Facultad de Ciencias

Estudio de idoneidad de la técnica *Random Forest* para la regionalización estadística de proyecciones de cambio climático

(On the suitability of the Random Forest technique for the statistical downscaling of climate change projections)

Trabajo Fin de Máster
para acceder al
Máster Universitario en Ciencia de Datos

Autor: Miguel Traspuesto Abascal
Director: Dr. Rodrigo García Manzanas
Junio - 2019

Resumen

Los Modelos Globales del Clima (GCM, por sus siglas en inglés) son las herramientas utilizadas hoy en día para la simulación del clima en las diferentes escalas temporales (desde 3-5 días vista hasta final de siglo). Debido a ciertas limitaciones físicas y a su alto coste computacional, la resolución espacial de los GCM actuales todavía es insuficiente. Para ayudar a solventar esta limitación se ha desarrollado en las últimas décadas una extensa batería de técnicas de regionalización (o downscaling).

En el marco de la iniciativa europea VALUE (http://www.value-cost.eu/), cuyo objetivo es el de comparar diferentes estrategias de regionalizacion para el estudio del cambio climático (Gutiérrez et al., 2018), han presentado recientemente la intercomparación de métodos de regionalizacion estadística más extensa (más de 50 técnicas) y rigurosa hasta la fecha sobre 86 estaciones repartidas por Europa.

En este Trabajo de Fin de Máster se perfila a random forest como otra opción válida a las técnicas presentes en Gutiérrez et al. (2018). Además se muestra como random forest también es otra opción viable para la regionalización estadística de proyecciones de cambio climático sin necesidad de una fase de selección de variables, obteniendo resultados más realistas que otras técnicas.

Palabras clave: *Regionalización estadística, Random Forest, Proyecciones climáticas.*

1

Abstract

Global Climate Models (GCMs) are the tools used today to simulate climate at different time scales (from 3-5 days seen until the end of the century). Due to certain physical limitations and their high computational cost, the spatial resolution of the current GCMs is still insufficient. In order too solve this limitation, an extensive battery of downscaling techniques has been developed in recent decades.

In the European initiative VALUE (http://www.value-cost.eu/), whose objective is to compare different downscaling strategies for the study of climate change (Gutiérrez et al., 2018), recently presented the intercomparison of methods of statistical downscaling most extensive (more than 50 techniques) and rigorous to date over 86 stations spread across Europe.

In this Master's Thesis, random forest is outlined as another valid option to the techniques present in Gutiérrez et al. (2018). It is also shown that random forest is also another viable option for the statistical downscaling of climate change projections without the need for a phase of selection of variables, obtaining more realistic results than other techniques.

Keywords: *Statistical Downscaling, Random Forest, Climate change Projections.*

1

Índice general

CAPÍTULO 1

Introducción

1.1. Motivación

Los modelos climáticos son modelos matemáticos que resuelven las ecuaciones que gobiernan la dinámica del sistema climático y los intercambios de materia, calor y momento que tienen lugar entre sus distintos componentes (atmósfera, hidrosfera, criosfera, litosfera y biosfera). Estas ecuaciones forman un complejo sistema no-lineal sin solución analítica; por tanto, se resuelven numéricamente, lo que requiere discretizar el espacio en celdas tridimensionales y el tiempo en "pasos". La resolución espacial de estos modelos está determinada por el horizonte de predicción que se pretenda alcanzar (predicción a corto plazo, estacional, proyecciones climáticas, etc.) y la potencia de cómputo disponible. Por ejemplo, en el caso de simulaciones para la generación de escenarios de cambio climático, los modelos se integran en períodos de tiempo muy largos (del orden de siglos) bajo diversas hipótesis de forzamiento radiativo (duplicación de emisiones de CO_2, por ejemplo). En estos casos, la resolución horizontal es típicamente entre 1^o y 2.5^o ($\sim 100 - 250$ km), claramente insuficiente para el estudio de impactos a nivel local (hidrología, agricultura, salud, ecosistemas, etc.).

Para solventar esta limitación, se han desarrollado durante las últimas décadas técnicas que permitan proyectar las simulaciones de los modelos climáticos a la escala local requerida, proceso conocido como *downscaling* (regionalización) en el contexto de las Ciencias Atmosféricas. Existen dos metodologías principales a tal fin, una

basada en técnicas dinámicas (acoplar modelos de área limitada y mayor resolución a las salidas de modelos de baja resolución) y otra basada en técnicas estadísticas. Estas últimas engloban tres familias diferentes, los métodos de *corrección del sego*, los *generadores del tiempo* y los métodos de *perfect pronosis*. En este Trabajo Fin de Máster (TFM de aquí en adelante) se consideran únicamente estos últimos.

En el marco de la iniciativa europea VALUE (`http://www.value-cost.eu/`), cuyo objetivo es el de comparar diferentes métodos de regionalización *perfect prognosis* (o simplemente *perfect prog*) para el estudio del cambio climático, Gutiérrez et al. (2018) han presentado recientemente la intercomparación de métodos más extensa y rigurosa hasta la fecha. En concreto, en este trabajo se analizan más de 50 técnicas (que cubren las tres grandes familias anterioramente mencionadas) sobre 86 estaciones de temperatura y precipitación repartidas por toda Europa. En este mismo estudio se manifiesta que *"[...] un trabajo interesante podría ser incluir métodos de machine learning no lineales."* El primer objetivo de este TFM es, por tanto, el de extender este estudio probando una técnica de machine learning; random forests. Por tratarse de la variable más problemática (de entre las más demandadas por los usuarios; temperatura y precipitación), aquí nos centraremos únicamente en la precipitación. Nótese que, además de su carácter mixto (ocurrencia/cantidad), esta variable presenta una alta variabilidad espacial, por lo que su regionalización supone un reto mucho mayor que para el caso de la temperatura (véase, por ejemplo, Schmidli et al., 2007; Bundel et al., 2011; Murphy, 1999; Schoof and Pryor, 2001; Schmidli et al., 2007).

Más allá de algunos estudios focalizados en zonas muy concretas para el downscaling de temperatura (Pang et al., 2017; Wu and Li, 2019; Hamed et al., 2018) y precipitación (He et al., 2016; Shi and Song, 2015), el uso de la técnica random forest en el contexto de la meteorología ha sido muy limitado hasta la fecha.

1.2. Objetivos

- Como ya se ha dicho anteriormente, el primer gran objetivo de este TFM es el de extender el trabajo presentado en Gutiérrez et al. (2018) añadiendo una técnica nueva de machine learning, random forest. En particular, siguiendo el mismo marco experimental que en el mencionado trabajo, se comparará el rendimiento en condiciones perfectas de random forest con el de otra técnica de regionalización mucho más clásica, los Modelos Lineales Generalizados (GLM; ver Sección 3.4) —que también se utilizan en Gutiérrez et al. (2018)— sobre 86 estaciones de precipitación repartidas por toda Europa. Este estudio se aborda

a lo largo de la Sección 4.1.

- Una vez completado el punto anterior, el segundo gran objetivo de este TFM será el de analizar la idoneidad de random forest para la regionalización de proyecciones climáticas. Para ello se considerará el marco experimental presentado en Manzanas et al. (2019) para un conjunto de 41 estaciones de precipitación en Malawi, y se estudiará si random forest puede ser de utilidad para solucionar los problemas de extrapolación encontrados para el caso de los GLM en el citado trabajo, que pueden dar lugar a proyecciones futuras muy poco plausibles. Este análisis se presenta en la Sección 4.2.

1.3. Estructura

En los Capítulos 2 y 3 se hace una descripción detallada de los datos (observaciones, reanálisis y GCMs) y métodos (árboles de clasificación y regresión, random forest, modelos lineales generalizados) utilizados en este TFM, respectivamente.

Los resultados obtenidos se exponen a lo largo del Capítulo 4, que comprende dos secciones principales (Secciones 4.1 y 4.2). Cada una de ellas está ligada a uno de los dos grandes objetivos que se han propuestos (sección anterior).

Por último, en el Capítulo 5 se resumen las principales conclusiones obtenidas de la elaboración de este TFM, y se enuncian algunas líneas de trabajo futuro.

CAPÍTULO 2

Datos

2.1. Observaciones

Para el estudio de idoneidad en condiciones perfectas (primer objetivo del TFM; Sección 4.1) se utilizan las observaciones diarias de precipitación para el período 1979-2008 sobre las 86 estaciones que se consideraron en Gutiérrez et al. (2018), las cuales se encuentran uniformemente distribuidas por toda Europa, cubriendo adecuadamente los diferentes climas de la región (ver Tabla 2.1). Como se puede ver en la Figura 2.1 se han divido las 86 estaciones en las 8 regiones de PRUDENCE (Christensen and Christensen, 2007): Islas Británicas, Iberia, Francia, Europa central, Escandinavia, Alpes, Mediterráneo y Europa del este. Cabe destacar que los datos de estas estaciones provienen de ECA&D (Klein Tank et al., 2002) y están públicamente accesibles en formato *csv* a través de `http://www.value-cost.eu/data`.

Tabla 2.1: Reproducción de la Tabla 1 en Gutiérrez et al. (2018), en la que se indica, para las 86 estaciones consideradas (ordenadas por latitud): identificador ECA&D, nombre, longitud (°), latitud (°), elevación (m), país y tipo de clima de Köppen–Geiger.

#	ID	Nombre	Lon.	Lat.	Elev.	País	Köppen
1	231	Málaga	-4.49	36.67	7	España	Csa
2	63	Methoni	21.70	36.83	51	Grecia	Csa
3	214	Lisboa-Geofisica	-9.15	38.72	77	Portugal	Csa
4	229	Badajoz/Talavera-La-Real	-6.83	38.88	185	España	Csa
5	175	Cagliari	9.05	39.23	21	Italia	Csa
6	3919	Palma-De-Mallorca	2.74	39.56	8	España	BSk

Tabla 2.1: Reproducción de la Tabla 1 en Gutiérrez et al. (2018), en la que se indica, para las 86 estaciones consideradas (ordenadas por latitud): identificador ECA&D, nombre, longitud (°), latitud (°), elevación (m), país y tipo de clima de Köppen–Geiger.

#	ID	Nombre	Lon.	Lat.	Elev.	País	Köppen
7	59	Corfu	19.92	39.62	11	Grecia	Csa
8	62	Larissa	22.45	39.65	72	Grecia	BSk
9	3946	Madrid-Barajas	-3.56	40.47	609	España	BSk
10	232	Navacerrada	-4.01	40.78	1894	España	Csb
11	236	Tortosa-Observatorio-Ebro	0.49	40.82	44	España	Csa
12	176	Roma-Ciampino	12.58	41.78	105	Italia	Csa
13	212	Braganca	-6.73	41.80	690	Portugal	Csb
14	1394	Santiago-De-Compostela	-8.41	42.89	370	España	Cfb
15	1686	Hvar	16.45	43.17	20	Croacia	Csa
16	234	San-Sebastián-Igueldo	-2.04	43.31	251	España	Cfb
17	39	Marseille-Marignane	5.23	43.44	5	Francia	Csa
18	800	Toulouse-Blagnac	1.38	43.62	151	Francia	Cfa
19	355	Mont-Aigoual	3.58	44.12	1567	Francia	Cfb
20	2062	Constanta	28.63	44.22	13	Rumanía	Cfa
21	219	Bucuresti-Baneasa	26.08	44.52	90	Rumanía	Cfa
22	1684	Gospic	15.37	44.55	564	Croacia	Cfb
23	1687	Zavizan	14.98	44.82	1594	Croacia	Dfc
24	177	Verona-Villafranca	10.87	45.38	68	Italia	Cfa
25	173	Milan	9.19	45.47	150	Italia	Cfa
26	450	Sibiu	24.15	45.80	444	Rumanía	Cfb
27	21	Zagreb-Gric	15.98	45.82	156	Croacia	Cfa
28	242	Lugano	8.97	46.00	300	Suiza	Cfa
29	217	Arad	21.35	46.13	116	Rumanía	Cfb
30	1662	Sion-2	7.33	46.22	482	Suiza	Cfb
31	15	Sonnblick	12.95	47.05	3106	Austria	ET
32	32	Bourges	2.37	47.07	161	Francia	Cfb
33	12	Graz	15.45	47.08	366	Austria	Cfb
34	951	Iasi	27.63	47.17	102	Rumanía	Cfa
35	243	Saentis	9.35	47.25	2502	Suiza	ET
36	13	Innsbruck	11.40	47.27	577	Austria	Cfb
37	244	Zueriswitzerland	8.57	47.38	556	Suiza	Cfb
38	4002	Oberstdorf	10.28	47.40	806	Alemania	Cfb
39	58	Zugspitze	10.99	47.42	2964	Alemania	ET
40	239	Basel-Binningen	7.58	47.55	316	Suiza	Cfb
41	14	Salzburg	13.00	47.80	437	Austria	Cfb
42	48	Hohenpeissenberg	11.01	47.80	977	Alemania	Cfb
43	322	Rennes	-1.73	48.07	36	Francia	Cfb
44	16	Wien	16.35	48.23	198	Austria	Cfb
45	38	Paris-14e	2.34	48.82	75	Francia	Cfb
46	2762	Rheinstetten	8.33	48.97	116	Alemania	Cfb
47	4004	Regensburg	12.10	49.04	365	Alemania	Cfb
48	3991	Giessen-Wettenberg	8.65	50.60	203	Alemania	Cfb
49	17	Uccle	4.37	50.80	100	Bélgica	Cfb
50	483	Dresden-Klotzsswitzerlande	13.76	51.13	227	Alemania	Cfb
51	274	Oxford	-1.27	51.77	63	Reino Unido	Cfb
52	2006	Brocken	10.62	51.80	1142	Alemania	Dfc
53	333	Siedlce	22.25	52.25	152	Polonia	Dfb
54	54	Potsdam	13.06	52.38	81	Alemania	Cfb
55	42	Bremen	8.80	53.05	4	Alemania	Cfb

Tabla 2.1: Reproducción de la Tabla 1 en Gutiérrez et al. (2018), en la que se indica, para las 86 estaciones consideradas (ordenadas por latitud): identificador ECA&D, nombre, longitud (°), latitud (°), elevación (m), país y tipo de clima de Köppen–Geiger.

#	ID	Nombre	Lon.	Lat.	Elev.	País	Köppen
56	351	Waddington	0.52	53.17	68	Reino Unido	Cfb
57	350	Valley	-4.53	53.25	11	Reino Unido	Cfb
58	468	Helgoland	7.89	54.18	4	Alemania	Cfb
59	1020	Lazdijai	23.52	54.23	133	Lituania	Dfb
60	3994	Arkona	13.44	54.68	42	Alemania	Cfb
61	332	Leba	17.53	54.75	2	Polonia	Dfb
62	200	Kaunas	23.83	54.88	77	Lituania	Dfb
63	272	Eskdalemuir	-3.20	55.32	242	Reino Unido	Cfb
64	201	Klaipeda	21.07	55.73	6	Lituania	Cfb
65	113	Tranebjerg	10.60	55.85	11	Dinamarca	Cfb
66	1009	Birzai	24.77	56.20	60	Lituania	Dfb
67	107	Vestervig	8.32	56.77	18	Dinamarca	Cfb
68	465	Visby	18.33	57.67	42	Suecia	Cfb
69	462	Goteborg	11.99	57.72	5	Suecia	Cfb
70	349	Stornoway	-6.32	58.33	9	Reino Unido	Cfb
71	275	Wick	-3.08	58.45	36	Reino Unido	Cfb
72	192	Faerder	10.53	59.03	6	Noruega	Cfb
73	194	Utsira-Fyr	4.88	59.31	55	Noruega	Cfb
74	28	Helsinki-Kaisaniemi	24.95	60.18	4	Finlandia	Dfb
75	708	Jokioinen-Jokioisten	23.50	60.81	104	Finlandia	Dfb
76	5585	Salen	13.26	61.17	360	Suecia	Dfc
77	191	Kjoeremsgrende	9.05	62.10	626	Noruega	Dfc
78	330	Fokstua	9.28	62.12	952	Noruega	Dfc
79	1051	Tafjord	7.42	62.23	15	Noruega	Csb
80	29	Jyvaskyla-Lentoasema	25.68	62.40	139	Finlandia	Dfc
81	7682	Siikajoki-Revonlahti	25.09	64.68	48	Finlandia	Dfc
82	339	Haparanda	24.14	65.83	5	Suecia	Dfc
83	1427	Jackvik	17.00	66.38	430	Suecia	Dfc
84	30	Sodankyla-Lapin-Ilmatiet	26.63	67.37	179	Finlandia	Dfc
85	190	Karasjok	25.50	69.47	129	Noruega	Dfc
86	195	Vardoe	31.08	70.37	14	Noruega	ET

Para el estudio de idoneidad de random forest en condiciones de cambio climático (segundo gran objetivo de este TFM; Sección 4.2), se utiliza también la precipitación diaria para el período 1971-2012 en 41 estaciones uniformemente distribuídas sobre Malawi (Figura 2.2). Estos datos son los mismos que se utilizan en Manzanas et al. (2019), y han sido proporcionados por el Departamento de Cambio Climático y Servicios Meteorólogicos (DCCMS).

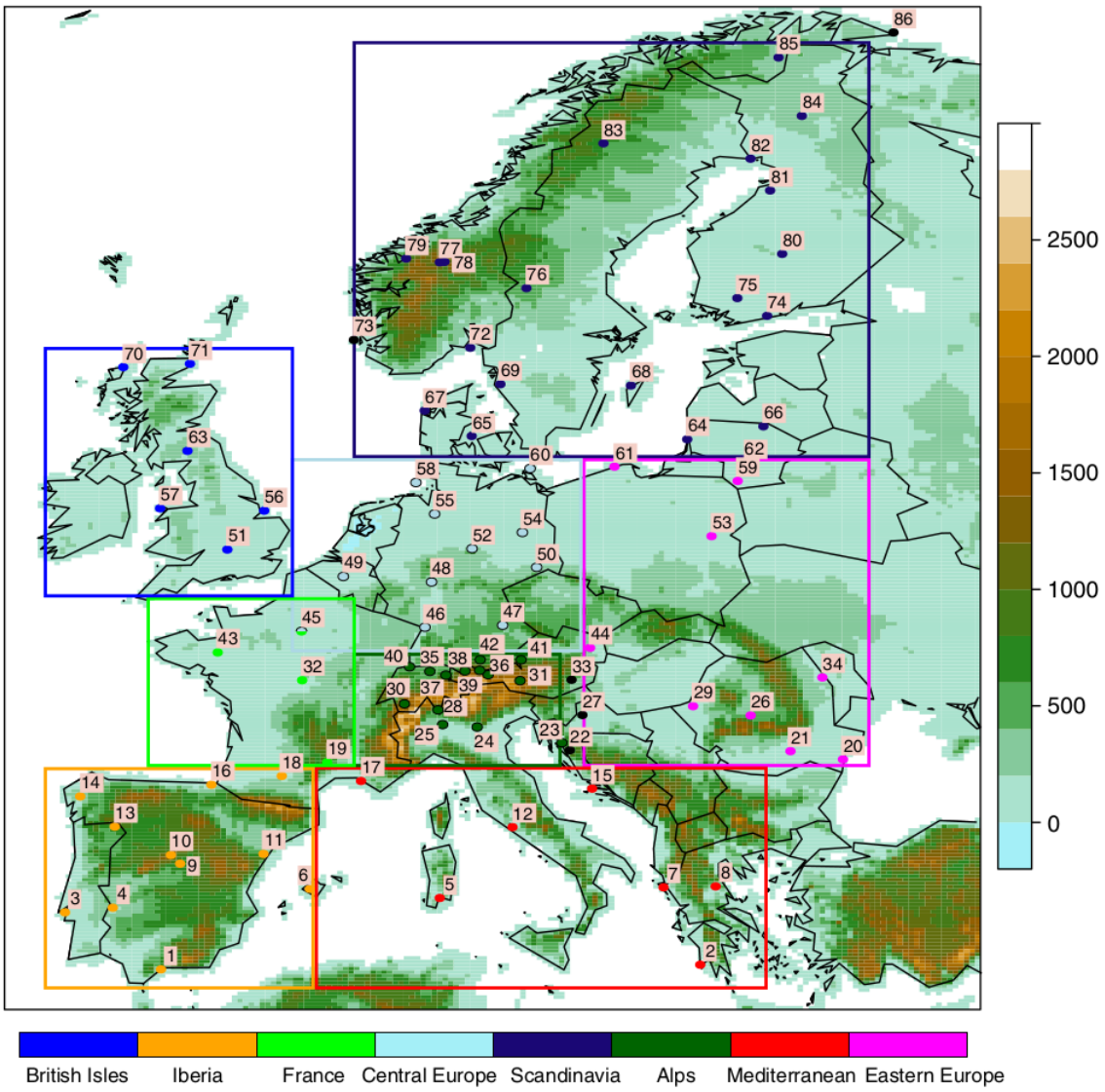

Figura 2.1: Figura 1 de Gutiérrez et al. (2018), mostrando la distribución espacial de las 86 estaciones utilizadas en este TFM, etiquetadas de acuerdo a su latitud (ver Tabla 2.1). La barra de colores a la derecha se refiere a la orografía del terreno (en metros). Las diferentes regiones PRUDENCE se indican con rectángulos de diferente color (barra de colores en la parte de abajo).

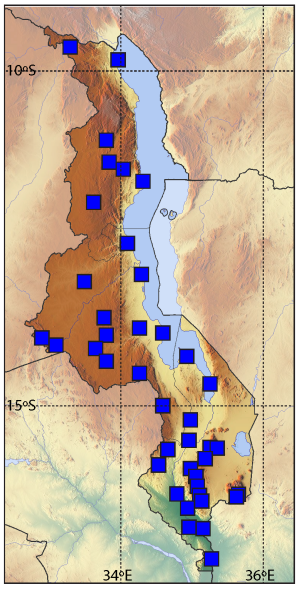

Figura 2.2: Distribución espacial de las 41 estaciones en Malawi consideradas en este TFM (adaptado de la Figura 1 en Manzanas et al. (2019)).

2.2. Reanálisis

Los reanálisis combinan observaciones y modelos numéricos del clima con el objetivo de generar una estimación sintética lo más cercana posible al estado "real" de la atmósfera en cada momento. Típicamente, los reanálisis se integran (en modo retrospectivo) a nivel global, proporcionando datos tridimensionales (desde la superficie terrestre hasta más allá de la estratosfera) para las últimas décadas. Estos registros son usados en múltiples investigaciones y servicios climáticos.

Al igual que en Gutiérrez et al. (2018) y Manzanas et al. (2019), en este TFM se utiliza el reanálisis europeo de referencia, ERA-Interim (Dee et al., 2011), que proporciona datos para un conjunto enorme de variables atmosféricas, tanto en la superficie como en distintos niveles de presión (niveles verticales). Este dataset puede descargarse libremente a través del servidor de datos del Grupo de Meteorología de Santander (UDG-TAP: `http://www.meteo.unican.es/udg-tap/home`) mediante el paquete *loadeR* de *climate4r* (Iturbide et al., 2019). Los datos se encuentran en una rejilla regular de 2º de resolución espacial, la cual compatible con la resolución de los GCMs usados para la generación de escenarios de cambio climático (ver siguiente sección). La Tabla 2.2 indican las variables, niveles de presión, unidades y agregación temporal del conjunto de predictores típico para tareas de regionalización estadística (ver Tabla 2 de Gutiérrez et al. (2018)). Los colores se asignan en función de cómo se han usado en este TFM (ver Capítulo 4).

Variable	Código	Niveles	Unidades	Agregación temporal
Precipitación	PRC	-	mm	Acumulación diaria
Presión a nivel del mar	PSL	-	Pa	Media diaria
Temperatura 2m	2T	2m	K	Media diaria
Geopotencial	Z	200 250 500 700 850 1000 mb	m2 s-2	Media diaria
Temperatura	T	200 250 500 700 850 1000 mb	K	Media diaria
Componente zonal del viento	U	200 250 500 700 850 1000 mb	m s-1	Media diaria
Componente meridional del viento	V	200 250 500 700 850 1000 mb	m s-1	Media diaria
Humedad específica	Q	200 250 500 700 850 1000 mb	g kg-1	Media diaria
Humedad relativa	R	200 250 500 700 850 1000 mb	%	Media diaria

Tabla 2.2: Descipción de las variables, niveles de presión, unidades y agregación temporal del conjunto de predictores típicos (ver Tabla 2 de Gutiérrez et al. (2018)). Los colores se asignan en función de cómo se han usado en este TFM (ver Capítulo 4).

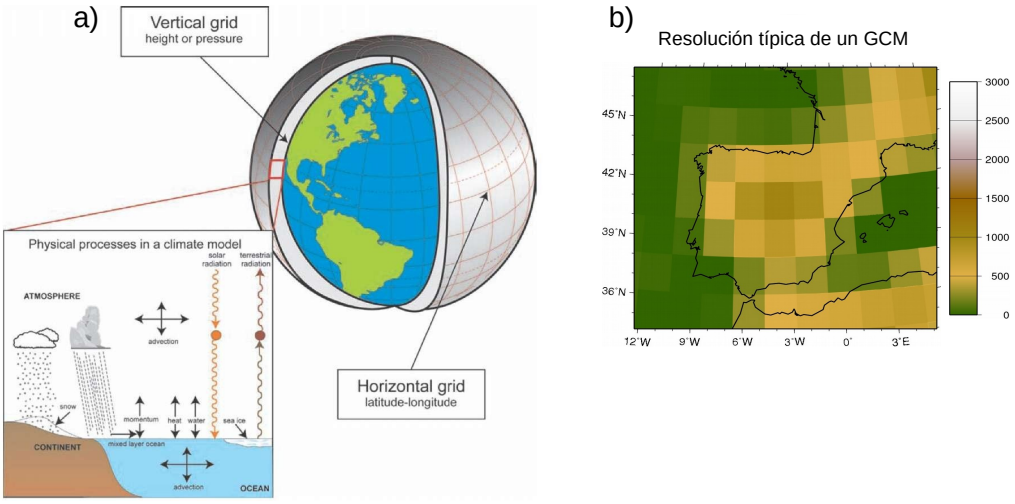

Figura 2.3: a) Rejilla tridimensional típica de un GCM sobre la Tierra. b) Ejemplo de la orografía que ve un GCM sobre la Península Ibérica.

2.3. Modelos Globales del Clima (GCMs)

Los Modelos Globales del Clima (GCMs) son herramientas que describen los procesos e interacciones (intercambios de calor, flujo y energía) más importantes que tienen lugar entre los diferentes componentes del sistema climático (atmósfera, océano, criosfera, litosfera, etc.). Para ello tienen que resolver un sistema de ecuaciones altamente no lineal (sin solución analítica) por medio de métodos numéricos, lo que requiere discretizar el espacio (en celdas tridimensionales) y el tiempo (en "pasos"). Como consecuencia de esta discretización, la resolución espacial de estos modelos no suele superar, para el caso de simulaciones de cambio climático (que se integran por cientos de años), los 200 km (ver Figura 2.3).

La evolución más reciente de los GCM incorpora el modelado específico de procesos que hasta la fecha no se tenían en cuenta, como la interacción entre la atmásfera y la vegetación, y el ciclo del carbono (Taylor et al., 2012). Para diferenciarlos de los GCM anteriores, a estos nuevos GCM se les suele denominar *Earth System Models (ESMs)*. Sin embargo, al igual que en gran parte de la literatura actula, en este TFM utilizaremos los términos GCM y ESM indistintamente.

Para simular el clima futuro, estos ESM se integran bajo distintas trayectorias de forzamiento radiativo (o *Representative Concentration Pathways (RCPs)*), que representan una serie de escenarios futuros plausible que han sido definidos por la comunidad ciéntifica en base a ciertos parámetros (emisiones de gases de efecto invernadero, evolución de la demografía, del uso de energías verdes, etc.) —ver Figura 2.4.— Los RCP utilizados en el quinto informe del IPCC (*Intergovernmental Panel on Climate Change*, `https://www.ipcc.ch/report/ar5/wg1/`)) van desde un forzamiento radiativo de 2.6 W/m^2 para el final de siglo (el más benévolo de todos) hasta otro de 8.5 W/m^2 —se puede ver una breve comparativa en la Tabla 2.3.— Para el estudio de regionalización en condiciones de cambio climático de la Sección 4.2, se utilizará el RCP85 (Taylor, 2001) por ser el que más se aproxima a las condiciones observadas en la actualidad (estos RCP fueron definidos en el año 2000). En la Tabla 2.4 se indican los 4 ESMs del CMIP5 (5th phase of the Coupled Model Intercomparison Project: `https://esgf-node.llnl.gov/projects/cmip5/`) que se consideran en este TFM (de entre los utilizados en Manzanas et al. (2019)).

Al igual que para ERA-Interim, el acceso a estos datos se ha hecho a través del UDG-TAP mediante el paquete *loadeR* de *climate4r*. Hay que decir que cada ESM tiene una resolución horizontal nativa distinta, por lo que, para obtener un marco de trabajo general común, se ha interpolado cada uno de ellas a la misma rejilla regular de 2º (mediante *interpolación bilineal*) utilizada para el reanálisis ERA-Interim.

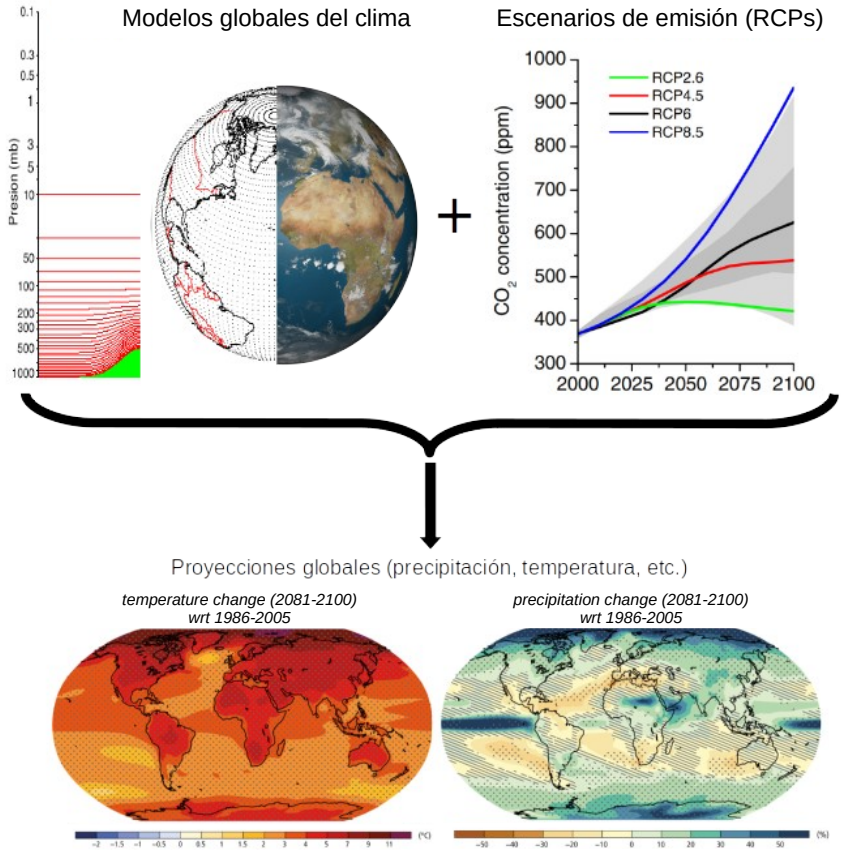

Figura 2.4: Esquema del proceso de generación de proyecciones globales de cambio climático.

	RCP 2.6	RCP 8.5
Dependencia con combustibles fósiles	No se usan	Se usan demasiado
Población mundial para el año 2100	9 mil millones	12 mil millones
Emisiones de CO_2 para el año 2100	Como hasta ahora hasta 2020, después se descarta y negativo para 2100	Tres veces las emisiones actuales
Emisiones de metano	Reducidas un 40 %	Incremento rápido
Uso de campos de cultivo	Se incrementa por la producción de bio-energía	Determinado por el incremento en la población

Tabla 2.3: Comparación entre el RCP más optimista (2.6) y el más pesimista (8.5).

Tabla 2.4: Los 4 ESMs del CMIP5 considerados en este TFM.

Nombre	***Institución***	***Resolución***	***Referencia***
CanESM2	CCCMA	$2.8° \times 2.8°$	Chylek et al. (2011)
CNRM-CM5	CNRM-CERFACS	$1.4° \times 1.4°$	Voldoire et al. (2011)
GFDL-ESM2M	NOAA GFDL	$2.5° \times 2°$	Dunne et al. (2012)
NorESM1-M	NCC	$1.5 \times 1.9°$	Kirkevag et al. (2008)

CAPÍTULO 3

Métodos

3.1. Regionalización Estadística

Como se ha mencionado en la Sección 2.3 los GCM nos permiten simular el estado del sistema climático sobre un espacio discretizado de celdas tridimensionales, por lo que, para poder llegar a la escala local/regional necesaria en la mayoría de aplicaciones de impacto es necesario aplicar algún tipo de regionalización o *downscaling*.

La regionalización estadística se basa en modelos matemáticos que relacionan de forma empírica las variables de circulación atmosférica a gran escala dadas por los GCM (*predictores*) con las variables locales observadas en superficie (*predictandos*), en nuestro caso precipitación. Típicamente, se distinguen dos fases diferentes en cualquier método de downscaling estadístco; el de ajuste/calibración ó y el de aplicación. El primero de ellos es utilizado para inferir la relación predictor-predictando. Para ello es necesario disponer de un registro suficientemente largo (20 años o más) de osbervaciones de las variables que se desean regionalizar (ya sea en localidades puntuales ó en rejillas interpoladas) y de las variables de larga escala que se consideren. En el segundo, las relaciones halladas se aplican con fines predictivos.

En los métodos de *perfect prog* (Wilks, 2006), la fase de ajuste/calibración se lleva a cabo con predictores provenientes de un reanálisis, lo que asegura una correspondencia temporal (día a día) entre predictor y predictando —los reanálisis proporcionar una estimación del estado real de la atmósfera, día a día, en un período

retrospectivo.— Esta correspondencia es la que hace posible inferir una relación físicamente verosímil entre predictor y predictando.
Una vez el método ha sido calibrado en condiciones perfectas, la relación predictor-predictando aprendida puede ser aplicada a nuevos predictores provenientes de un GCM; por ejemplo, proyecciones de cambio climático, adaptando así sus salidas crudas de baja resolución (no regionalizadas) a la escala local/regional de interés.

Para cada predictando y región de interés, la tarea más costosa es la selección de una combinación de variables predictoras adecuadas, que deben ser definidas sobre un dominio geográfico adecuado. Más en concreto, las variables que se seleccionen como predictores deberían 1) conseguir explicar una parte importante de la variabilidad en el predictando y 2) ser adecuadamente reproducidas en el GCM (en comparación con el reanálisis) (Wilby et al., 2004). Por lo general, esto se consigue de forma empírica, probando diferentes combinaciones de predictores y/o dominios y evaluando su rendimiento en función de la/s métrica/s de validación de interés (véase, por ejemplo Gutiérrez et al., 2013; San-Martín et al., 2016).

3.2. Árboles de Clasificación y Regresión

La técnica con el que se trabajará en este TFM para encontrar las relaciones entre predictor y predictando es *Random Forest*. Esta técnica se basa en la conjunción de varios árboles de clasificación y regresión (*CART*, del inglés *Classfication and Regresion Trees*). Por lo tanto en este apartado se mostrará qué es un *CART*.

En un principio los árboles se crearon para resolver problemas de clasificación. Es decir, el objetivo es clasificar una variable categórica basada en unos ciertos predictores. Un árbol de decisión está formado por *nodos*, *ramas* y *hojas*. Cada nodo es una evaluación sobre una variable predictora. Cada rama es un valor que puede tomar la variable predictora. Cada hoja representa la clase final, es decir, un valor de la variable predictando. Un ejemplo de árbol de clasificación sencillo se puede ver en la Figura 3.1

Para poder construir un árbol de decisión existen distintos algoritmos. Todos ellos tienen una idea común: se van formando nodos en función del *poder de separación* que aporten. El algoritmo original en los árboles es el **ID3**, que se basa en la *ganancia de información* (IG, del inglés *Information Gain*, ver Ecuación 3.1) como criterio de saparación. El objetivo es maximizar esta ganancia reduciendo la incertidumbre en la división (entropía, H).

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

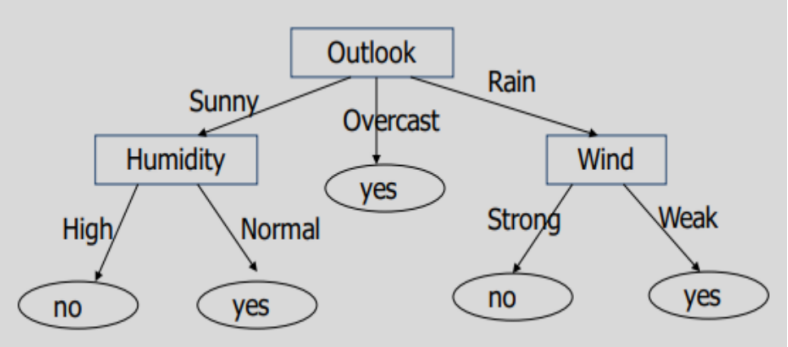

Figura 3.1: Ejemplo de árbol de clasificación para el clásico conjunto de datos *play tennis* (`https://github.com/sjwhitworth/golearn/blob/master/examples/datasets/tennis.csv`).

$$IG(X) = H(X) - H(X \mid Y), \tag{3.1}$$
$$H(X) = -\sum_{x \in X} p(x) log_2(p(x)), \tag{3.2}$$
$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) log_2(p(x \mid y)). \tag{3.3}$$

Existen otros criterios para cuantificar el poder de separación, por ejemplo el *índice de Gini* (ver Ecuación 3.4). En este caso, sólo se prueban divisiones binarias en cada nodo del árbol. Al contrario que con la ganancia de información, el objetivo en este caso es minimizar el índice de Gini.

$$GINI(X) = 1 - \sum_{x \in X} p_x. \tag{3.4}$$

También existen árboles de regresión. Estos son iguales que los árboles de clasificación salvo por dos diferencias. La primera y obvia es que la tarea a resolver en este caso es encontrar una relación entre predictores y predicanto siendo este último continuo, no discreto. La segunda es cómo se hacen las particiones en los nodos. En este caso el poder de separación, se calcula con el error cuadrático (Ecuación 3.5),

$$\sum_{j=1}^{J} \sum_{i:\, x_i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{3.5}$$

en donde J representaría el número de regiones en los que se particiona el espacio de características. Sin embargo, como sería imposible considerar todas las posibles particiones se sigue la técnica de *separación binaria recursiva*. Es descir, se construye el árbol nodo a nodo desde la raíz hasta las hojas. En un primer paso, se considera la región completa del espacio de características y se trataría de particionar esta en dos regiones. Para hacer esta separación se seleccionaría el predictor, j y el valor de separación, s, que minimicen la Ecuación 3.5 para $J = 2$. Es decir, la siguiente ecuación

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

en donde se tiene que

$$R_1(j, s) = \{X \mid X_j \leq s\},$$
$$R_2(j, s) = \{X \mid X_j > s\}.$$

Posteriormente se seguirá el mismo procedimiento con las regiones R_1 y R_2 hasta que se minimice el valor de la Ecuación 3.5, clasificando correctamente el conjunto de datos. Se puede ver una breve animación de este proceso en `https://bit.ly/2K1IBXh`.

Una de las principales ventajas de CART es que los árboles son herramientas muy intuitivas y fáciles de explicar, puesto que se pueden representar gráficamente. Sin embargo, también presentan desventajas importantes, siendo su alta sensibilidad a la partición train/test escogida la mayor de ellas. Además, si no se emplean ciertas restricciones (por ejemplo, el uso de técnicas de poda), los árboles tienden al sobreajustarse con facilidad, resultando en una capacidad de generalización (capacidad predictiva para nuevos datos que no han sido vistos en el entrenamiento) baja.

3.3. Random Forest

Con el fin de aliviar algunas de las limitaciones que se acaban de señalar para CART, nacen las técnicas de *ensembles*, que se basan en la idea de que juntando de alguna manera las predicciones de varios modelos, el rendimiento puede mejorar bastante. Existen dos tipos principales de técnicas de *ensembles*, las de *bagging* y las de *boosting*. En este TFM se trabajará únicamente con únicamente con técnicas de *bagging*, cuya idea es la siguiente: dada una muestra de datos se construyen t submuestras mediante remuestreo con repetición (*bootstrapping*) y sobre cada una de estas t submuestras se ajustan t modelos. Posteriormente la predicción es calculada a partir de las t predicciones de cada modelo (p. e. para regresión con la media de las predicciones o para clasificación con la categoría que más votos reciba). Se pude ver una ilustración de esta metodología en la Figura 3.2. En cada modelo individual, al conjunto de sucesos que no se seleccionan se les llama *"fuera de la bolsa"* (*OOB*, del inglés *Out-Of-Bag*). Se puede obtener una estimación de la capacidad de generalización de la técnica promediando los errores OOB que cometen los distintos modelos individuales.

Una de las técnicas más habituales de *bagging* es *random forest*, que fue por primera vez propuesta en Ho (1995) con la idea de mejorar el rendimiento ofrecido por CART. Un random forest se compone de un número dado (*ntree*) de árboles individuales, que se dejan crecer sobre *ntree* submuestras de datos obtenidas por medio de *bootstrapping* sobre la muestra orignal. Si sobre cada una de estas submuestras construyesemos un árbol como se ha descrito en la Sección 3.2 todos ellos serían

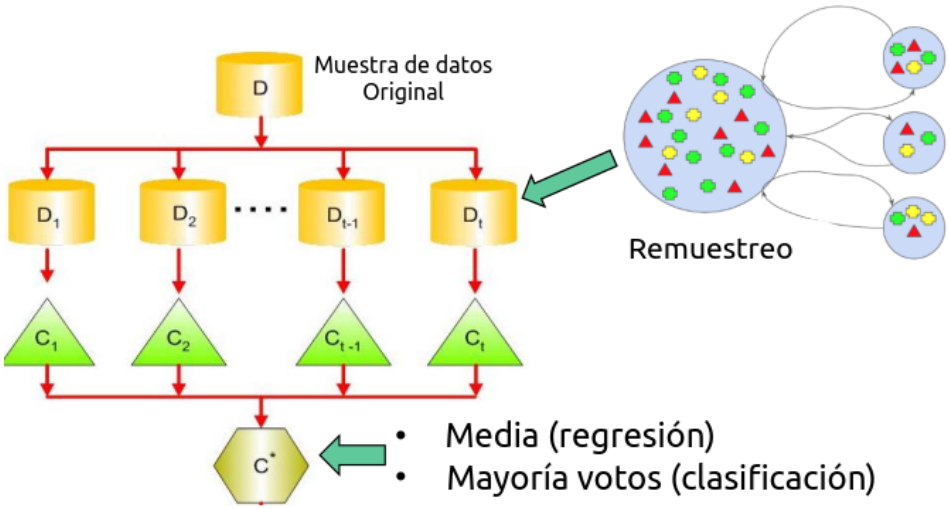

Figura 3.2: Ilustración de cómo funciona un método de *bagging*.

muy similares, pero los métodos de agrupación funciona mejor cuando los modelos son poco similares. Para ello, en random forest, al realizar el calculo del poder de separación en un nodo, este se realizará además sobre una submuestra aleatoria de $mtry$ predictores. Siendo p el número de predictores, típicamente los valores para $mtry$ son: Clasificación: $mtry = \sqrt{p}$, regresión: $mtry = p/3$. El resultado final predicho por random forest será la media de cada uno de los resultados obtenidos por los distintos árboles individuales:

$$P(y \mid x) = \frac{1}{ntree} \sum_{i=1}^{ntree} P_i(y \mid x),$$

donde $P_i(y \mid x)$ para cada $i = 1, \ldots, ntree$ es la predicción de cada árbol individual.

En este TFM se usa la implementación de random forest del paquete de *R*, *randomForest* (ver Liaw and Wiener (2002)). Nótese que, dado el carácter mixto de la precipitación, será necesario en nuestro caso combinar el resultado de dos random forest distintos, uno para la predicción del evento binario de *ocurrencia* (0 = no lluvia, 1 = lluvia, caracterizado por un umbral de 1mm,) y otro para el evento continuo *cantidad*.

3.4. Modelos Lineales Generalizados (GLMs)

La regresión lineal ordinaria predice el valor esperado de una variable aleatoria respuesta (predictando) como una combinación lineal de un conjunto de variables independientes (predictores). Esto implica que un cambio constante en un predictor

conduce a un cambio constante en la variable respuesta, lo cual resulta apropiado para variables gaussianas (como la temperatura, por ejemplo). Sin embargo, este comportamiento no se adecua a variables no gaussianas, como la precipitación.

Los modelos lineales generalizados (GLMs, del inglés *Generalized Linear Models*) son una generalización de la regresión lineal formulada por Nelder and Wedderburn (1972) en la década de los 70, que permiten modelar el valor esperado de variables aleatorias respuesta, Y, cuyas distribuciones pertenezcan a la familia exponencial, a través de una función arbitraria llamada *función de enlace*, g, y una serie de *parámetros desconocidos*, β, de acuerdo con la Ecuación 3.6:

$$E(Y) = \mu = g^{-1}(X\beta) \tag{3.6}$$

donde X es la variable independiente y $E(Y)$ el valor esperado de Y. Los parámetros desconocidos, β, siempre se pueden calcular por máxima verosimilitud, usando para ello un algoritmo iterativo de ajuste de mínimos cuadrados.

En el caso de la precipitación, y dado su carácter mixto, cualquier método basado en este tipo de técnicas ha de combinar los resultados de dos GLMs independientes, uno para la predicción del evento binario *ocurrencia* (0 = no lluvia, 1 = lluvia, caracterizado por un umbral de 1mm,) y otro para la del evento continuo *cantidad* —nótese que la ocurrencia sigue una distribución Binomial y la cantidad una Gamma.—

3.5. Análisis de Componentes Principales

Las técninas de regionalización estadística pueden ser ajustadas utilizando información sinóptica o local (ver Sección 4.1.3). En el primer caso, dado el gran volumen de datos que involucran habitualmente los problemas en meteorología, es conveniente aplicar algún tipo de técnica que permita comprimir dicho volumen sin una pérdida importante de información. Esto es lo que se consigue con el análisis de Componentes Principales (PCA, Preisendorfer (1988)) que permite eliminar información redundante, reduciendo por tanto la dimensionalidad en los datos y filtrando parte del ruido que no aporta nada a la hora de construir un modelo predictivo. Para los métodos de regionalización utilizados en este TFM que utilizan las componentes principales (PCs), se ha establecido como criterio usar las n primeras PCs que sean necesarias para explicar el 95 % de la varianza en el conjunto total de predictores.

3.6. Marco de Validación

3.6.1. Validación Cruzada

Con el fin de evitar el sobreajuste y tener por tanto una idea real de la capacidad de generalización de cualquier modelo predictivo es necesario considerar un marco de de validación cruzada adecuado. Una aproximación sencilla sería separar la muestra inicial en *train/test*; por ejemplo el 80 % de la muestra para el ajuste/trin y el 20 % restante para la evaluación/test.

La separación en *train/test* descrita anteriormente tiene el problema de que los resultados obtenidos pueden ser sensibles a cómo se haga dicha separación (particularmente en el caso de CART). Para suplir esta limitación se propuso la validación cruzada en k subconjuntos o *k-fold cross validation* (Kohavi et al., 1995). Esta metodología consiste en dividir el conjunto con el que los modelos son calibrados en k subconjuntos distintos. Para cada $1 \leq i \leq k$ se construye un modelo calibrado con los subconjuntos j-ésimos, $j \in \{1, \ldots, k\} \setminus \{i\}$, y se predice el conjunto i-ésimo. De este modo, tomando cada predicción es posible reconstruir la serie original completa.

Al igual que en Gutiérrez et al. (2018), en este TFM se utilizará para el estudio de idoneidad en condiciones perfectas (Sección 4.1) una validación cruzada con $k = 5$. En concreto, se trabajará con cinco bloques de seis años consecutivos cada uno (recuérdese que el período completo de estudio es 1979-2008).

3.6.2. Métricas de Validación

La validación de las predicciones de lluvia no es trivial debido a su carácter dual; es necesario considerar por separado el evento binario *ocurrencia* y el continuo *cantidad*. Para la validación de la *ocurrencia* se considera la siguiente métrica:

R01: Cociente entre el número de días húmedos (lluvia $>$ 1mm) predichos y observados.

Para la validación de la *cantidad* sólo se consideran los días de lluvia, tanto en la serie observada como en la predicha. Las métricas consideradas en este caso son:

SDII: Cociente entre la media predicha y la observada.

$$SDII = \frac{\overline{pred}}{\overline{obs}}$$

RV: Cociente entre la varianza predicha y la observada.

$$RV = \frac{\sigma_{pred}}{\sigma_{obs}}$$

R95P: Cociente entre el percentil 95 predicho y el observado.

Adicionalmente se calcula la **correlación**, que da una idea de la correspondencia día a día entre la serie observada y la predicha. Para esta métrica se consideran las series completas (ocurrencia multiplicada por cantidad), puesto que ambas tienen efecto sobre este estadístico. Se considera el *coeficiente de correlación de Spearman*, más adecuado que el de Pearson para variables no gaussianas.

Tanto **R01**, **SDII** como la correlación de Spearman son métricas que se utilizan en Gutiérrez et al. (2018). En este TFM se consideran también **RV** y **R95P** para evaluar la similitud entre las distribuciones predicha y observada. Nótese que el valor perfecto sería 1 para todas las métricas utlizadas aquí.

CAPÍTULO 4

Resultados y Discusión

En primer lugar, y siguiendo el marco experimental propuesto en Gutiérrez et al. (2008), la Sección 4.1 presenta un estudio de idoneidad en condiciones perfectas para la técnica random forest, comparándola con otra técnica de regionalización estadística más clásica (los GLM) sobre un conjunto de 86 estaciones repartidas por toda Europa.
A continuación, tomando como referencia el trabajo de Manzanas et al. (2019), se estudia la viabilidad de random forest para la generación de proyecciones locales de cambio climático. En particular, para un conjunto de 41 estaciones en Malawi, se analiza si random forest podría ayudar a solventar los problemas de extrapolacíon que sufren los GLM cuando se consideran ciertas variables predictoras.

4.1. Regionalización en Condiciones Perfectas

Con el fin de encontrar la configuración óptima de la técnica random forest, en esta sección se aborda tres estudios de sensibilidad: 1) al tamaño del bosque, 2) al tipo de dato utilizado en la calibración —anual frente a estacional— y 3) al carácter espacial de los predictores —local frente a regional.—

4.1.1. Sensibilidad al Tamaño del Bosque

Para agilizar los tiempos de cómputo, para este análisis se han considerado únicamente 2 estaciones pertenecientes a cada una de las ocho zonas PRUDENCE

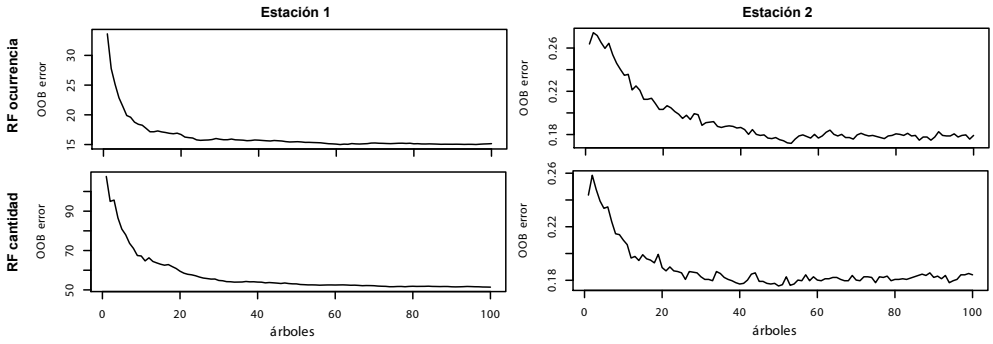

Figura 4.1: Evolución del error OOB en relación al tamaño del bosque para dos estaciones ilustrativas (en columnas). La primera fila representa el OOB cometido en la clasificación de la ocurrencia y la segunda en la estimación de la cantidad.

(Figura 2.1), que se indican en verde en la Tabla 2.1. Hay dos parámetros principales que pueden configurarse en la técnica random forest. El primero de ellos es el número de predictores que se consideran en cada división en los árboles, para el cual se mantiene el valor por defecto en este TFM (ver Sección 3.3). El segundo es el número de árboles del que consta el bosque. Para ajustar este parámetro, se hace uso de los errores *out-of-bag* (OOB), que se introdujeron en la Sección 3.3. La primera (segunda) fila de la Figura 4.1 muestra como disminuye el error OOB a medida que crece el tamaño del bosque en la predicción de la ocurrencia (cantidad) de precipitación, para dos estaciones (en columnas) en las que el error comienza a saturar para bosques de unos 50 árboles (en otras estaciones hay que llegar hasta los 200 árboles). Por tanto, como compromiso entre el error OOB y los tiempos de cómputo se utilizan en todo el TFM bosques formados por 100 árboles.

4.1.2. Entrenamiento Anual vs. Estacional

Mientras algunos estudios previos sugieren que la calibración de las técnicas de downscaling estadístico debería hacerse independientemente para cada estación del año (Maraun et al., 2010), hay otros que indican que es mejor utilizar a la vez todo el período disponible para el ajuste (Imbert and Benestad, 2005; Teutschbein et al., 2011). Actualmente no hay un consenso sobre este tema. Por tanto, esta sección se dedica a estudiar la influencia del tipo de dato utilizado (estacional frente a anual) en la calibración de las técnicas de regionalización empleadas en este TFM, tanto random forest como GLM (se utilizan las mismas 16 estaciones que en la sección anterior). Nótese que, en ambos casos (calibración anual y estacional), la validación se presenta por separado para las cuatro estaciones del año (Figuras de la 4.2 a la

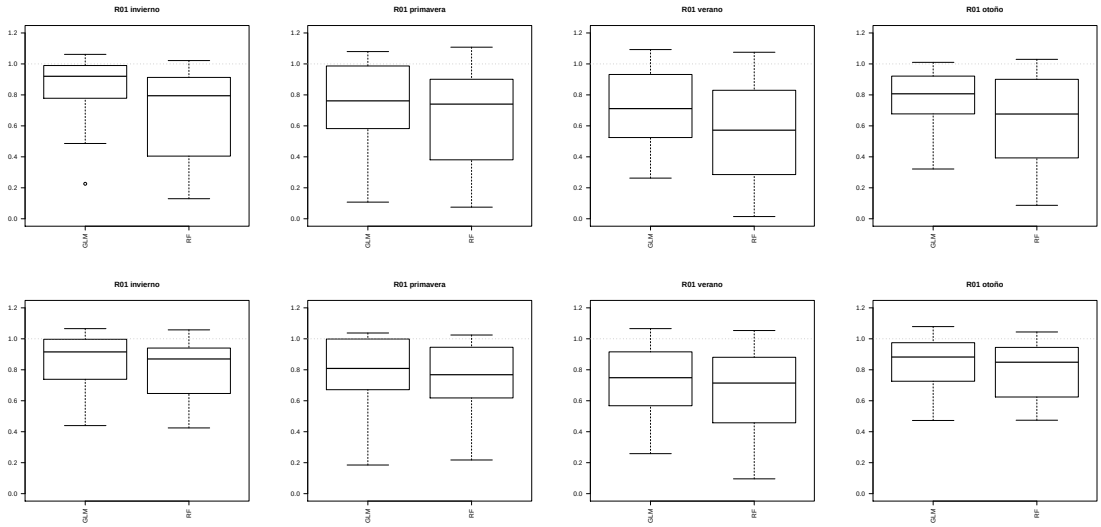

Figura 4.2: Comparación entre utilizar dato anual (fila de arriba) y estacional (fila de abajo) para el ajuste de GLM y random forest, en términos de la métrica R01. En ambos casos los resultados se muestran por separado para cada estación del año (columnas).

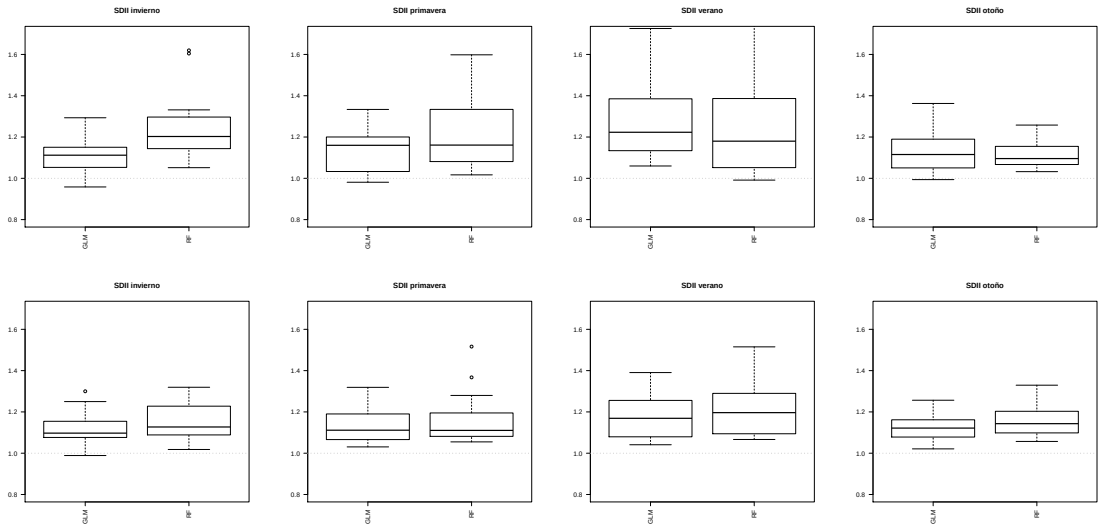

Figura 4.3: Como la Figura 4.2, pero para la métrica SDII.

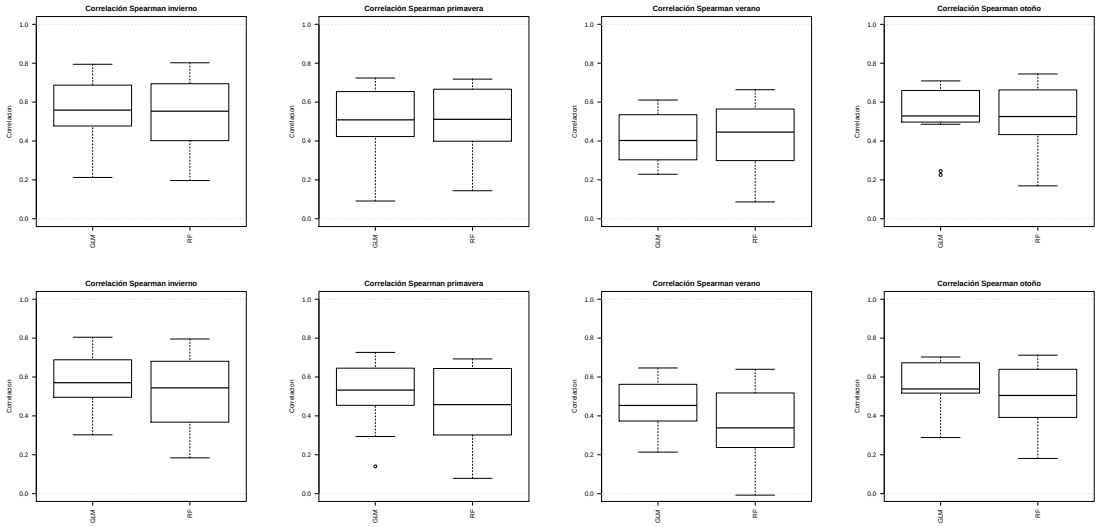

Figura 4.4: Como la Figura 4.2, pero para la correlación de Spearman.

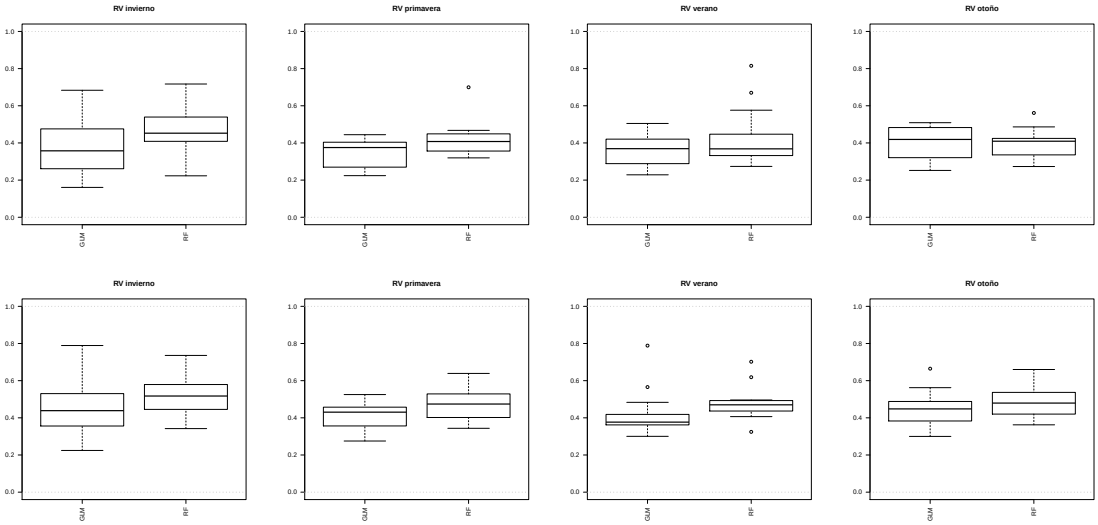

Figura 4.5: Como la Figura 4.2, pero para la métrica RV.

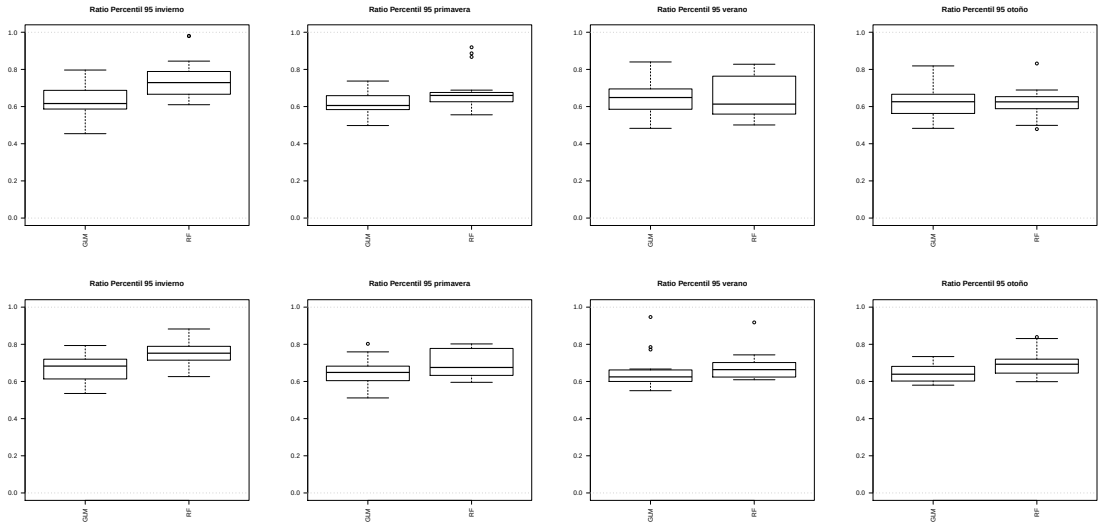

Figura 4.6: Como la Figura 4.2, pero para la métrica R95P.

4.6): Invierno (Diciembre-Febrero), primavera (Marzo-Mayo), verano (Junio-Agosto) y otoño (Septiembre-Noviembre).

Tanto para random forest como para GLM, los resultados son en general ligeramente mejores cuando se ajusta un modelo independiente para cada estación (fila de abajo), siendo la dispersión a lo largo de los 16 estaciones menor —en particular para el caso de la métrica SDII en verano (Figura 4.3).— Por este motivo, para el resto de análisis presentados en esta sección, todas las técnicas se entrenan con dato estacional.

Además, cabe destacar que, aunque en general los resultados hallados para GLM y random forest son bastante similares —la comparación entre las dos técnicas se hará de forma más exhaustiva (para las 86 estaciones) en la Sección 4.1.4,— estos últimos se comportan algo mejor para las métricas RV y R95P (Figuras 4.5 y 4.6, respectivamente).

4.1.3. Carácter Local vs. Regional de los Predictores

Las técnicas *perfect prog* pueden considerar predictores locales y/o regionales, utilizando el valor de las variables consideradas en las celdas cercanas y/o las Componentes Principales (PC, del inglés) correspondientes a las Funciones Ortogonales Empíricas (Preisendorfer, 1988) en un dominio geográfico representativo (que también debe ser determinado convenientemente). El uso de uno u otro tipo de predictores (o la combinación de ambos) depende de la aplicación. Por lo general, las

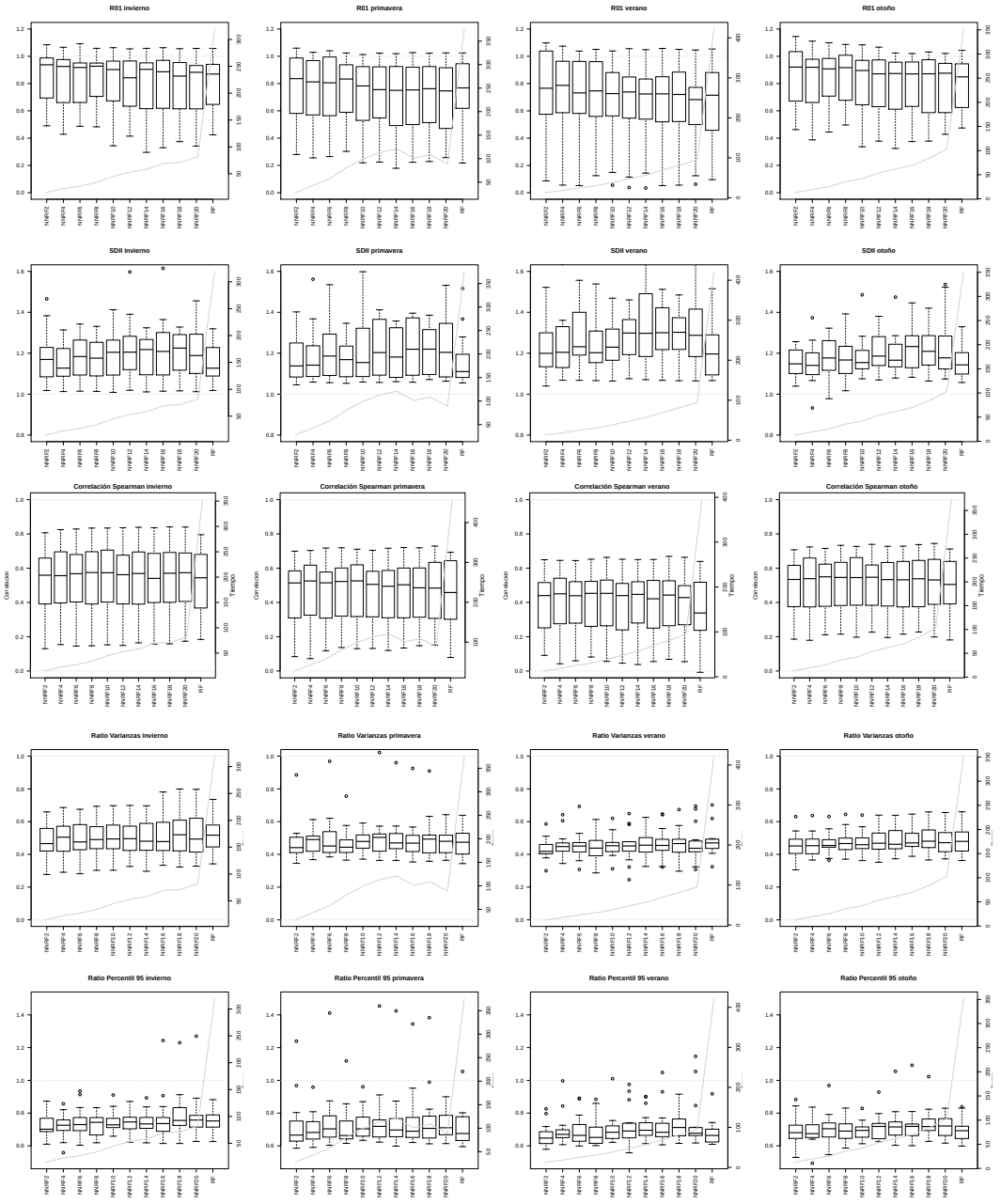

Figura 4.7: Estudio de sensibilidad al carácter espacial de los predictores para la técnica random forest, en función de las diferentes métricas consideradas (en filas), por estaciones del año (en columnas). Los boxplots 1-10 muestran resultados para el caso de predictores locales, incrementando la cantidad de celdas vecinas consideradas desde 2 hasta 20 (en saltos de 2). El boxplot más a la derecha corresponde al caso de predictores regionales, descritos por las PCs explicando el 95 % de la varianza del patrón completo. La línea gris muestra los tiempos de cómputo requeridos por cada configuración (escala en el lado derecho, en minutos).

PCs son más informativas en aquellos casos en los que el clima local está determinado principalmente por fenómenos sinópticos (por ejemplo, frentes fríos), mientras que la información en puntos cercanos suele ser útil en aquellos casos en los que los fenómenos de pequeña escala (por ejemplo, tormentas) son relevantes. Trabajar con PCs permite filtrar la variabilidad de alta frecuencia en los predictores de larga escala que puede no estar correctamente vinculada a la escala local.

Las Figura 4.7 muestra los resultados obtenidos para las distintas métricas de validación (en filas) para random forests que utilizan 1) predictores locales, en concreto anomalías estandarizadas en las celdas más cercanas, yendo de 2 celdas a 20 (boxplots 1-10) y 2) las PCs explicando el 95 % de la varianza de todos los predictores considerados sobre la correspondiente zona PRUDENCE (boxplot más a la derecha). Nótese que para este análisis se vuelven a utilizar las mismas 16 estaciones que se utilizaron para las dos secciones anteriores.

En general, puede observarse que no hay diferencias signficativas entre considerar predictores locales o regionales. De hecho, en el primer caso, apenas hay cambios apreciables entre considerar únicamente la información en las 2 o en las 20 celdas más cercanas. Sin embargo, las diferencias en los tiempos de cómputo (representados por una línea gris, escala secundaria situada a la derecha, en minutos) sí son significativas. Como es de esperar, a medida que se incrementa el número de predictores, aumentan los tiempos de cómputo. En particular, mientras que para el caso de considerar únicamente 2 celdas de predictores serían unos pocos minutos, en el caso de las PCs se puede llegar a varias horas. Es importante destacar que, al hacer una comparativa con GLM (no se muestra en la memoria), se vio que, por debajo de 10 celdas cercanas, los random forests utilizados en este TFM son más rápidos que los GLM utilizados en Gutiérrez et al. (2018).

También conviene destacar que, la gran dispersión encontrada en los resultados obtenidos —especialmente para R01 y la correlación de Spearman (primera y tercera fila de la Figura 4.7)— se debe a lo desbalanceada que está la muestra de 16 estaciones seleccionadas. Por ejemplo, las dos estaciones de la zona del Mediterraneo poseen muy pocos días de lluvia, por lo que el ajuste del random forest de cantidad es complicado, pudiendo dando lugar a resultados de validación anómalos (este efecto es especialmente apreciable en verano).

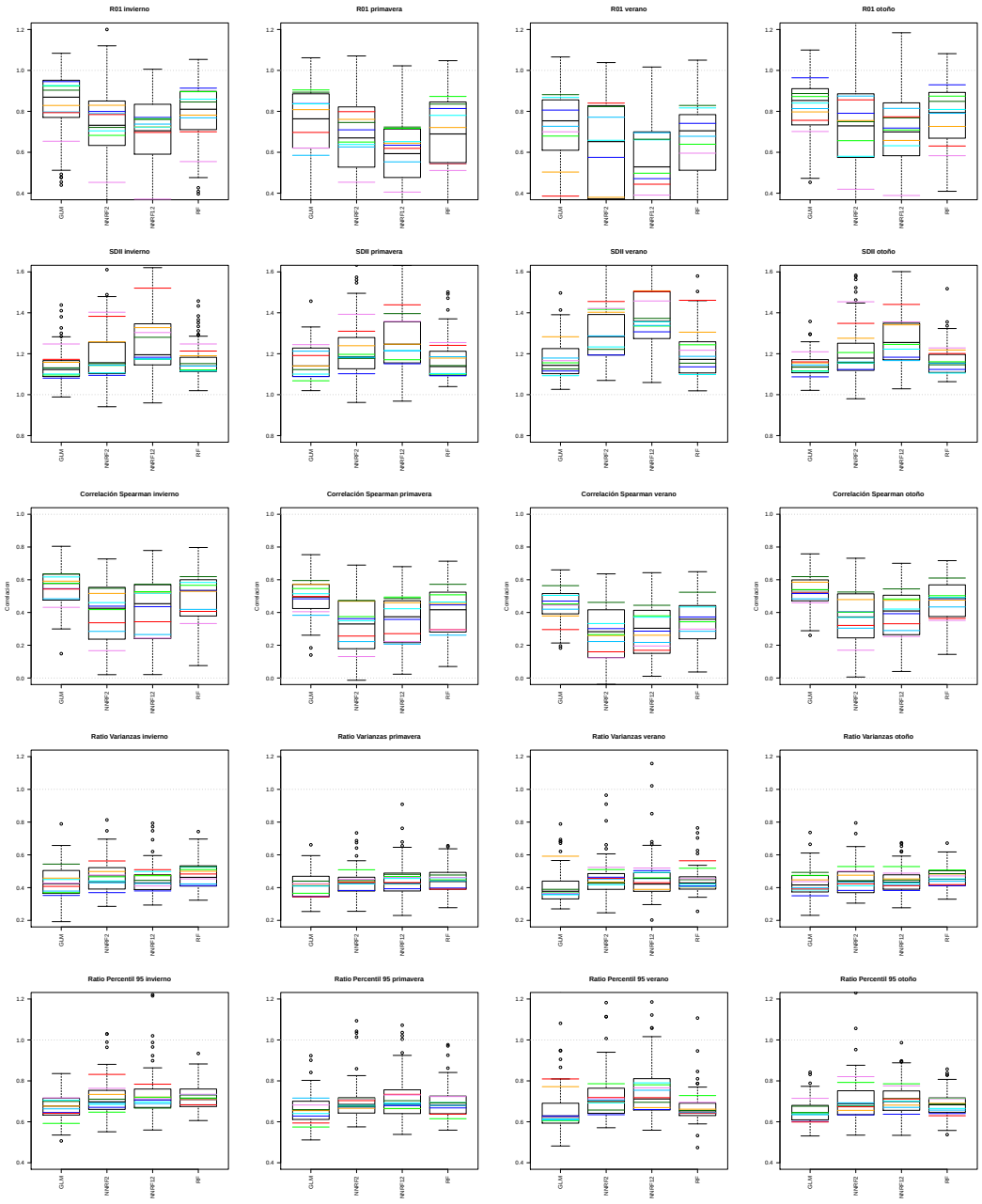

Figura 4.8: Estudio de idoneidad en condiciones perfectas para GLM y random forest, aplicadas sobre las 86 estaciones que se consideran en Gutiérrez et al. (2018), en función de las distintas métricas de validación consideradas en este TFM (en filas), por estaciones del año (en columnas). En el caso de random forest se consideran dos implementaciones, una que considera predictores locales en las dos celdas más cercanas y otra que considera PCs (veáse el texto para mayor detalle).

4.1.4. Comparación Random Forest vs GLM

Los resultados presentados en las secciones anteriores para la submuestra de 16 estaciones dejan intuir que random forest parece ser una técnica competitiva (comparada con GM) para la regionalización de precipitación. Para corrobar esta hipótesis, en esta sección se presenta una comparación más exhaustiva entre las dos técnicas en la que se consideran las 86 estaciones disponibles sobre toda Europa (Tabla 2.1). Los resultados obtenidos para las distintas métricas de validación consideradas se muestran en la Figura 4.8 (en filas). Nótese que en esta ocasión cada boxplot está acompañado de 8 líneas horizontales. Estas líneas representan la media en cada una de las regiones PRUDENCE, siendo el color identificativo de la región en cuestión (ver Figura 2.1). Basándonos en los resultados de las secciones anteriores, se consideran tanto un random forest que usa predictores locales en las 2 celdas más próximas a la estación sobre la que se quiere hacer el downscaling (opción más rápida) como un random forest que utiliza las PCs que explican el 95 % de todos los predictores considerados sobre la zona PRUDENCE que contiene cada estación (en ocasiones puede interesar mantener información regional más amplia). Adicionalmente, como opción intermedia, se decide incluir en el estudio también un random forest que considera predictores locales en las 12 celdas más cercanas (para el cual los tiempos de cómputo son similares a los de GLM). Estos tres tipos de random forest se comparan contra el GLM utilizado en Gutiérrez et al. (2018). Todas las técnicas son calibradas con dato estacional (ver Sección 4.1.2).

A diferencia de lo que observamos para la submuestra de 16 estaciones utilizada en la Sección 4.7 (Figura 4.7), podemos comprobar aquí que el hecho de considerar información local o regional en los predictores sí tiene un efecto. En particular, en el caso de predictores locales, los resultados son bastante más variables que cuando se consideran PCs, que resutan ser más robustas. Este resultado quizás tenga que ver con la gran variabilidad climática existente entre las 86 estaciones consideradas (mientras es de esperar que en algunas de ellas los fenómenos locales sean importantes, en otras el clima estará determinado mayormente por fenómenos sinópticos que sólo pueden capturarse mediante el uso de PCs en dominios grandes).
De todas maneras, si comparamos el random forest que usa PCs con GLM, vemos que, en general, se obtienen resultados muy similares en ambos casos. De hecho, las regiones para las que se obtienen los mejores (y los peores) resultados suelen coincidir en ambas técnicas. Incluso para casos particulares en los que ciertas estaciones presentan resultados de validación anómalamente raros (por ejemplo para el SDII en invierno), estos valores se repiten tanto en random forest como en GLM.

Más allá de la semejanza general, hay que señalar que los random forest muestran algunas de las desventajas típicas que ya se conocen para las técnicas basadas en modelos de regresión como GLM; en particular, la subestimación de la varianza observada (ver los resultados obtenidos para RV y R95P). En el caso de GLM, es práctica habitual para solventar esta limitación introducir un proceso de simulación estocástica a partir de los valores estimados similar al inflado de varianza que se suele realizar en la regresión ordinaria (Manzanas et al., 2015). Sin embargo, en principio, esta sería una limitación difícil de solventar en random forest. Además, y pese a que el sesgo tanto de GLM como de random forest es prácticamente nulo (no se muestra aquí), esto es resultado de una compensación entre la subestimación en el número de días de lluvia predicho (véase R01) y la sobrestimación en la cantidad media de lluvia predicha en los días húmedos (véase SDII). Por último, los GLM reproducen algo mejor que random forest las fluctuaciones observadas en las series diarias (mayores correlaciones).

Como resumen de esta sección, podemos concluir que random forest es, en general, una técnica competitva para la regionalización de precipitación, mostrando resultados similares a GLM para la mayoría de las métricas de validación consideradas, y pecando de algunas de las desventajas conocidas para las técnicas basadas en modelos de regresión.

4.2. Regionalización de Proyecciones Climáticas

Una vez las técnicas han sido calibradas (utilizando observaciones y predictores de reanálisis bajo un marco de validación cruzada; Sección 4.1), éstas pueden ser aplicadas tomando como predictores las simulaciones de un GCM, adaptando por tanto sus salidas de baja resolución a la escala local de interés. Conviene mencionar que cuando las técnicas de regionalización *perfect prog* son aplicadas a datos de GCM, los sesgos de los mismos deben tratarse adecuadamente (o *armonizarse*), es decir, deben hacerse compatibes con los datos de reanálisis utilizados para la calibración (Maraun et al., 2010); de lo contrario, se podrían obtener resultados engañosos. Por tanto, antes de entrar en cualquiera de los métodos *perfect prog* considerados en este TFM, a cada predictor de GCM se le corrige su sesgo (con respecto al reanálisis) mes a mes, forzándole así al menos a que siga el ciclo anual del reanálisis. Posteriormente, los datos del predictor GCM se estandarizan (celda a celda) restándoles su propia media y dividiendo por la desviación estándar del reanálisis. Por un lado, estos campos estandarizados se utilizan como datos de entrada en las implementaciones que consideran predictores en celdas cercanas. Por otro lado, para las que

consideran las PCs como predictores, las PCs del GCM se obtienen proyectando los campos estandarizados del GCM en las funciones ortogonales empíricas (EOF) del reanálisis utilizado para la calibración. También es importante señalar que en el caso de proyecciones de cambio climático (el que nos ocupa en esta sección) no es conveniente realizar el ajuste de las técnicas por separado para cada estación sino sobre todo el conjunto de datos disponible a la vez (ver discusión al respecto en la Sección 4.1.2), dado que las estaciones del futuro no tendrían porqué corresponderse con las observadas en el período histórico de calibración.

Usando un método de regionalización GLM que considera predictores locales en la celda más cercana a cada estación, Manzanas et al. (2019) presenta proyecciones locales de precipitación (hasta 2100) para un conjunto de 41 estaciones en Malawi (ver Figura 2.2). A pesar de dar lugar a resultados de validación muy similar en condiciones perfectas (es decir, cuando la técnica se entrena con el reanálisis ERA-Interim), en este trabajo se ve que ciertas variables predictoras pueden conducir a proyecciones futuras muy diferentes. En particular, la inclusión de la humedad específica entre los predictores da lugar a condiciones futuras mucho más húmedas que las proyectadas directamente (sin regionalizar) por los ESMs para la zona de estudio. Manzanas et al. (2019) argumentan que este efecto puede deberse al hecho de que esta variable aumente en los ESMs a niveles nunca vistos durante el período utilizado para la calibración (con reanálisis), lo que podría dar lugar a problemas de extrapolación en el caso de métodos de regionalización basados en regresión. Este efecto pernicioso desaparece cuando se sustituye la humedad específica por la relativa (variable acotada entre 0 % y 100 %), que da lugar a proyecciones locales más compatibles con la salida directa de los ESMs, y por tanto, más realistas (conviene dejar claro que, aunque se espera del downscaling que pueda introducir nuevos detalles locales no aportados por el ESM, la señal promedio dada por este último no debería verse significativamente alterada).

La Figura 4.9 —precipitación acumulada año a año, hasta el 2100, para el promedio de las 41 estaciones consideradas— corrobora, usando un GLM análogo al utilizado en Manzanas et al. (2019) los resultados obtenidos por en este trabajo cuando se utiliza un patrón predictor que contiene presión a nivel del mar (PSL) y humedad específica (Q) en 850 hPa —línea azul— y otro en el que la Q es remplazada por la humedad relativa (R) —línea verde.— Para comprobar si random forest presenta estos mismos problemas de extrapolación cuando se incluye la Q como predictor, se replicó el mismo experimento utilizando un random forest que también utiliza predictores locales en la celda más cercana a la estación en cuestión. La Figura 4.10 muestra que las proyecciones obtenidas con random forest son mucho más compa-

tibles con la salida cruda (sin regionalizar) de los distintos ESMs, proporcionando valores mucho más realistas que los mostrados en la Figura 4.9 y en Manzanas et al. (2019) para la técnica GLM. A la vista de estos prometedores resultados, decidimos calibrar el método random forest considerando el paquete completo de predictores que se explora en Manzanas et al. (2019), compuesto por un conjunto de variables de circulación y termodinámicas típicamente empleadas en problemas de regionalización de precipitación (véase, por ejemplo Charles et al., 1999; Timbal et al., 2003; Bürger and Chen, 2005; Haylock et al., 2006; Fowler et al., 2007; Hertig and Jacobeit, 2008; Sauter and Venema, 2011; San-Martín et al., 2016): presión a nivel del mar (PSL), temperatura (T) en 850 hPa, viento (U) en 250 hPa, humedad específica (Q) en 850 hPa y humedad relativa (R) en 850 hPa. La línea amarilla en la Figura 4.10 muestra los resultados obtenidos.

Al contrario de lo que se encuentra en Manzanas et al. (2019) para la técnica GLM, se obtienen valores intermedios, relativamente compatibles con la salida cruda de los distintos ESMs, lo que sugiere que random forest podría ser una alternativa robusta para la regionalización de proyecciones de cambio climático. Por tanto, con una visión muy optimista, uno se podría preguntar qué sucedería si se le dieran como entrada al método todos los predictores disponibles, olvidándose así por tanto del trabajo previo de selección de variables. Para comprobarlo, se repitió el experimento cogiendo un conjunto de predictores muy amplio disponible sobre la zona (variables identificadas con los colores azul, naranja y morado en la Tabla 2.2). En este caso, con el fin de aliviar el gran volumen de datos (y los consiguientes tiempos de cómputo), se consideran las PCs que explican el 95 % de la varianza del patrón completo. La línea roja muestra las proyecciones obtenidas.

Si bien es cierto que se aprecia una ligera tendencia positiva para la última parte del siglo en todos los ESMs, los resultados son muy prometedores, encontrando, salvo para el caso del CanESM2 —para el cual se obtiene una tendencia similar a las encontradas en Manzanas et al. (2019)— proyecciones mucho más realistas que las proporcionadas por la técnica GLM con un paquete de predictores más sencillos.

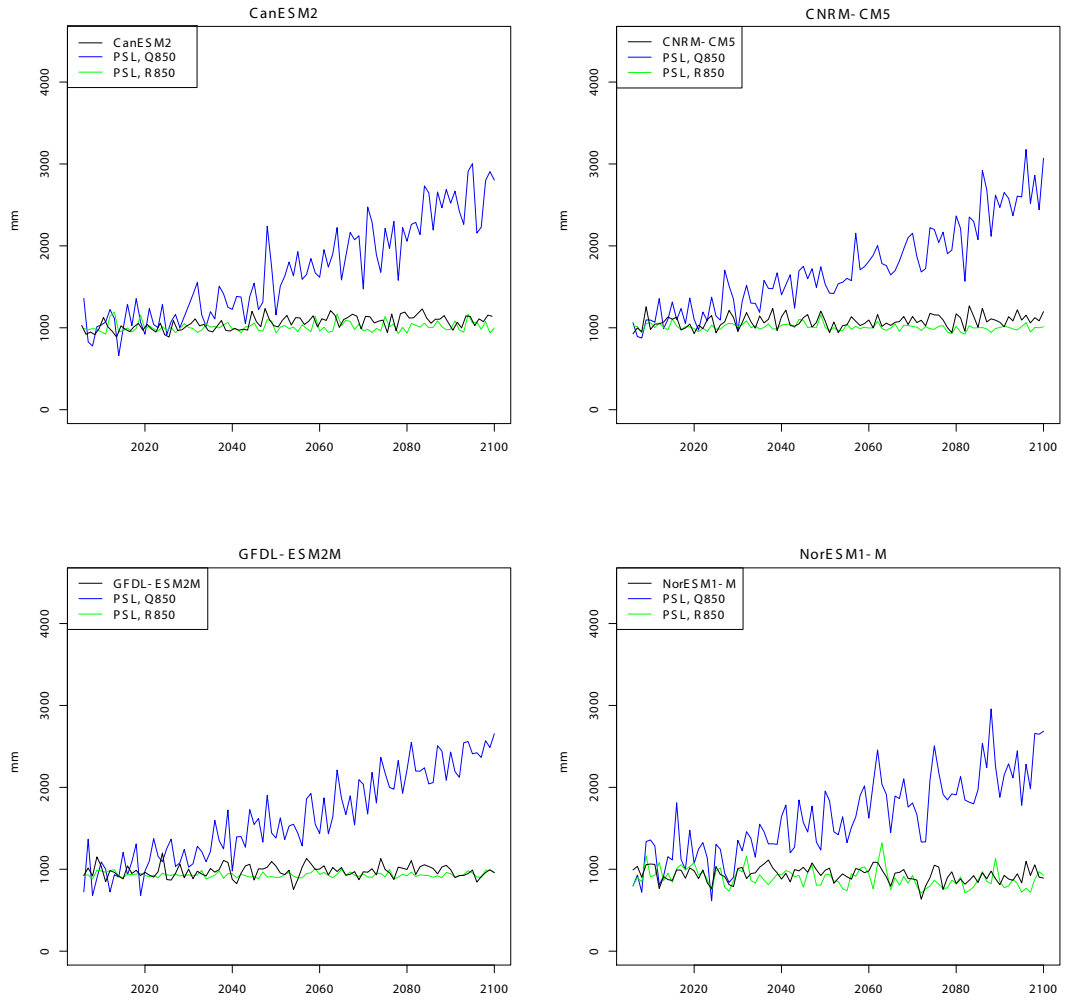

Figura 4.9: Precipitación acumulada, año a año, para el promedio de las 41 estaciones en Malawi, cuando se aplica la técnica GLM con distintos paquetes de predictores (distintos colores). Con fines comparativos, la salida cruda de los ESMs se muestra en negro (véase el texto para mayor detalle).

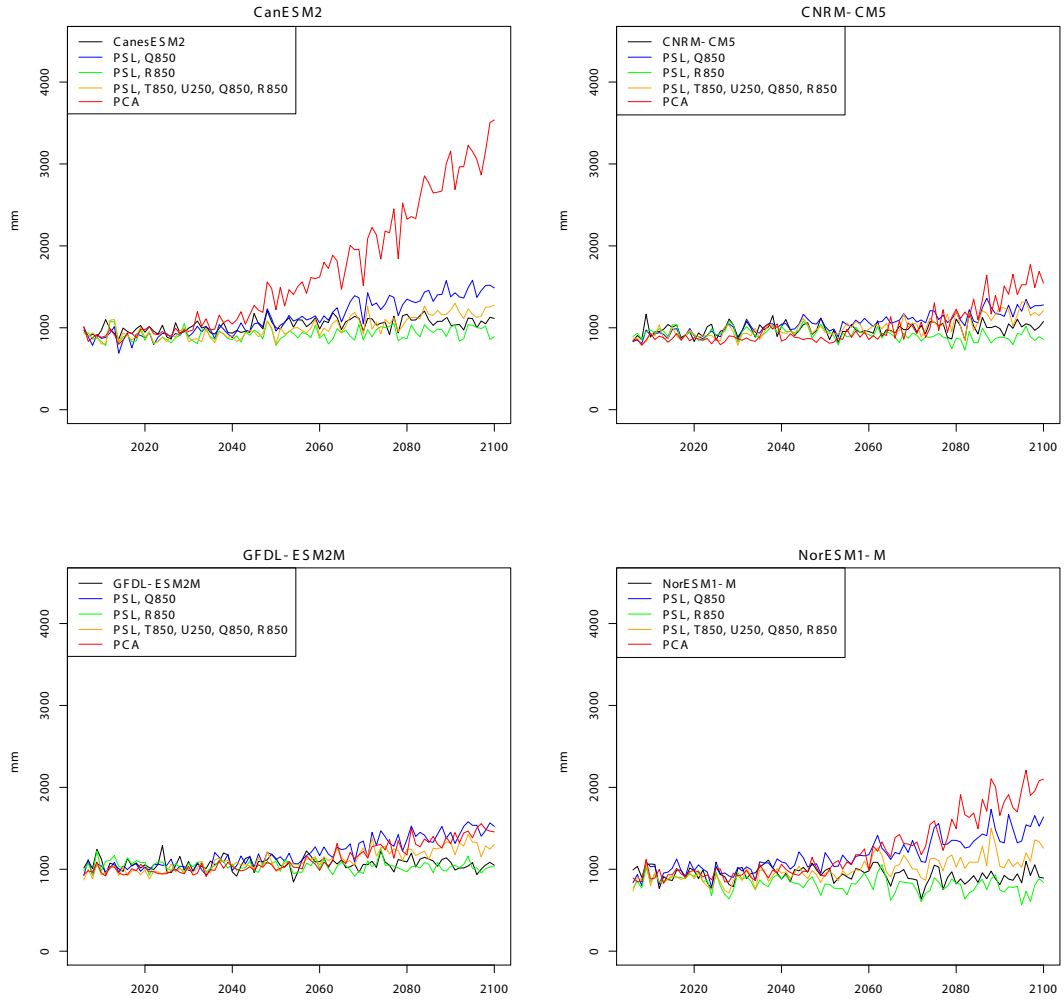

Figura 4.10: Precipitación acumulada, año a año, para el promedio de las 41 estaciones en Malawi, cuando se aplica la técnica random forest con distintos paquetes de predictores (distintos colores). Con fines comparativos, la salida cruda de los ESMs se muestra en negro (véase el texto para mayor detalle).

CAPÍTULO 5

Conclusiones y Trabajos Futuros

5.1. Conclusiones Principales

- Siguiendo el marco experimental propuesto en Gutiérrez et al. (2018), se ha evaluado en la Sección 4.1 el potencial de la técnica random forest para la regionalización estadística de precipitación en condiciones perfectas, encontrándose resultados similares a los proporcionados por otra técnica más clásica (GLM) para todas las métricas de validación consideradas en este TFM. Por tanto, random forest se perfila como otra opción válida para futuros estudios de regionalización de precipitación.

- En relación al punto anterior, Manzanas et al. (2019) han demostrado recientemente que la selección de predictores puede ser una fuente de incertidumbre muy importante en la regionalización de proyecciones de cambio climático. Por tanto, una vez vistos los resultados obtenidos en la Sección 4.1, en la Sección 4.2 se utiliza random forest para generar proyecciones de cambio climático en Malawi, siguiendo el marco experimental propuesto en Manzanas et al. (2019). Nuestros resultados indican que random forest da lugar a proyecciones más realistas que las otenidas con la técnica GLM en el citado estudio, sin necesidad de preocuprase por el trabajo previo de selección de predictores, que constituye una de las tareas más complejas en cualquier método de regionalización *perfect prog.*

5.2. Trabajos Futuros

- Actualmente, se está trabajando con modelos sencillos que involucran el uso de datos sintéticos para tratar de entender mejor los resultados obtenidos en la Sección 4.2; y, en particular, porqué las proyecciones con el modelo CanESM2 tienden a desviarse de lo esperado cuando se consideran PCs como predictores (Figura 4.10). La idea es recoger los resultados de este análisis, junto con parte de los mostrados en este TFM, en un artículo que tratará de publicarse en los próximos meses.

- En este TFM se ha considerado únicamente la precipitación como variable objetivo. Sin embargo, también queremos evaluar el potencial de la técnica random forest para la regionalización de otras variables de interés como la temperatura y el viento.

- Por último, más allá de las proyecciones de cambio climático, también nos gustaría probar esta técnica en el contexto de otro tipo de predicción con un horizontal temporal más corto, la predicción estacional (Doblas-Reyes et al., 2013), cuyas características son muy diferentes a las de las proyecciones de cambio climático tratadas en este TFM.

5.3. Reproducibilidad de Resultados

En este TFM se le ha prestado una atención especial a la reproducibilidad de los resultados, algo que se está convirtiendo (cada día más) en una preocupación importante en cualquier disciplina científica. Por tanto, con el fin de asegurar la transferencia y reproducibilidad de los resultados mostrados en esta memora, todo el código utilizado en la elaboración de este TFM se encuentra públicamente disponible en *GitHub*: `https://github.com/MiguiTE/TFM`. Siendo en las ramas *datasciencehub* y *validacion* donde se encuentra divido el código para la calibración de los modelos usados y la validación de los mismos, respectivamente. Además, todos los cálculos se han realizado en el lenguaje de programación *R*, de uso libre. En particular, se han utilizado los paquetes *randomForest* (Liaw and Wiener, 2002) y el entorno *climate4R* (Iturbide et al., 2019), por lo que cualquier usuario podría reutilizar el código proporcionado.

Bibliografía

Bundel, A. Y., Kryzhov, V. N., Min, Y. M., Khan, V. M., Vilfand, R. M., and Tishchenko, V. A. (2011). Assessment of probability multimodel seasonal forecast based on the APCC model data. 36(3):145–154.

Bürger, G. and Chen, Y. (2005). Regression-based downscaling of spatial variability for hydrologic applications. 311(1-4):299–317.

Charles, S. P., Bryson, C. B., Whetton, P. H., and Hughes, J. P. (1999). Validation of downscaling models for changed climate conditions: Case study of southwestern Australia. 12(1):1–14.

Christensen, J. H. and Christensen, O. B. (2007). A summary of the prudence model projections of changes in european climate by the end of this century. 81(1):7–30.

Chylek, P., Li, J., Dubey, M., Wang, M., and Lesins, G. (2011). Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM. 11:22893—-2290.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Koehler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. 137(656):553–597.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: Status and prospects. 4(4):245–268.

Dunne, J. P., Jasmin, G. J., Alistair, J. A., Stephen, M. G., Robert, W. H., Shevliakova, E., Ronald, J. S., Cooke, W., Krista, A. D., Matthew, J. H., John, P. K., Sergey, L. M., Milly, P. C. D., Peter, J. P., Lori, T. S., Bonita, L. S., Michael, J. S., Michael, W., Adrew, T. W., and Niki, Z. (2012). GFDL's ESM2 global coupled climate–carbon Earth System Models. Part I: Physical formulation and baseline simulation characteristics. 25():6646—-6665.

Fowler, H. J., Blenkinsop, S., and Tebaldi, C. (2007). Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. 27(12):1547–1578.

Gutiérrez, J., Primo, C., Rodríguez, M., and Fernández, J. (2008). Spatiotemporal characterization of ensemble prediction systems - the mean-variance of logarithmics (mvl) diagram. *Nonlinear Processes in Geophysics*, No. 15:109–114.

Gutiérrez, J. M. et al. (2018). An intercoparison of a large esemble of statistical downscaling methods over europe: Results from the value perfect predictor cross-validation experiment. *International Journal of Climatology*, pp. 1–36. URL `https://doi.org/10.1002/joc.5462`.

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., and Herrera, S. (2013). Reassessing statistical downscaling techniques for their robust application under climate change conditions. 26(1):171–188.

Hamed, N. H., Husein, H. N., and Mohammed, H. R. (2018). Land surface temperature downscaling using random forests in central baghdad. 10(7).

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., and Goodess, C. M. (2006). Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios. 26(10):1397–1415.

He, X., Chaney, N. W., Schleiss, M., and Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. 52(10):8217–8237.

Hertig, E. and Jacobeit, J. (2008). Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. 28(8):1025–1045.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282. IEEE.

Imbert, A. and Benestad, R. (2005). An improvement of analog model strategy for more reliable local climate change scenarios. 82(3-4):245–255.

Iturbide, M., Bedia, J., Herrera, S., Baño, J., Fernández, J., Frías, M. D., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A. S., and Gutiérrez, J. M. (2019). The R-based climate4R open framework for reproducible climate data access and post-processing. 111:42–54.

Kirkevag, A., Iversen, T., Seland, O., Debernard, J. B., Storelvmo, T., and Kristjansson, J. E. (2008). Aerosol-cloud-climate interactions in the climate model CAM-Oslo. 60(3):492–512.

Klein Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. 22(12):1441–1453.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, vol. 14, pp. 1137–1145. Montreal, Canada.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. 2(3):18–22. URL `https://CRAN.R-project.org/doc/Rnews/`.

Manzanas, R., Brands, S., San-Martín, D., Lucero, A., Limbo, C., and Gutiérrez, J. M. (2015). Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines. 28(10):4171–4184.

Manzanas, R., Fiwa, L., Vanya, C., Kanamaru, H., and Gutiérrez, J. M. (2019). Implausible regional climate change projections obtained from statistical downscaling: A case-study for precipitation in malawi. En revisión en Climatic Change.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. 48(3):n/a–n/a.

Murphy, J. (1999). An evaluation of statistical and dynamical techniques for downscaling local climate. 12(8):2256–2284.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. 135(3):370–384.

Pang, B., Yue, J., Zhao, G., and Xu, Z. (2017). Statistical downscaling of temperature with the random forest model. 2017.

Preisendorfer, R. (1988). *Principal component analysis in meteorology and oceanography*. Elsevier, 1st ed.

San-Martín, D., Manzanas, R., Brands, S., Herrera, S., and Gutiérrez, J. M. (2016). Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods.

Sauter, T. and Venema, V. (2011). Natural three-dimensional predictor domains for statistical precipitation downscaling. 24(23):6132–6145.

Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J., and Schmith, T. (2007). Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. 112(D4):n/a–n/a.

Schoof, J. and Pryor, S. (2001). Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. 21(7):773–790.

Shi, Y. and Song, L. (2015). Spatial downscaling of monthly trmm precipitation based on evi and other geospatial variables over the tibetan plateau from 2001 to 2012. 35(2):180–195.

Taylor, K. E. (2001). Summarizing multiple aspects of model performace in a single diagram. 106(D7):7183–7192.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of cmip5 and the experiment design. 93(4):485–498.

Teutschbein, C., Wetterhall, F., and Seibert, J. (2011). Evaluation of different downscaling techniques for hydrological climate-change impact studies at the catchment scale. 37(9-10):2087–2105.

Timbal, B., Dufour, A., and McAvaney, B. (2003). An estimate of future climate change for western France using a statistical downscaling technique. 20(7-8):807–823.

Voldoire, A., Sánchez-Gómez, E., Salas y Mélia, D., Decharme, B., and Cassou, C. (2011). The CNRM-CM5.1 global climate model: Description and basic evaluation.

Wilby, R. L., Charles, S., Zorita, E., Timbal, B., Whetton, P., and Mearns, L. (2004). Guidelines for use of climate scenarios developed from statistical downscaling methods. Tech. rep., IPCC-TGCIA.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences*. Amsterdam, Elsevier, 2nd ed.

Wu, H. and Li, W. (2019). Downscaling land surface temperatures using a random forest regression model with multitype predictor variables. 7:21904–21916.