. (2006). "Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series". *International Journal of Climatology* 26: 345–381.
- Borough PA, McDonnell RA (1998) *Principles of Geographical Information Systems*. Oxford University Press, Oxford
- Cano R, Gutierrez JM. (2004). "Relleno de lagunas y homogeneización de series de precipitación en redes densas a escala diaria". *El Clima entre el mar y la montaña*. Asociación Española de Climatología: Santander; 431–440.
- Cuadrat, J.M., Pita, M.F. (2000): *Climatología*. Ed. Cátedra.
- El-Kenawy, A. (2012): "Spatio-temporal variability of surface air temperature in Northeastern Spain". Tesis doctoral.
- El-Kenawy, A., López-Moreno, J.I., Stepanek, P., Vicente-Serrano, S. (2013): "An assessment of the role of homogenization protocol in the performance of daily temperature series and trends: application to northeastern Spain". *International Journal of Climatology*. 33: 87–108
- Esteban, P., Ninyerola, M., Prohom, M. (2009): "Spatial modelling of air temperature and precipitation for Andorra (Pyrenees) from daily circulation patterns". *Theoretical Applied Climatology* 96:43–56
- Feng, S., Hu, Q., Qian, W. (2004): "Quality control of daily meteorological data in China, 1951-2000: A new dataset". *International Journal of Climatology*,24(7),pp. 853-870

- García Barrón, L., Pita, M.F. (2001): “Propuesta metodológica para la determinación de inhomogeneidades relativas en las series de observaciones”, en *El tiempo del clima*, Publicaciones de la Asociación Española de Climatología, 2001, Serie A, nº 2, pp. 87-95
- González-Hidalgo, J.C., Brunetti, M. & de Luis, M. (2011), "A new tool for monthly precipitation analysis in Spain: MOPREDAS database (monthly precipitation trends December 1945-November 2005)", *International Journal of Climatology*, vol. 31, no. 5, pp. 715-731.
- Goovaerts, P. (1997) *Geostatistics for natural resources evaluation*. Oxford University Press, New York
- Guijarro, J.A. (2008): “Homogeneización de las series termopluviométricas de Baleares y su influencia en la evaluación de las tendencias”. En SIGRÓ J. et al. (Eds.), *Cambio climático regional y sus impactos*, Asociación Española de Climatología, A-6:151-160
- Hengl, T. (2009): *A practical guide to geostatistical mapping*. Office for Official Publications of the European Communities, Luxembourg.
- Herrera, S. (2011): “Desarrollo, validación y aplicaciones de Spain02: Una rejilla de alta resolución de observaciones interpoladas para precipitación y temperatura en España”. Tesis doctoral. <http://www.meteo.unican.es/en/tesis/herrera>
- Herrera, S., Gutiérrez, J.M., Ancell, R., Pons, M.R., Frías, M.D., Fernández, J. (2012): “Development and Analysis of a 50 year high-resolution daily gridded precipitation dataset over Spain (Spain02)”. *International Journal of Climatology* 32: 74–85.
- Hewitson, B. C. and Crane, R. (2005). “Gridded area-averaged daily precipitation via conditional interpolation”. *J. Climate*, 18:41–57.
- HOME (2011): Homepage of the COST Action ES0601 – “Advances in Homogenisation Methods of Climate Series: an Integrated Approach (HOME)”, available at: <http://www.homogenisation.org>, último acceso: 12 February 2013
- Hofstra, N., Haylock, M., New, M., Jones, P., and Frei, C. (2008). “The comparison of six methods for the interpolation of daily european climate data”. *J. Geophys. Res.*, 113:D21110.
- Isaaks E.H., Strivastava R.M. (1989) *An introduction to applied geostatistics*. Oxford University Press, Oxford
- Joly, D., Nilsen, L., Fury, R., Elvebakk, A., Brossard, T (2003): “Temperature interpolation at a large scale: test on a small area in Svalbard”, *International Journal of Climatology*. 23: 1637–1654.

- Martínez del-Castillo, E., Serrano-Notivoli, R., Novak, K., Longares-Aladrén, L.A., Arrechea, E., de-Luis, M., Saz-Sánchez, M.A. (2012): “Cuantificación de los gradientes climáticos altitudinales en la vertiente Sur del Macizo del Moncayo a partir de una nueva red de estaciones automáticas en altura”. *VIII Congreso AEC: Cambio Climático. Extremos e Impactos*. Salamanca
- Mestre, O., Gruber, C., Prieur, C., Caussinus, H., Jourdain, S. (2011): “SPLIDHOM: A Method for Homogenization of Daily Temperature Observations”. *J. Appl. Meteor. Climatol.*, 50, 2343–2358.
- Mitchell, T.; Jones, P. (2005): “An improved method of constructing a database of monthly climate observations and associated high-resolution grids”. *International Journal of Climatology* 25: 693–712.
- Pita, M.F. (2007): “Horizontes y retos de la ciencia climática”. En Cuadrat, J.M., Martín Vide, J.: *La Climatología española. Pasado, presente y futuro*. Prensas Universitarias de Zaragoza. 574 pp. 2007.
- R Development Core Team (2000): *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos Versión 1.0.1 (2000-05-16)*. <http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf>.
- Saz-Sánchez, M.A., Serrano-Notivoli, R., de-Luis, M., Longares-Aladrén, L.A. (2010): “Comparación de métodos de interpolación y de regresión para la cartografía de temperaturas máximas y mínimas absolutas: el caso de navarra (Norte de España) en 2009”. *VII Congreso AEC: Clima, ciudad y ecosistemas*, Madrid.
- Schmunk, R.: *Panoply Software*. NASA Goddard Institute for Space Studies. <http://www.giss.nasa.gov/tools/panoply>.
- Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.) (2007): *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Stepanek, P. (2003): “Homogeneización de las series de temperatura del aire en la República Checa durante el periodo instrumental”. *Geographicalia*, nº43, pp. 5-24.
- Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquafredda, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and



- Brandsma, T. (2012): “Benchmarking homogenization algorithms for monthly data”, *Clim. Past*, 8, 89-115
- Vincent, L. A., X. Zhang, B.R. Bonsal, and W.D. Hogg (2002): “Homogenization of daily temperatures over Canada”. *J. Climate*, 15, 1322-1334.
- Vicente, S.M., Saz, M.A. y Cuadrat, J.M. (2003). “Comparative analysis of interpolation methods in the middle Ebro valley (Spain): application to annual precipitation and temperature”. *Climate Research* 24, pp 161-180.
- Willmott, C. J., and K. Matsuura (2006), “On the use of dimensioned measures of error to evaluate the performance of spatial interpolators”, *Int. J. Geogr. Inf. Sci.*, 20, 89–102.

# **ANEXO I**

## **ESTACIONES UTILIZADAS**

**Tabla 4.** Observatorios utilizados. Identificador, nombre, altitud (*metros*), longitud, y latitud.

INDICATIVO	NOMBRE	ALTITUD	LONGITUD	LATITUD
2043	NOLAY	1070	554069	4597477
9284	CERVERA DEL RIO ALHAMA	543	587024	4651010
9290	FITERO	438	594537	4656998
9292	CORELLA	370	600930	4663349
9299	TARAZONA	480	606455	4639552
9301	MONTEAGUDO	410	608519	4646247
9302	TUDELA AZUCARERA	263	615930	4657717
9308	VERUELA	650	608516	4629804
9313	BISIMBRE	320	630060	4634391
9315	GALLUR	254	639709	4636425
9348	MONTEAGUDO DE LAS VICARIAS	783	569422	4579628
9350	ARIZA	765	579517	4574427
9371	TERRER AZUCARERA	570	607751	4575839
9395	CALATAYUD	534	614335	4578593
9402	ANIÑON	729	608591	4589610
9425	ALPARTIR	510	635133	4586509
9432	EPILA (MONTARRON)	336	641730	4607677
2296A	ÓLVEGA-CAMINO VEGAFRÍA	1039	585048	4625363
9287A	SAN PEDRO MANRIQUE	1096	563674	4652992
9287F	CORNAGO	704	574853	4657564
9299B	TARAZONA (AYUNTAMIENTO)	475	605598	4639848
9299C	TARAZONA (COMARCAL)	475	605967	4641334
9299X	TARAZONA	475	605873	4641096
9301X	MONTEAGUDO	410	608516	4646041
9302C	TUDELA (AYUNTAMIENTO)	295	615063	4656880
9302X	TUDELA I	295	615063	4656880
9302Y	TUDELA II	295	615063	4656880
9304E	CABANILLAS	257	621734	4654421
9307A	SAN MARTIN DEL MONCAYO (D.G.A.)	816	600216	4632520
9309A	BULBUENTE	530	615703	4630657
9310I	AINZON (COMARCAL)	455	622987	4629853
9311B	BORJA 3	448	621396	4632449
9311C	BORJA (AYUNTAMIENTO)	440	622021	4632336
9314E	NOVILLAS (D.G.A.)	247	632838	4643729
9314I	SANTA ENGRACIA	290	638202	4645127
9315E	MALLEN (D.G.A.)	295	631811	4639483
9336L	FUENDEJALON (D.G.A.)	455	627162	4624836
9337A	PRADILLA DE EBRO	228	644107	4635648

9350A	ARIZA P F E	700	579125	4574114
9352A	ALMAZUL	1000	571673	4602917
9354E	ALHAMA DE ARAGON (DGA)	703	592616	4572084
9354X	ALHAMA DE ARAGON	703	592497	4571879
9367B	ATECA (DGA)	608	601413	4576827
9394U	CALATAYUD AGUAS	600	613364	4576400
9394X	CALATAYUD	600	613364	4576400
9395A	CALATAYUD (COMARCAL)	525	614137	4579361
9401A	VILLARROYA DE LA SIERRA (D.G.A)	740	601140	4591321
9404H	SAVIÑAN (DGA)	445	619537	4588733
9407A	ARANDA DE MONCAYO (D.G.A.)	860	600571	4603868
9407E	MAIDEVERA (EMBALSE)	742	603290	4603259
9408O	ILLUECA (D.G.A.)	581	613978	4599718
9416B	RICLA COOP SAN PEDRO	377	633549	4595581
9425C	LA ALMUNIA (LA REDONDA)	398	636634	4590671
9427B	LA ALMUNIA EITA (COMARCAL)	370	635002	4593510
9427X	LA ALMUNIA DE DOÑA GODINA	370	635182	4593433
9428E	CALATORAO COOPERATIVA	360	640147	4596230
9431E	EPILA (ONDIVIELA)	330	642236	4603152
9432A	EPILA (COLEGIO)	382	643066	4606871
9432B	EPILA (D.G.A.)	320	643191	4606441
AGRAMONT E	Agramonte	1090	597896	4630342
CALCENA	Calcena	811	606791	4612380
MAJADA	Majada	1798	600210	4622801
P014	Fuentes de Ágreda	1092.5	589855	4628925
P018	Beratón	1394.5	599077.6	4619432
P084	Añón de Moncayo	855	605541	4626407
R022	Repetidor de Val	764	599528	4636355

# **ANEXO II**

## **SÍNTESIS DE FUNCIONES PROGRAMADAS EN R**

```

4. #####
5. ## CÁLCULO DE VALORES Z
6. #####
7.
8.
9. ## CARGAR FUNCIONES
10.
11. mmovil <- function(x,ini,end){
12.   m <- rep(NA,length=length(x))
13.   for(i in ini:end){
14.     if(is.na(x[i])) next else{
15.       before <- i-14
16.       after <- i+14
17.       if(before<1) before=1;if(after>length(x)) after=length(x)
18.       m[i] <- mean(x[before:after],na.rm=T)
19.     }
20.   }
21.   return(m)
22. }
23. sdmovil <- function(x,ini,end){
24.   m <- rep(NA,length=length(x))
25.   for(i in ini:end){
26.     if(is.na(x[i])) next else{
27.       before <- i-14
28.       after <- i+14
29.       if(before<1) before=1;if(after>length(x)) after=length(x)
30.       m[i] <- sd(x[before:after],na.rm=T)
31.     }
32.   }
33.   return(m)
34. }
35. estandarizacion <- function(x,xmean,xdesv,sdcan,meancan,ini,end){
36.   m <- rep(NA,length=length(x))
37.   for(i in ini:end){
38.     if(is.na(x[i])) next else{
39.       m[[i]] <- ((x[i]-xmean[i])/xdesv[i])*sdcan[i]+meancan[i]
40.     }
41.   }
42.   return(m)
43. }
44. estandarizacionPred <- function(x,xmean,xdesv,sdcan=scan,meancan=mcan){
45.   m <- rep(NA,length=length(x))
46.   for(i in 1:length(x)){
47.     if(is.na(x[i])) next else{
48.       m[[i]] <- ((x[i]-xmean)/xdesv)*sdcan+meancan
49.     }
50.   }
51.   return(m)
52. }
53. predichosSinNA <-
54.   function(xx,sts=sts,stsInfo=stsInfo,var=var,n=ncol(EstandarMovil),pred,datocand){
55.     if(pred=='alt'){newdata=data.frame(x=as.numeric(as.character(sts$ALTITUD[which(sts$INDICATIVO==colnames(var)[cc]))])); j=1}
56.     if(pred=='lat'){newdata=data.frame(x=as.numeric(as.character(sts$Y[which(sts$INDICATIVO==colnames(var)[cc]))])); j=3}
57.     if(pred=='lon'){newdata=data.frame(x=as.numeric(as.character(sts$X[which(sts$INDICATIVO==colnames(var)[cc]))])); j=2}
58.     m <- numeric(length=8)
59.     if(sum(!is.na(xx))==1){ #si solo hay un vecino, pone ese dato
60.       m[1] <- xx[!is.na(xx)]
61.       m[5] <- 1
62.       m[2:4] <- NA
63.       m[6:8] <- NA
64.     } else if(sum(!is.na(xx))==2){

```

```

64.     m[1] <- mean(xx,na.rm=T)
65.     m[5] <- 1
66.     m[2:4] <- NA
67.     m[6:8] <- NA
68.   } else {
69.     if(sum(is.na(xx))==n) m <- rep(NA,8) else{
70.       w <- which(!is.na(as.numeric(xx)))
71.       y <- xx[w] #variable independiente (temperatura)
72.       if(length(w)>4) orden <- 4 else orden <- length(w)-1
73.       x <- stsInfo[j,w] #variable dependiente
74.       for(degree in 1:orden){ #degree tiene que ser menor que el n?mero de puntos
75.         fm <- lm(y~poly(x,degree,raw=TRUE))
76.         if(is.na(summary(fm)$coefficients[degree+1])){
77.           m[degree] <- NA
78.           m[(degree+4)] <- NA
79.         } else {
80.           m[degree] <- predict(fm, newdata)[1]#predicho
81.           m[(degree+4)] <- summary(fm)$r.squared #R2
82.         }
83.       }
84.     }
85.   }
86.   return(m)
87. }
88.
89.
90. Zetas <- function(cc,var){
91.
92.   ##
93.   ##1. COMIENZO##
94.
95.   pb <- tkProgressBar(min = 0, max = nrow(var),width = 800)
96.   setTkProgressBar(pb, cc, label=paste('Configurando estacion
',colnames(var)[cc],sep=''))
97.   cand <- var[,cc]
98.
99.   vecinos <- var[,-c(cc)]
100.
101.   for(h in 1:length(cand)){if(is.na(cand[h])) next else{ini <- h; break}}
102.   for(h in length(cand):1){if(is.na(cand[h])) next else{end <- h; break}}
103.
104.
105.   #2. Estandarizacion
106.   #2.1.periodo comun de la candidata con los vecinos
107.   ComunCAND <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
108.   colnames(ComunCAND) <- names(vecinos)
109.   ComunVECI <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
110.   colnames(ComunVECI) <- names(vecinos)
111.
112.
113.   for(j in 1:ncol(vecinos)){
114.     suma <- cand+vecinos[,j]
115.     w <- which(!is.na(suma))
116.     ComunCAND[w,j] <- cand[w]
117.     ComunVECI[w,j] <- vecinos[w,j]
118.   }
119.   subCC <- as.numeric(which(colSums(is.na(ComunCAND))==length(fechas)))
120.   ComunCAND <- ComunCAND[,-c(subCC)]
121.   ComunVECI <- ComunVECI[,-c(subCC)]
122.   vecinos <- vecinos[,-c(subCC)]
123.
124.   #media
125.   setTkProgressBar(pb, cc, label=paste('Calculando medias de
',colnames(var)[cc],sep=''))
126.   MMCAND <-sfApply(ComunCAND,2,mmovil,ini,end)
127.   MMVECI <- sfApply(ComunVECI,2,mmovil,ini,end)
128.   #desviacion est?ndar

```

```

129.     setTkProgressBar(pb, cc, label=paste('Calculando desvest de
',colnames(var)[cc],sep=''))
130.     SDCAND <- sfApply(ComunCAND,2,sdmovil,ini,end)
131.     SDVECI <- sfApply(ComunVECI,2,sdmovil,ini,end)
132.
133.     setTkProgressBar(pb, cc, label=paste('Estandarizando vecinos
',colnames(var)[cc],sep=''))
134.     EstandarMovil <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
135.     colnames(EstandarMovil) <- names(vecinos)
136.     for(j in 1:ncol(vecinos)){
137.         EstandarMovil[,j] <- estandarizacion(x=vecinos[,j]
,xmean=MMVECI[,j],xdesv=SDVECI[,j],
138.         sdcan=SDCAND[,j],meancan=MMCAND[,j],ini=ini,end=end)
139.     }
140.     #info de estaciones
141.     stsInfo <- matrix(NA,ncol=ncol(vecinos),nrow=3)
142.     colnames(stsInfo) <- colnames(vecinos)
143.     for(i in 1:ncol(vecinos)){
144.         stsInfo[1,i] <-
as.numeric(as.character(sts$ALTITUD[which(as.character(sts$INDICATIVO)==colnames(stsIn
fo)[i])]))#alturas
145.         stsInfo[2,i] <-
as.numeric(as.character(sts$X[which(as.character(sts$INDICATIVO)==colnames(stsInfo)[i]
)]))#X
146.         stsInfo[3,i] <-
as.numeric(as.character(sts$Y[which(as.character(sts$INDICATIVO)==colnames(stsInfo)[i]
)]))#Y
147.     }
148.
149.     #2.4 PREDICHOS DE CADA VARIABLE
150.     predAlt <- matrix(NA,ncol=8,nrow=nrow(EstandarMovil))
151.     predLat <- matrix(NA,ncol=8,nrow=nrow(EstandarMovil))
152.     predLon <- matrix(NA,ncol=8,nrow=nrow(EstandarMovil))
153.
154.
155.     setTkProgressBar(pb, 7500, label=paste('Calculando predichos de altitud en la
estacion ',colnames(var)[cc],sep=''))
156.     predAlt[ini:end,] <-
t(sfApply(x=EstandarMovil[ini:end,],fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInf
o,
157.         var=var,n=ncol(EstandarMovil),pred='alt'))
158.     setTkProgressBar(pb, 15000, label=paste('Calculando predichos de latitud en la
estacion ',colnames(var)[cc],sep=''))
159.     predLat[ini:end,] <-
t(sfApply(x=EstandarMovil[ini:end,],fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInf
o,
160.         var=var,n=ncol(EstandarMovil),pred='lat'))
161.     setTkProgressBar(pb, 23011, label=paste('Calculando predichos de longitud en la
estacion ',colnames(var)[cc],sep=''))
162.     predLon[ini:end,] <-
t(sfApply(x=EstandarMovil[ini:end,],fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInf
o,
163.         var=var,n=ncol(EstandarMovil),pred='lon'))
164.
165.     #Elegir predicho en funcion de la R2
166.     PredFinal <- matrix(NA,nrow=nrow(vecinos),ncol=7)
167.     colnames(PredFinal) <- c('PredFinal','PredAlt','PredLon','PredLat',
168.         'r1','r2','r3')
169.     for(i in ini:end){
170.         if(sum(is.na(predAlt[i,5:8]))==4) next else{
171.             kk <- max(predAlt[i,5:8],na.rm=T)
172.             wpAlt <- which(max(predAlt[i,5:8],na.rm=T)==kk)
173.             PredFinal[i,2] <- predAlt[i,wpAlt]
174.             PredFinal[i,5] <- kk
175.             kk <- max(predLon[i,5:8],na.rm=T)
176.             wpLon <- which(max(predLon[i,5:8],na.rm=T)==kk)

```



```

177.     PredFinal[i,3] <- predLon[i,wpLon]
178.     PredFinal[i,6] <- kk
179.     kk <- max(predLat[i,5:8],na.rm=T)
180.     wpLat <- which(max(predLat[i,5:8],na.rm=T)==kk)
181.     PredFinal[i,4] <- predLat[i,wpLat]
182.     PredFinal[i,7] <- kk
183.   }
184. }
185. #normalizar cada predicho primero
186. PredNormaliz <- matrix(NA,nrow=nrow(vecinos),ncol=3)
187. scan <- sd(cand,na.rm=T)
188. mcan <- mean(cand,na.rm=T)
189.
190. xm <- mean(PredFinal[,2],na.rm=T)
191. xsd <- sd(PredFinal[,2],na.rm=T)
192. PredNormaliz[,1] <-
  estandarizacionPred(x=PredFinal[,2],xmean=xm,xdesv=xsd,scan=scan,meancan=mcan)
193. xm <- mean(PredFinal[,3],na.rm=T)
194. xsd <- sd(PredFinal[,3],na.rm=T)
195. PredNormaliz[,2] <-
  estandarizacionPred(x=PredFinal[,3],xmean=xm,xdesv=xsd,scan=scan,meancan=mcan)
196. xm <- mean(PredFinal[,4],na.rm=T)
197. xsd <- sd(PredFinal[,4],na.rm=T)
198. PredNormaliz[,3] <-
  estandarizacionPred(x=PredFinal[,4],xmean=xm,xdesv=xsd,scan=scan,meancan=mcan)
199. rm(xsd,xm)
200.
201. #calcular predicho final
202. predRaw <- matrix(NA,ncol=1,nrow=nrow(vecinos))
203. for (i in 1:nrow(EstandarMovil)){
204.   predRaw[i,1] <- ((PredNormaliz[i,1]*PredFinal[i,5])+
205.                   (PredNormaliz[i,2]*PredFinal[i,6])+
206.                   (PredNormaliz[i,3]*PredFinal[i,7]))/
207.   (PredFinal[i,5]+PredFinal[i,6]+
208.    PredFinal[i,7])
209. }
210. #normalizar predicho final
211. xm <- mean(predRaw[,1],na.rm=T)
212. xsd <- sd(predRaw[,1],na.rm=T)
213. PredFinal[,1] <-
  estandarizacionPred(x=predRaw[,1],xmean=xm,xdesv=xsd,scan=scan,meancan=mcan)
214. rm(xm,xsd,predRaw)
215.
216.
217. #Diferencias
218. Diferencias <- PredFinal[,1]-var[,cc]
219. Z <- Diferencias/sd(Diferencias,na.rm=T)
220. close(pb)
221. return(Z)
222. }

```

```

##*****
## FUNCIÓN DE RELLENO
##*****

Relleno <- function(cc,var,ori,ncpu=8){

  if(ncpu>1) sfInit(parallel=TRUE,cpus=ncpu)

  ## CARGAR FUNCIONES

  estandarizacion <- function(x,xmean,xdesv,sdcan,meancan){
    m <- rep(NA,length=length(x))
    for(i in 1:length(x)){
      if(is.na(x[i])) next else{
        m[[i]] <- ((x[i]-xmean)/xdesv)*sdcan+meancan
      }
    }
    return(m)
  }
  estandarizacionPred <- function(x,xmean,xdesv,sdcan=scan,meancan=mcanc){
    m <- rep(NA,length=length(x))
    for(i in 1:length(x)){
      if(is.na(x[i])) next else{
        m[[i]] <- ((x[i]-xmean)/xdesv)*sdcan+meancan
      }
    }
    return(m)
  }
  predichosSinNA <-
function(xx,sts=sts,stsInfo=stsInfo,var=var,n=ncol(EstandarMovil),pred,datocand){

  if(pred=='alt'){newdata=data.frame(x=as.numeric(as.character(sts$ALTITUD[which(sts$INDICATIVO==colnames(var)[cc]])))); j=1}

  if(pred=='lat'){newdata=data.frame(x=as.numeric(as.character(sts$Y[which(sts$INDICATIVO==colnames(var)[cc]])))); j=3}

  if(pred=='lon'){newdata=data.frame(x=as.numeric(as.character(sts$X[which(sts$INDICATIVO==colnames(var)[cc]])))); j=2}
    m <- numeric(length=8)
    if(sum(!is.na(xx))==1){ #si solo hay un vecino, pone ese dato
      m[1] <- xx[!is.na(xx)]
      m[5] <- 1
      m[2:4] <- NA
      m[6:8] <- NA
    } else if(sum(!is.na(xx))==2){
      m[1] <- mean(xx,na.rm=T)
      m[5] <- 1
      m[2:4] <- NA
      m[6:8] <- NA
    } else {
      if(sum(is.na(xx))==n) m <- rep(NA,8) else{
        w <- which(!is.na(as.numeric(xx)))
        y <- xx[w] #variable independiente (temperatura)
        if(length(w)>4) orden <- 4 else orden <- length(w)-1
        x <- stsInfo[j,w] #variable dependiente
        for(degree in 1:orden){ #degree tiene que ser menor que el n?mero de puntos
          fm <- lm(y~poly(x,degree,raw=TRUE))
          if(is.na(summary(fm)$coefficients[degree+1])){
            m[degree] <- NA
            m[(degree+4)] <- NA
          } else {
            m[degree] <- predict(fm, newdata)[1]#predicho
            m[(degree+4)] <- summary(fm)$r.squared #R2
          }
        }
      }
    }
  }
  return(m)
}

```

```

##
##1. COMIENZO##

pb <- tkProgressBar(min = 0, max = nrow(var),width = 800)
setTkProgressBar(pb, cc, label=paste('Configurando estacion ',colnames(var)[cc],sep=''))
cand <- var[,cc]
vecinos <- as.data.frame(var[,-c(cc)])

for(h in 1:length(cand)){if(is.na(cand[h])) next else{ini <- h; break}}
for(h in length(cand):1){if(is.na(cand[h])) next else{end <- h; break}}

#2. Estandarizaci?n
#2.1.periodo com?n de la candidata con los vecinos
ComunCAND <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
colnames(ComunCAND) <- names(vecinos)
ComunVECI <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
colnames(ComunVECI) <- names(vecinos)
ww <- list()
  for(j in 1:ncol(vecinos)){
    if(sum(!is.na(vecinos[ini:end,j]))>29) ww[[j]] <- j else next
  }
if(length(ww)==0) result=cand else { #si no tiene vecinos con mas de 29 dias, pasa a la
siguiente candidata
  subCC <- unlist(ww)
  vecinos <- as.data.frame(vecinos[,subCC])
  names(vecinos) <- colnames(ComunCAND)[subCC]
  ComunCAND <- as.matrix(ComunCAND[,subCC])
  colnames(ComunCAND) <- names(vecinos)
  ComunVECI <- as.matrix(ComunVECI[,subCC])
  colnames(ComunVECI) <- names(vecinos)
  w <- which(!is.na(cand[ini:end]))
  for(j in 1:ncol(vecinos)){
    #suma <- cand+vecinos[,j]
    #w <- which(!is.na(suma))
    ComunCAND[(ini:end)[w],j] <- cand[ini:end][w]
    ComunVECI[(ini:end)[w],j] <- vecinos[ini:end,j][w]
  }
  #media
  setTkProgressBar(pb, cc, label=paste('Calculando medias de
',colnames(var)[cc],sep=''))
  #MMCAND <- sfApply(ComunCAND,2,mean, na.rm=T)
  #MMVECI <- sfApply(ComunVECI,2,mean, na.rm=T)
  MMCAND <- apply(ComunCAND,2,mean, na.rm=T)
  MMVECI <- apply(ComunVECI,2,mean, na.rm=T)
  #desviacion est?ndar
  setTkProgressBar(pb, cc, label=paste('Calculando desvest de
',colnames(var)[cc],sep=''))
  #SDCAND <- sfApply(ComunCAND,2,sd, na.rm=T)
  #SDVECI <- sfApply(ComunVECI,2,sd, na.rm=T)
  SDCAND <- apply(ComunCAND,2,sd, na.rm=T)
  SDVECI <- apply(ComunVECI,2,sd, na.rm=T)

  setTkProgressBar(pb, cc, label=paste('Estandarizando vecinos
',colnames(var)[cc],sep=''))
  EstandarMovil <- matrix(NA,nrow=nrow(vecinos),ncol=ncol(vecinos))
  colnames(EstandarMovil) <- names(vecinos)
  for(j in 1:ncol(vecinos)){
    EstandarMovil[,j] <- estandarizacion(x=vecinos[,j] ,xmean=MMVECI[j],xdesv=SDVECI[j],
sdcan=SDCAND[j],meancan=MMCAND[j])
  }
  #info de estaciones
  stsInfo <- matrix(NA,ncol=ncol(vecinos),nrow=3)
  colnames(stsInfo) <- names(vecinos)
  for(i in 1:ncol(vecinos)){
    stsInfo[1,i] <-
as.numeric(as.character(sts$ALTITUD[which(as.character(sts$INDICATIVO)==colnames(stsInfo)[
i])]))#alturas
    stsInfo[2,i] <-
as.numeric(as.character(sts$X[which(as.character(sts$INDICATIVO)==colnames(stsInfo)[i])]))
#X

```

```

      stsInfo[3,i] <-
as.numeric(as.character(sts$Y[which(as.character(sts$INDICATIVO)==colnames(stsInfo)[i])]))
#Y
    }

#2.4 PREDICHOS DE CADA VARIABLE

nas <- which(is.na(ori[,cc])) #calculamos predichos para los dias en los que no hay
dato en la candidate

sfExport('cc')
setTkProgressBar(pb, 7500, label=paste('Calculando predichos de altitud en la estacion
',colnames(var)[cc],sep=''))
predAlt <-
t(sfApply(x=EstandarMovil,fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInfo,
          var=var,n=ncol(EstandarMovil),pred='alt'))
setTkProgressBar(pb, 15000, label=paste('Calculando predichos de latitud en la estacion
',colnames(var)[cc],sep=''))
predLat <-
t(sfApply(x=EstandarMovil,fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInfo,
          var=var,n=ncol(EstandarMovil),pred='lat'))
setTkProgressBar(pb, 23011, label=paste('Calculando predichos de longitud en la estacion
',colnames(var)[cc],sep=''))
predLon <-
t(sfApply(x=EstandarMovil,fun=predichosSinNA,margin=1,sts=sts,stsInfo=stsInfo,
          var=var,n=ncol(EstandarMovil),pred='lon'))

#Elegir predicho en funcion de la R2
PredFinal <- matrix(NA,nrow=nrow(vecinos),ncol=7)
colnames(PredFinal) <- c('PredFinal','PredAlt','PredLon','PredLat',
                        'r1','r2','r3')

for (i in nas){
  if(sum(is.na(predAlt[i,5:8]))==4) next else{
    kk <- max(predAlt[i,5:8],na.rm=T)
    wpAlt <- which(max(predAlt[i,5:8],na.rm=T)==kk)
    PredFinal[i,2] <- predAlt[i,wpAlt]
    PredFinal[i,5] <- kk
    kk <- max(predLon[i,5:8],na.rm=T)
    wpLon <- which(max(predLon[i,5:8],na.rm=T)==kk)
    PredFinal[i,3] <- predLon[i,wpLon]
    PredFinal[i,6] <- kk
    kk <- max(predLat[i,5:8],na.rm=T)
    wpLat <- which(max(predLat[i,5:8],na.rm=T)==kk)
    PredFinal[i,4] <- predLat[i,wpLat]
    PredFinal[i,7] <- kk
  }
}
rm(predAlt,predLat,predLon)
#normalizar cada predicho primero
PredNormaliz <- matrix(NA,nrow=nrow(vecinos),ncol=3)

scan <- sd(cand,na.rm=T)
mcan <- mean(cand,na.rm=T)

#agilizar esta estandarizacion d?ndole la media y la desvest antes de que empiece
xm <- mean(PredFinal[,2],na.rm=T)
xsd <- sd(PredFinal[,2],na.rm=T)
PredNormaliz[,1] <-
estandarizacionPred(x=PredFinal[,2],xmean=xm,xdesv=xsd,sdcan=scan,meancan=mcan)
xm <- mean(PredFinal[,3],na.rm=T)
xsd <- sd(PredFinal[,3],na.rm=T)
PredNormaliz[,2] <-
estandarizacionPred(x=PredFinal[,3],xmean=xm,xdesv=xsd,sdcan=scan,meancan=mcan)
xm <- mean(PredFinal[,4],na.rm=T)
xsd <- sd(PredFinal[,4],na.rm=T)
PredNormaliz[,3] <-
estandarizacionPred(x=PredFinal[,4],xmean=xm,xdesv=xsd,sdcan=scan,meancan=mcan)
rm(xsd,xm)
#calcular predicho final

```

```

predRaw <- matrix(NA,ncol=1,nrow=nrow(vecinos))
for (i in nas){
  predRaw[i,1] <- ((PredNormaliz[i,1]*PredFinal[i,5])+
                  (PredNormaliz[i,2]*PredFinal[i,6])+
                  (PredNormaliz[i,3]*PredFinal[i,7]))/
                (PredFinal[i,5]+PredFinal[i,6]+
                 PredFinal[i,7])
}
#normalizar predicho final
xm <- mean(predRaw[,1],na.rm=T)
xsd <- sd(predRaw[,1],na.rm=T)
PredFinal[,1] <-
estandarizacionPred(x=predRaw[,1],xmean=xm,xdesv=xsd,sdcan=scan,meancan=mcan)
rm(xm,xsd,predRaw)
#relleno del dato
result <- numeric(length=nrow(PredFinal))

result[nas] <- PredFinal[nas,1]
result[which(!is.na(cand))] <- cand[which(!is.na(cand))]
rm(PredFinal,PredNormaliz)

}
close(pb)
return(result)

sfStop()
}

```

# **ANEXO III**

## **CARTOGRAFÍAS DE EJEMPLO**

(1 a 31 de marzo de 2011)

