PhD THESIS

DEEP CONVOLUTIONAL NEURAL NETWORKS FOR STATISTICAL DOWNSCALING OF CLIMATE CHANGE PROJECTIONS

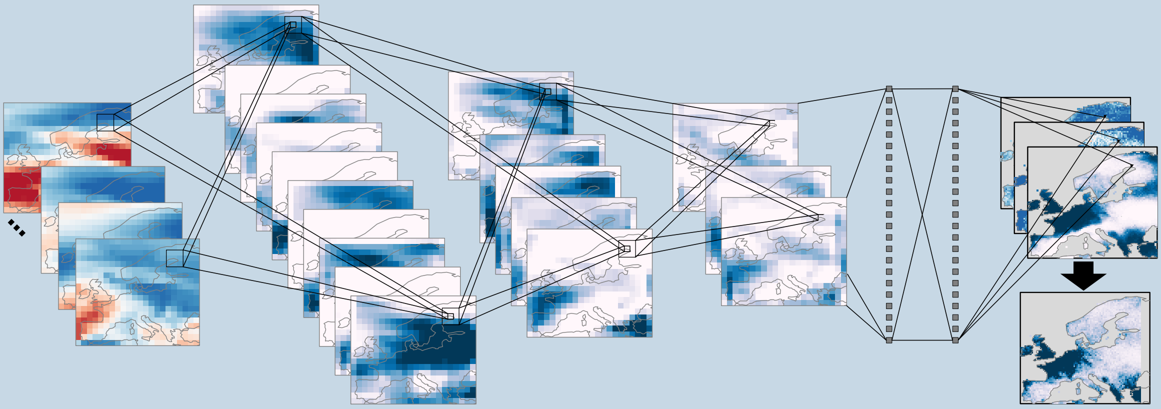
TESIS DOCTORAL

REDES NEURONALES DE CONVOLUCIÓN PROFUNDAS PARA LA REGIONALIZACIÓN ESTADÍSTICA DE PROYECCIONES DE CAMBIO CLIMÁTICO

UNIVERSIDAD DE CANTABRIA

PROGRAMA DE DOCTORADO EN CIENCIA Y TECNOLOGÍA

2021



AUTOR: Jorge Baño Medina

DIRECTORES: José Manuel Gutiérrez Llorente
Rodrigo García Manzanas

PhD Thesis

Deep convolutional neural networks for statistical downscaling of climate change projections

PhD Programme in Science and Technology

Presented by:

Jorge Baño Medina

Under the supervision of:

Dr. José Manuel Gutiérrez Llorente
Dr. Rodrigo García Manzanas

University of Cantabria

September 2021

Deep convolutional neural networks for statistical downscaling of climate change projections

Redes neuronales profundas de convolución para la regionalización estadística de proyecciones de cambio climático

Jorge Baño Medina
Santander Meteorology Group
Institute of Physics of Cantabria
(CSIC–University of Cantabria)
Santander, Spain
September 2021

“There is a pleasure in the pathless woods;
There is a rapture on the lonely shore;
There is society when none intrudes;
By the deep sea, and music in its roar;
I love not man the less, but Nature more...”

Lord Byron

Contents

Acknowledgements

I would like to thank the many people that helped me along these years.

First of all, it is a pleasure to express my most sincere gratitude to all my colleagues at the Santander Meteorology Group. Thank you for welcoming me to the group and for your valuable support, advise and guidance all these years.

Especially, I am deeply grateful to my supervisors, Dr. José Manuel Gutiérrez Llorente and Dr. Rodrigo García Manzanas. Thank you for considering me as a PhD candidate in the first place, and for taking your time to train me, not only in the topic of interest, but also in valuable competencies relevant in any professional and personal environment. Also, thank you for your constant support and for your supervision of immeasurable worth.

At a more personal level, I would like to thank my family and friends, not only for their support all these years, but also for letting me share this magnificent adventure called life, with them. Also I would like to thank all the people that helped me to feel the city of Santander in a very special way.

Jorge Baño-Medina
September 2021

Acronyms

The following list contains the acronyms most used throughout the Thesis:

AD Activation Difference

BDL Bayesian Deep Learning

CGAN Conditional Generative Adversarial Network

CMIP5 Coupled Model Intercomparison Project Phase 5

CNN Convolutional Neural Network

CNRM Centre National de Recherches Météorologiques

CORDEX COordinated Regional Downscaling EXperiment

CVAE Conditional Variational Auto-Encoder

DD Dynamical Downscaling

DL Deep Learning

DLR Deutsches Zentrum für Luft- und Raumfahrt (German Aeroespace Center)

DRN Deep Residual Network

ECMWF European Center for Medium-Range Weather Forecasts

EURO-CORDEX EUROpean COordinated Regional Downscaling EXperiment

EURO-CORDEX ESD EUROpean COordinated Regional Downscaling EXperiment Empirical Statistical Downscaling

GAN Generative Adversarial Network

GCM Global Climate Model

GHG Greenhouse Gases

GLM Generalized Linear Model

IPCC Intergovernmental Panel on Climate Change

JCR Journal of Citation Reports

KS Kolmogorov-Smirnov test

LSTM Long-Short Term Memory Network

MIP Model Intercomparison Project

MLP Multi-Layer Perceptron

MOS Model Output Statistics

NN Neural Network

PC Principal Component

PDA Prediction Difference Analysis

PP Perfect Prognosis

P02 Percentile 2^{nd}

P98 Percentile 98^{th}

P98Wet Percentile 98^{th} of the wet-days distribution

RBG Red-Blue-Green

RCM Regional Climate Model

RCP Representative Concentration Pathway

RMSE Root Mean Squared Error

ROCSS Relative Operating Characteristic Skill Score

SD Statistical Downscaling

SDII Simple Daily Intensity Index

SMG Santander Meteorology Group

UN United Nations

US United States

VAE Variational Auto-Encoder

WG Weather Generator

CHAPTER 1

Resumen en Español

1.1 Contexto y Objetivos de la Tesis

Los modelos numéricos son la principal herramienta usada actualmente para estudiar la evolución del clima a diferentes escalas temporales, desde unos pocos días (predicción meteorológica) hasta varias décadas (proyecciones de cambio climático) en el futuro. Estos modelos resuelven numéricamente las ecuaciones que describen la dinámica del sistema climático (conservación de la masa y de la energía, ecuaciones de Navier-Stokes, etc.) sobre una malla tridimensional discretizada en el espacio formada por puntos de rejilla. Los requerimientos computaciones de estos modelos limitan su resolución espacial y temporal. Por ejemplo, los modelos globales del clima (GCM[1]) desarrollados bajo el paraguas del CMIP5[2] (Taylor et al., 2012) presentan resoluciones espaciales que van desde 1° hasta 3° (entre 100 y 300 km en latitudes ecuatoriales, aproximadamente). Estos modelos simulan la evolución del sistema climático a escala global en base a diferentes forzamientos[3] naturales y antropogénicos. A pesar de que los GCMs actuales reproducen satisfactoriamente gran parte de los patrones climáticos que tienen lugar a escalas sinópticas (del orden de los cientos de km), son incapaces de representar procesos importantes que ocurren en escalas espaciales menores que el tamaño del punto de rejilla usado en el modelo (p.e., la precipitación convectiva). La obtención de campos climáticos de alta resolución que

[1] Todos los acrónimos que aparecen a lo largo de este capítulo han sido definidos de acuerdo a sus siglas en inglés.

[2] El experimento de intercomparación de modelos acoplados (CMIP) representa la iniciativa más ambiciosa en estudios de modelización del clima a escala global. CMIP5 es la quinta fase/versión de CMIP.

[3] El término forzamiento se refiere a cualquier mecanismo que tiene el potencial de alterar el clima del planeta a través de cambios en su balance energético, por ejemplo erupciones volcánicas o cambios en la concentración de gases de efecto invernadero, entre otros.

resuelvan/implementen esta variabilidad local es clave para el desarrollo de aplicaciones de impacto en diversas actividades socio-económicas de interés como la energía, la agricultura, la hidrología o la sanidad. Además, la disponibilidad de simulaciones climáticas de alta resolución para las próximas décadas es clave para el desarrollo de políticas de sostenibilidad y planes de adaptación y mitigación ante el cambio climático. Por estos motivos, recientemente han surgido varias iniciativas internacionales cuyo principal objetivo es el de coordinar a la comunidad científica del clima para la generación de escenarios locales/regionales de cambio climático mediante distintas técnicas (p.e., ver Jacob et al. (2020)).

En este contexto, en las últimas décadas se han desarrollado dos enfoques diferentes para incrementar la resolución espacial de los GCMs de cara a su uso en estudios de impacto: la regionalización (o *downscaling*) dinámica y la estadística. Por un lado, el downscaling dinámico (DD) está basado en el uso de modelos numéricos regionales (RCM) que resuelven un conjunto de ecuaciones similar al usado en los GCMs pero a resoluciones espaciales más altas sobre una determinada región del mundo. Para ello se usan como condiciones de frontera las salidas de los GCMs (Rummukainen, 2010). Por otro lado, el downscaling estadístico (SD) construye modelos estadísticos o algoritmos que relacionan un conjunto de variables atmosféricas de larga escala y baja resolución (predictores) con un registro de observaciones a escala local (predictandos; típicamente temperatura y/o precipitación en superficie) sobre un área de interés. Esta Tesis se centra en un tipo en particular de SD —el cual es considerablemente menos costoso en términos computacionales que el DD,— conocido como *"Perfect-Prognosis"* (PP).

La principal particularidad del PP es que se utilizan observaciones, tanto para los predictores como para los predictandos, en la construcción de los modelos estadísticos o algoritmos. En el caso de los predictores es muy común recurrir a datos de reanálisis[4], mientras que para los predictandos se suelen utilizar tanto rejillas de observación a alta resolución como registros meteorológicos en estaciones puntuales. Una vez que la relación estadística es establecida en condiciones "perfectas" —es decir, usando observaciones (o *quasi*-observaciones) para predictor y predictando— esta puede ser aplicada a los predictores de baja resolución dados por los GCMs para distintos escenarios de forzamiento radiativo —que se definen en base a diferentes trayectorias de concentración de gases de efecto invernadero,— obteniendo así las correspondientes proyecciones de cambio climático de alta resolución hasta el final de siglo. Es importante destacara que el PP se construye sobre tres hipótesis clave que tienen que ver con la calidad del modelo estadístico inferido

[4] Un reanálisis es un *dataset* definido sobre una malla regular que cubre todo el globo y combina observaciones con predicciones meteorólogicas a corto plazo a través de un proceso de asimilación. Son la herramienta más precisa que existe en la actualidad para describir el estado de la atmósfera en un instante de tiempo determinado.

y su transferabilidad desde el campo de las observaciones hasta el mundo del modelo climático (Maraun and Widmann, 2018; Maraun et al., 2019). Estas hipótesis establecen que los predictores de baja resolución empleados en la construcción del modelo estadístico tienen que 1) ser suficientemente informativos para describir la variabilidad local del predictando de interés, y 2) ser realísticamente simulados por los GCMs —a nivel de compatibilidad distribucional con respecto a sus equivalentes en el reanálisis.— Además, con vistas a su utilidad en condiciones de cambio climático, 3) los modelos estadísticos deben mostrar cierta capacidad de extrapolación (con respecto a las condiciones en que se hayan calibrado). Hasta la fecha, una gran variedad de técnicas han sido utilizadas para ligar la larga y la pequeña escalas, por ejemplo los modelos lineales (Gutiérrez et al., 2019), los análogos (Hewitson and Crane, 1996), las *support vector machines* (Tripathi et al., 2006) y los *random forests* (Hutengs and Vohland, 2016), entre otras. A pesar de sus respectivos éxitos, ninguna de estas técnicas es capaz de tratar automáticamente la alta dimensionalidad del espacio de entrada (campo de predictores) sin sobreajustar. Por este motivo es habitual recurrir a técnicas de selección de variables o de compresión del espacio de los predictores como paso previo a la construcción del modelo estadístico (Gutiérrez et al., 2019). Este proceso, que a menudo es guiado por el conocimiento experto humano, suele conllevar cierta pérdida de información que puede ser relevante para explicar las fluctuaciones temporales del predictando local de interés.

Por ello, la comunidad científica del clima está dirigiendo su atención últimamente hacia las redes neuronales profundas (DL, Goodfellow et al. (2016)), y en particular hacia las redes de convolución (CNN)[5], que ya han demostrado ser de gran utilidad en otras disciplinas que involucran el uso de grandes volúmenes de datos tales como el reconocimiento de voz o la visión por ordenador. En concreto, algunos estudios previos en el campo de estas aplicacioens han puesto de manifiesto que las CNNs son capaces de 1) aprender patrones espaciales complejos de los datos, y 2) tratar automáticamente, y de forma eficiente, con espacios de alta dimensionalidad sin sobreajustar. Estas propiedades convierten a este tipo de redes en potenciales candidatas para una gran variedad de aplicaciones climáticas, incluyendo el downscaling estadístico. Sin embargo, hasta la fecha muy poco estudios —mayormente centrados en casos de estudio sintéticos (ver Vandal et al. (2018b) para un ejemplo ilustrativo),— han analizado la aplicabilidad del DL (y en concreto, las CNNs) a este problema. Estos primeros trabajos muestran resultados prometedores y una buena capacidad a la hora de reproducir las fluctuaciones de ciertas variables (p.e., temperatura y precipitación) a escala local, pero las topologías analizadas 1) no son aplicables en el marco del PP ya que se incumplen algunas de las hipótesis mencionadas anteriormente, 2)

[5] Una CNN es un tipo específico de red neuronal que es comúnmente empleada para aprender patrones no lineales en *datasets* que presentan cierta estructura espacial (LeCun et al., 1995).

en caso de verificar las hipótesis del PP, no han sido rigurosamente estudiadas (por ejemplo, atendiendo a los diferentes aspectos que determinan la calidad de una predicción), y 3) no han sido evaluadas en el espacio del modelo climático, es decir, usando predictores de GCMs.

De acuerdo con estas consideraciones, esta Tesis se centra en evaluar la idoneidad de los modelos de DL, en particular las CNNs, para el downscaling estadístico de simulaciones de cambio climático sobre Europa bajo el paradigma PP. En particular, se plantean los siguientes objetivos:

1. Analizar la aplicabilidad y el rendimiendo de las CNNs para el downscaling estadístico del clima en condiciones "perfectas" —es decir, usando datos de reanálisis como predictores.— Una de las principales cualidades a examinar será la capacidad de estos modelos para tratar espacios de alta dimensionalidad, inherentes en la mayoría de aplicaciones climáticas.

2. Evaluar los beneficios y las desventajas de topologías CNN de tipo *multi-site* —es decir, en las que las predicciones se realizan simultáneamente en varias localidades o *sites* con el mismo modelo estadístico,— frente a sus equivalentes *single-site*. En particular, se analizará la regularización implícita que ocurre en este tipo de topologías.

3. Mejorar la interpretabilidad sobre el comportamiento interno de las CNNs, las cuales son típicamente vistas como modelos tipo "caja negra". En concreto, se estudiará en detalle la relación predictor-predictando, midiendo la influencia que cada patrón ejerce sobre los resultados devueltos por la red.

4. Estudiar la idoneidad de las CNNs para el downscaling estadístico de escenarios de cambios climático. Para ello, se evaluará en primer lugar el rendimiento de las CNN para reproducir el clima observado cuando se hace downscaling del escenario histórico (hasta el 2005) de los GCMs. A continuación se explorará el potencial de estos modelos para producir proyecciones futuras de cambio climático, prestando especial a su plausabilidad, para lo que se compararán los resultados proporcionados por las CNNs con las propias salidas de un conjunto de GCMs y RCMs.

1.2 Principales Resultados y Conclusiones

Presentamos a continuación el marco metodológico considerado para el desarrollo de la Tesis, así como una discusión sobre los principales resultados y conclusiones obtenidos de la misma, en relación a los objetivos anteriormente expuestos.

En concreto, en la sección 1.2.1 se describe y expone el experimento llevado a cabo para evaluar el rendimiento de las CNNs en condiciones "perfectas", además de dos estudios de interpretabilidad que pretenden aportar conocimiento sobre la relación predictor-predictando que se establece en las redes. Esta sección se basa en el artículo titulado *"Configuration and intercomparison of deep learning neural models for statistical downscaling"*, publicado en la revista *Geoscientific Model Development*, y en tres artículos publicados en los proceedings del congreso internacional —en los años 2018, 2019 y 2020— denominado *Climate Informatics*: *"Deep convolutional networks for feature selection in statistical downscaling"*, *"Understanding deep learning decisions in statistical downscaling models"* y *"The importance of inductive bias in convolutional models for statistical downscaling"*.

En la sección 1.2.2 se describen los experimentos que abordan la idoneidad de las CNNs para el downscaling estadístico de GCMs con el fin de generar escenarios regionales de cambio climático sobre Europa. Esta sección se basa en los artículos titulados *"On the suitability of deep convolutional neural networks for downscaling climate change projections"* —publicado en la revista *Climate Dynamics*,— y *"DeepESD: An ensemble of regional climate change projections over Europe based on deep learning downscaling"* —en proceso de revisión en la revista *Nature Scientific Data*.—

1.2.1 Downscaling en Condiciones "Perfectas"

En esta sección, 1) se analiza el rendimiento de diversas topologías de CNNs utilizando como referencia modelos lineales generalizados (GLM), 2) se evalúan los beneficios de las arquitecturas *multi-site* frente a las *single-site*, y 3) se producen una serie de *saliency maps*[6] —utilizando para ello la técnica del análisis de las diferencias en las predicciones (PDA, Zintgraf et al. (2017))— con el fin de ganar interpretabilidad en la relación predictor-predictando de las CNNs desarrolladas en esta Tesis.

Para ello se ha seguido el marco experimental que se definió para el Experimento 1 de VALUE[7]. En consecuencia, nuestro interés se centra en la generación de predicciones diarias de temperatura y precipitación sobre Europa para el período 1979-2008. Para ello utilizamos el *dataset* de observaciones E-OBS (Cornes et al., 2018) —que cubre todo el continente a una resolución espacial de 0.5°— como predictando y una serie de variables de

[6] El término *saliency map* hace referencia a cualquier tipo de transformación que consiga trasladar la información contenida en un cierto espacio en el que se establecen relaciones complejas entre distintas variables a otro en el que la interpretabilidad sea mayor. Este tipo de herramientas han sido ampliamente utilizadas para el estudio de distintas topologías de DL (Simonyan et al., 2014; Zhou et al., 2016; Zintgraf et al., 2017; Montavon et al., 2018; Larraondo et al., 2019; Reimers et al., 2019; Toms et al., 2021).

[7] VALUE es una COST action europea diseñada con el fin de proporcionar un marco experimental que permitiese evaluar e intercomparar diferentes técnicas de SD para estudios de cambio climático. Esta iniciativa une a climatólogos, informáticos, científicos y empresarios con el fin de facilitar la transferencia de conocimiento entre sectores y mejorar la calidad de la investigación en este ámbito.

larga escala —altura geopotencial, temperatura del aire, humedad específica y velocidad del viento a distintas alturas— provenientes del reanálisis ERA-Interim (Dee et al., 2011) como predictores, estos últimos sobre una rejilla de 2°. La elección de predictores se hizo en base a la literatura previa existente (Huth, 2002, 2005; Gutiérrez et al., 2013; San-Martín et al., 2017; Gutiérrez et al., 2019). Nuestra principal aportación en este punto fue el diseño de un conjunto de modelos de DL que implementan distintas topologías —a grandes rasgos consisten en tres capas convolucionales seguidas (o no) de capas densas— con grados crecientes de complejidad, y su comparación con otras técnicas más tradicionales para el downscaling estadístico a nivel continental.

En comparación con los métodos de downscaling considerados como referencia —dos configuraciones distintas de GLMs que dieron muy buenos resultados en el mayor experimento de intercomparación de métodos de downscaling estadístico realizado hasta la fecha sobre Europa (Gutiérrez et al., 2019),— las CNNs desarrolladas en esta Tesis muestran una mayor capacidad explicativa de la variabilidad local, tanto para la temperatura como para la precipitación (especialmente para esta última). Esto es concecuencia de la habilidad de estas redes para 1) aprender patrones complejos y no-lineales que están presentes en los datos, y 2) tratar eficientemente y de forma automáticamente espacios de entrada (campo de predictores) de alta dimensionalidad. Este último aspecto constituye una clara ventaja con respecto a los métodos tradicionales de SD, puesto que el uso de técnicas de reducción de la dimensionalidad —que implican cierta pérdida de información— deja de ser necesario. Sin embargo, las razones que explican el buen comportamiento encontrado para las CNNs siguen siendo, en parte, desconocidas. Por ello, se describen a continuación los dos estudios de interpretabilidad llevados a cabo en esta Tesis con el fin de arrojar luz sobre el carácter de "caja negra" que habitualmente se le confiere a las redes neuronales.

En primer lugar construimos versiones *multi-site* y *single-siste* de las CNNs que mejores resultados obtuvieron en el estudio intercomparativo anterior —tres capas convolucionales de 50, 25 y 1 mapa de características, respectivamente.— Los resultados de este experimento indican que mientras las CNNs *single-site* son propensas a sobreajustar en algunas localidades, sus equivalentes *multi-site* muestran cierta capacidad de regularización implícita. Esto permite a las redes *muti-site* tratar simultáneamete todo el espacio de los predictores, evitando el sobreajuste. Además, el uso de modelos CNN *multi-site* puede traducirse en una mejora en la reproducibilidad de la escala local, especialmente en la predicción de la cantidad de precipitación.

En segundo lugar estudiamos la relación predictor-predictando que se establece en las CNNs de acuerdo a un análisis basado en PDA (ver Zintgraf et al. (2017) para más detalles sobre esta técnica). PDA evalúa la influencia que cada patrón de entrada tiene en la salida de los modelos de downscaling midiendo la diferencia en las predicciones según se utilice o no como predictor. En el caso de la precipitación, los resultados muestran una gran dependencia en la humedad específica del aire, lo cual es consistente con previos estudios sobre Europa de SD (San-Martín et al., 2017; Gutiérrez et al., 2019), aunque otras variables como la velocidad del viento y la altura geopotencial también ejercen cierta influencia en la predicción. A diferencia de la precipitación, la temperatura muestra una dependencia clara (y casi exclusiva) de los campos sinópticos de temperatura en las capas más bajas de la atmósfera. Este comportamiento también es consistente con lo descrito en estudios previos (Huth, 1999, 2002, 2004). Además, tanto para precipitación como para temperatura, únicamente un área de 5x5 puntos de rejilla centrados en la localidad de interés parece ser relevante en el proceso de downscaling, lo cual también está en la línea de lo encontrado en otros estudios (Timbal and McAvaney, 2001; Timbal et al., 2003; Gutiérrez et al., 2004; Brands et al., 2011b; Gutiérrez et al., 2013; San-Martín et al., 2017).

En líneas generales, los experimentos descritos anteriormente han permitido 1) analizar el rendimiento de las CNNs en condiciones "perfectas", encontrando una mayor capacidad explicativa para la escala local que en los métodos de SD tradicionalmente usados por la comunidad hasta la fecha y, 2) arrojar luz sobre el funcionamiento interno de las CNNs, lo cual resulta clave para que la comunidad científica gane confianza en el uso de este tipo de técnicas para aplicaciones climáticas.

1.2.2 Downscaling de Modelos Globales del Clima

En esta sección se analiza la idoneidad de las CNNs desarrolladas en esta Tesis para el downscaling estadístico de las simulaciones climáticas dadas por los GCMs, comparándolas contra métodos de SD más tradicionales como GLMs. Para ello, nos acogemos al marco experimental propuesto en EURO-CORDEX-ESD[8] y hacemos downscaling de la simulación número 12 del GCM denominado EC-Earth, llevando sus salidas de baja resolución a la rejilla de 0.5° de E-OBS, produciendo así campos diarios de temperatura y precipitación sobre toda Europa, tanto para el escenario histórico (1979-2008) como para el RCP8.5[9] (2071-2100). En el primer caso, los campos de alta resolución producidos son directamente validados contra E-OBS. Sin embargo, para el RCP8.5, dado que no existe un *dataset* de observaciones futuras, las propias salidas de baja resolución del EC-Earth (conveniente interpoladas a la rejilla de E-OBS) son usadas como "pseudo-realidad" contra la que comparar las proyecciones generadas con los distintos métodos de SD empleados. Este mismo

[8] EURO-CORDEX-ESD es una evolución de EURO-CORDEX —la rama europea del experimento coordinado de downscaling a escala regional (CORDEX), una iniciativa global que busca desarrollar escenarios de cambio climático de alta resolución a través del uso de RCMs— que se centra en SD.

[9] RCP8.5 describe un escenario extremo de emisiones en el que se seguirían liberando a la atmósfera gases de efecto invernadero sin restricción alguna hasta 2100, alcanzando para ese momento una presión radiativa promedio de 8.5 W/m^2.

enfoque ha sido ampliamente usado en la literatura (ver por ejemplo Vrac et al. (2007b); Quesada-Chacón et al. (2021)) y se basa en la idea de que variaciones significativas en la señal de cambio climático obtenida por métodos de SD —con respecto a la mostrada por los GCMs que se intentan regionalizar; en este caso el EC-Earth— pueden ser un indicador de la implausabilidad de las proyecciones generadas (a menos que esté justificado por procesos físicos conocidos).

Los modelos estadísticos analizados —tres configuraciones de GLMs y las CNNs que mejores resultados obtuvieron en el experimento en condiciones "perfectas" (sección 1.2.1),— permiten reducir los sesgos sistemáticos que exhibe el EC-Earth (al compararlo con E-OBS) en el período histórico. A pesar de ello, alguna de las variables incluidas en el conjunto de predictores utilizado parece incumplir la condición de PP que establece que el GCM debe ser similar al reanálisis, al menos en términos de distribuciones. Como consecuencia, se encuentran algunos sesgos en el caso de los GLMs, en concreto para ciertos índices relacionados con la precipitación. Este efecto indeseado se reduce cuando los predictores locales (información en un reducido número de puntos de rejilla cercanos a la localidad de interés) se sustituyen por predictores representativos de un dominio espacial más extenso —p.e., usando la técnica de análisis de componentes principales— o cuando se utilizan CNNs en lugar de GLMs. Además, en comparación con los GLMs, el downscaling sobre el escenario RCP8.5 (período 2071-2100) por parte de las CNNs produce patrones de cambio climático que son notoriamente más compatibles con la "pseudo-realidad" mostrada por el EC-Earth, tanto para la temperatura como para la precipitación. Estos resultados ponen de manifiesto que las CNNs son capaces de generar escenarios regionales de cambio climático a escalas continentales, sin la necesidad de hacer una elección "óptima" de predictores.

El análisis descrito anteriormente para el EC-Earth se extendió en su segundo experimento a un conjunto de ocho GCMs actuales incluídos en el CMIP5. En este caso, además de los propios GCMs regionalizados, se ha considerado también un subconjunto de RCMs de EURO-CORDEX como "pseudo-realidad" con el fin de evaluar la plausabilidad de nuestros campos de alta resolución. El resultado es un *ensemble* de proyecciones de cambio climático de precipitación y temperatura diarias sobre toda Europa para el siglo XXI que ha sido acuñado como *DeepESD*.

Nuestros resultados muestran que *DeepESD* reproduce satisfactoriamente los campos de precipitación y temperatura observados sobre Europa en el período histórico, lo que otorga cierta confianza en las proyecciones futuras —de hecho, en líneas generales, no se encuentran diferencias significativas entre las señales de cambio climático proyectadas por *DeepESD* y las dadas por los GCMs y RCMs considerados como "pseudo-realidad".— Sin embargo, *DeepESD* proyecta menores niveles de calentamiento para el futuro lejano

(2071-2100) que el simulado por el conjunto de GCMs. Este aspecto será analizado en detalle en un trabajo futuro que tratará de definir si estas diferencias son consecuencia de una mejora en la reproducibilidad de la escala local por parte de *DeepESD* o si por el contrario podrían deberse a violaciones en la condición de estacionariedad.

1.3 Principales Logros

1.3.1 Publicaciones

Los principales resultados de esta Tesis han derivado en una serie de publicaciones en revistas internacionales de alto impacto relacionadas con las ciencias atmosféricas y la inteligencia artificial. En particular, la sección 1.2.1 está basada en:

- **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "Configuration and intercomparison of deep learning neural models for statistical downscaling", *Geoscientific Model Development*, vol. 13, pp. 2109–2124, 2020, DOI: 10.5194/gmd-2019-278 (primer decil en JCR[10])

- **J. Baño-Medina** and J. M. Gutiérrez, "The importance of inductive bias in convolutional models for statistical downscaling", *Proceedings of the 9th International Workshop on Climate Informatics: CI 2019*, 2019, DOI: 10.5065/y82j-f154

- **J. Baño-Medina** and J. M. Gutiérrez, "Deep convolutional networks for feature selection in statistical downscaling", *Proceedings of the 8th International Workshop on Climate Informatics: CI 2018*, 2018, DOI: 10.5065/D6BZ64XQ

- **J. Baño-Medina**, "Understanding deep learning decisions in statistical downscaling models", *Association for Computing Machinery, New York, NY, USA, p 79–85*, 2020, DOI: 10.1145/3429309.3429321

  Por otro lado, los resultados descritos en la sección 1.2.2 están basados en las siguientes publicaciones:

- **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "On the suitability of deep convolutional neural networks for downscaling climate change projections", *Climate Dynamics*, 2021, DOI:10.1007/s00382-021-05847-0 (primer cuartil en JCR)

- **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "DeepESD: An Ensemble of Regional Climate Change Projections over Europe based on Deep Learning Downscaling", En revisión en *Nature Scientific Data* (primer cuartil en JCR)

[10] Journal of Citation Reports (JCR) es una herramienta incluída en la plataforma Web of Science (WOS) que permite cuantificar la importancia de una revista, dentro del conjunto de revistas que versan sobre la misma temática, en base en el número de citas promedio que reciben los artículos que publica.

Adicionalmente, como resultado de las actividades llevadas a cabo en el Grupo de Meteorología de Santander (SMG) en paralelo al desarrollo de la Tesis se han publicado dos artículos más relacionadas con el desarrollo de software para el tratamiento de datos climáticos (incluyendo herramientas para SD):

- M. Iturbide, J. Bedia, S. Herrera, **J. Baño-Medina**, J. Fernández, M.D. Frías, R. Manzanas, D. San-Martín, E. Cimadevilla, A.S. Cofiño and J.M. Gutiérrez, "The R-based climate4R open framework for reproducible climate data access and post-processing", *Environmental Modelling & Software, vol. 111, pp. 42-54*, 2019, DOI: 10.1016/j.envsoft.2018.09.009 (primer cuartil JCR)

- J. Bedia, **J. Baño-Medina**, M.N. Legasa, M. Iturbide, R. Manzanas, S. Herrera, D. San-Martín, A.S. Cofiño and J.M. Gutiérrez, "Statistical downscaling with the downscaleR package (v3.1.0): Contribution to the VALUE intercomparison project", *Geoscientific Model Development*, 2019, DOI: 10.5194/gmd-2019-224 (primer decil en JCR)

Todos estos artículos han sido desarrollado en base a los principios FAIR[11] de transparencia y reproducibilidad, ingredientes clave en la promoción de la ciencia de alta calidad. Para ello hemos creado un repositorio de GitHub (`https://github.com/SantanderMetGroup/DeepDownscaling`) que almacena el código que permite reproducir todos los resultados de la Tesis. Además, la publicación en abierto de dicho código asegura que el mismo pueda ser fácilmente adaptado por cualquier usuario en base en función de sus intereses particulares.

1.3.2 DeepESD

En base al conocimiento adquirido durante la Tesis, hemos desarrollado *DeepESD*, el primer *dataset* basado en DL que proporciona proyecciones de cambio climático de precipitación y temperatura diarias a partir de un conjunto de ocho GCMs, a una resolución de 0.5° sobre Europa. *DeepESD* se ha publicado en abierto y está disponible a través del "Earth System Grid Federation (ESGF)", en el nodo de la Universidad de Cantabria (`https://data.meteo.unican.es/thredds/catalog/esgcet/collections/CORDEX-DeepESD-EE/catalog.html`[12]). Por un lado, esperamos que este *dataset* por lo que se espera que constituya una referencia para la comunidad científica del clima para el estudio en profundidad

[11] Recientemente, la comunidad científica se ha unido para definir una serie de principios denominados FAIR (*findable*, *accesible*, *interoperable* y *reusable*) que sirvan de guía para promover el aprovechamiento por parte de cualquier usuario de los datos y/o el código generado en cualquier trabajo científico (Wilkinson et al., 2016)

[12] Esta dirección web es temporal dado que el artículo que describe *DeepESD* está actualmente en proceso de revisión en *Nature Scientific Data*. Una vez se haya publicado, se publicará una versión final del *dataset* en otra URL.

de las ventajas e inconvenientes que las redes neuronales puedan presentar para la generación de proyecciones climáticas de alta resolución (p.e., capacidad de extrapolación y condición de estacionariedad). Además, *DeepESD* proporciona un nuevo *ensemble* plausible de escenarios de cambio climático que complementa los ya existentes (p.e., las simulaciones númericas de los RCMs de EURO-CORDEX) y que, junto con aquellos, podrían ser utilizados tanto para el desarrollo de actividades de impacto en distintos sectores socio-económicos (p.e. energía, agricultura, salud, turismo, etc.) como para el diseño de políticas adecuadas de mitigación frente al cambio climático.

1.3.3 Software

Esta Tesis se construye sobre (y contribuye a) *climate4R* (C4R), un conjunto de librerías de *R* desarrolladas por el SMG que permiten abordar las particularidades y requerimientos de (casi) cualquier aplicación en estudios del clima. Para mayor detalle, referimos al lector al artículo de referencia en el que se describe C4R (Iturbide et al., 2019) y/o al siguiente repositorio de *GitHub*: `https://github.com/SantanderMetGroup/climate4R`.

Además de colaborar en el desarrollo de diferentes librerías de C4R —especialmente en `downscaleR` (Bedia et al., 2020), que permite aplicar fácilmente distintas técnicas de SD— la principal contribución de esta Tesis a dicho *framework* es `downscaleR.keras`, que proporciona una interfaz a `Keras` (Chollet et al., 2015), la librería de referencia hoy por hoy en el campo del *deep learning* que permite diseñar (casi) cualquier tipo de topología. Por tanto, `downscaleR.keras` permite incorporar sofisticadas arquitecturas de redes neuronales al conjunto de métodos tradicionales de SD disponibles en `downscaleR`. Se puede encontrar más información sobre `downscaleR.keras` en `https://github.com/SantanderMetGroup/downscaleR.keras`.

1.3.4 Premios

Esta Tesis ha obtenido el tercer premio en el *Doctoral consortium* celebrado en Granada (España) en 2018 por la Asociación Española de Inteligencia Artificial (AEPIA).

1.4 Líneas de Trabajo Futuro

Parte de los resultados de esta Tesis han abierto la puerta al desarrollo de nuevos estudios que constituyen una continuación natural de algunos de los análisis presentados en esta memoria.

Por ejemplo, una extensión que nos planteamos consistiría en evaluar la idoneidad de las CNNs para SD en el resto de dominios de CORDEX (más allá de Europa), con la idea de generar un *dataset* global de proyecciones de cambio climático de alta resolución basado en

DL. Además, también nos planteamos analizar la viabilidad de las CNNs para hacer SD a resoluciones espaciales más finas que los 0.5° considerados en esta Tesis (0.5°). Es bastante probable que estos estudios impliquen tener que implementar cambios en las topologías desarrolladas durante esta Tesis, lo cual permitirá adquirir un mayor conocimiento sobre los potenciales beneficios y limitaciones de las CNNs para aplicaciones de SD.

Además, hemos visto a lo largo de la Tesis como la falta de poder explicativo por parte de los predictores de larga escala para reproducir la variabilidad local —especialmente para la precipitación— resulta en una infraestimación de los extremos en las predicciones. Actualmente, la manera de atajar este problema es realizando un remuestreo aleatorio desde las distribuciones estimadas, lo que conlleva una pérdida de estructura espacio-temporal en los campos de alta resolución. Sin embargo, algunas topologías de DL como los *Variational Auto-Encoders* (VAE, Kingma and Welling (2013)) o los modelos generativos (GAN, Goodfellow et al. (2014)) podrían ser de utilidad para la generación de predicciones estocásticas espacialmente consistentes, por lo que serán investigadas en detalle próximamente.

Además, es crucial seguir avanzando en el estudio de la interpretabilidad de las redes neuronales —nótese que este aspecto ha sido parcialmente resuelto en esta Tesis— con el fin incrementar la confianza de la comunidad en este tipo de modelos, lo que podría promover su uso en distintas aplicaciones climáticas (más allá del SD).

Finalmente, en el marco de una colaboración internacional entre SMG y el *Centre National de Recherches Météorologiques* (CNRM) que se inició durante una de las estancias realizadas a lo largo de esta Tesis, contemplamos abrir una nueva línea de investigación centrada en el uso de DL para construir emuladores estadísticos. Esta línea tratará de dar respuestas a algunas de las preguntas clave que han sido formuladas en el *Flagship Pilot Study* (FPS) sobre convección de CORDEX[13]: 1) ¿Es un modelo de DL capaz de aprender eficazmente el sistema de ecuaciones que caracterizan a un RCM? 2) Una vez que se ajusta la red neuronal para una combinación GCM-RCM particular, ¿tiene sentido su utilización con el find de emular el mismo RCM pero acoplado a otros GCMs? 3) ¿Tiene sentido utilizar en estudios climáticos una red que ha sido ajustada teniendo en cuenta un escenario de emisión que difiere claramente del que se espera para el futuro? Los resultados preliminares que se obtuvieron durante la estancia en el CNRM para un caso de estudio concreto muestran que las CNNs desarrolladas en esta Tesis podrían suponer una alternativa real al uso de RCMs.

[13] Véase `https://www.hymex.org/cordexfps-convection/wiki/doku.php?id=home` para más detalles

CHAPTER 2

Context, Objectives and Structure

2.1 Context

Numerical models are the main tool used nowadays to study the evolution of climate at different time-scales, from a few days into the future (weather forecasting) to several decades (climate change projections). These models solve numerically the equations that describe the dynamics of the climate system (energy and mass conservation, Navier-Stokes equations, etc.) over a discretized three-dimensional space formed by gridboxes. Computational limitations constrain the temporal and spatial resolution these models can achieve. For instance, the Global Climate Models (GCMs) of the Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al. (2012)) present spatial resolutions in between 1° and 3°. These GCMs provide simulations for the entire globe based on different natural and anthropogenic forcings[1]. Despite current GCM have proved robust to reproduce key large-scale circulation patterns, they usually misrepresent important processes that occur at spatial scales smaller than the size of the model gridbox (e.g., convective precipitation). To produce high-resolution climate fields which explicitly resolve regional-to-local features is key to different socio-economic sectors such as energy, agriculture, hydrology and health, which are in need of these products for the development of their particular activities (e.g., high-resolution wind fields are needed for energy generation, accurate precipitation estimates are needed for local water management). Moreover, they have become crucial in the design of efficient adaption plans, to elaborate sustainable environmental policies, and to assess the possible impacts of climate change at the regional-to-local level. For

[1] The term *forcing* refers to any mechanism that has the potential to alter the Earth's climate, for instance changes in the energy balance of the planet due to a volcano's eruption or to an alteration in the concentration of greenhouse gases.

these reasons, several worldwide initiatives aimed at providing accurate and top-quality high-resolution climate products have recently emerged (e.g. the Coordinated Regional Downscaling EXperiment (CORDEX)).

To improve the usability of the GCM outputs for impact applications, two conceptually different approaches have been developed in the last decades: dynamical and statistical downscaling. On the one hand, Dynamical Downscaling (DD) is based on the use of numerical Regional Climate Models (RCMs) which simulate regional features of the climate at a higher resolution over a limited area, driven at the boundaries by the coarse-resolution GCM outputs (Rummukainen, 2010). On the other hand, Statistical Downscaling (SD) relies on statistical models/algorithms which link the coarse-resolution outputs from the GCMs (predictors) with the local observations (predictands) over the area of interest (Maraun and Widmann, 2018). Despite the relative merits and disadvantages of both approaches, they are seen as complementary rather than mutually exclusive (see Vrac et al. (2012); Casanueva et al. (2019) for illustrative examples). This Thesis focuses on a particular type of SD —which is drastically cheaper than DD in terms of computational resources— known as "Perfect-Prognosis" (PP).

The main particularity of PP is that the statistical models/algorithms linking the predictors and the predictand are built based on observations. For the predictors, reanalysis data[2] are typically considered, whilst for the predictand either high-resolution gridded data or station-scale records can be used. Once the statistical model is fitted in these "perfect" conditions, it can be subsequently applied to GCM predictors to derive the corresponding high-resolution products up to the end of the century, based on a number of possible socio-economic pathways. PP downscaling builds on three key assumptions with regards to the quality of the statistical model inferred and its transferability from the observational to the climate model space (Maraun and Widmann, 2018; Maraun et al., 2019). These assumptions state that the low-resolution predictors considered to build the statistical model 1) have to be informative enough for the local-scale, and 2) must be realistically simulated by the GCMs at a distributional level. Moreover, for a meaningful use under conditions not seen during the calibration phase (e.g., in climate change scenarios), 3) the statistical models should exhibit moderate extrapolation capabilities. To date, a variety of techniques which include (generalized) linear models (Gutiérrez et al., 2019), analogs (Hewitson and Crane, 1996), support vector machines (Tripathi et al., 2006), random forests (Hutengs and Vohland, 2016) and very shallow neural networks (Quesada-Chacón et al., 2021) have been employed to establish the link between the large

[2] A reanalysis is a global gridded dataset that combines observations with short-range weather forecasts through data assimilation, providing the most reliable representation of the actual state of the atmosphere at a given time. Reanalyses are widely used by the climate community in different applications, especially in regions with low density and/or quality of observational records.

and the local-scale. Despite their several merits, none of these methods has the capability to automatically handle high-dimensional input spaces without leading to overfitting, reason why the predictor space needs to undergo tedious and "human-guided" feature selection and/or reduction procedures before entering the statistical model for training/-calibration/fit (Gutiérrez et al., 2019). This typically implies a loss of information which can be relevant to explain the local variability of the target predictand.

In this context, the community has moved its attention to deep neural networks or Deep Learning (DL, Goodfellow et al. (2016)), with a special focus on convolutional-based[3] topologies. DL models have already beaten part of the existing battery of machine learning models —especially in computer vision and natural speech recognition— and have proved capable to 1) learn complex spatial patterns from data, and 2) automatically deal with high-dimensional input spaces without leading to overfitting. These properties make DL models a potentially powerful candidate for a range of climate-oriented applications, including statistical downscaling. Nevertheless, very few attempts —mostly focused on synthetic use-cases based on image-super-resolution architectures (see Vandal et al. (2018b) for an illustrative example)— tackle the use of DL for this problem to date. These first works show promising results and a good skill to reproduce local precipitation and/or temperature fields, but they 1) are directly not applicable for PP downscaling due to some methodological constraints, 2) lack from a rigorous analysis of their performance when applied in PP mode, and 3) miss an evaluation of their appropriateness for downscaling of GCM scenarios.

2.2 Objectives

Following from the previous considerations, this Thesis focuses on assessing the suitability of deep learning topologies, in particular Convolutional Neural Networks (CNNs), for the downscaling of GCMs over Europe under the PP paradigm. The following main objectives will be addressed:

1. To test the applicability and performance of CNNs for climate downscaling in "perfect" conditions —i.e. based on reanalysis predictors.— In this regard, one of the key features to examine will be their ability to deal with high-dimensional input spaces.

2. To evaluate the benefits and disadvantages of CNN multi-site topologies, as compared to the equivalent single-site versions. We will analyze for this aim the implicit regularization that occurs in multi-site architectures.

[3] A Convolutional Neural Network (CNN, LeCun et al. (1995)) is a specific type of neural-based topology which is commonly employed to learn non-linear patterns in datasets in which the underlying spatial structure is important (e.g., for images). CNNs are the main focus of this Thesis.

3. To gain understanding about the internal functioning of CNNs, which are typically seen as "black-box" models. To do this, we will focus on the study of the predictor-predictand link (i.e., influence of every input feature in the downscaling model outputs).

4. To study the suitability of CNNs to downscale future climate change scenarios. To do so, we will first evaluate the ability of CNNs to reproduce the observed climate based on the historical scenario of a GCM. Then, we will explore their potential for moderate and coherent extrapolation under one emission scenario, based on various GCMs.

2.3 Structure

To accomplish the above goals, the Thesis is structured in four main parts: Introduction (Part I), Data and Methods (Part II), Main Results (Part III) and Concluding Remarks (Part IV).

Part I is formed by two introductory chapters. Chapter 3 presents SD as a way to bridge the gap between the coarse resolution provided by the current GCMs and the regional-to-local information required by most of impact applications. We devote special interest to PP-SD —assumptions, techniques and limitations of this paradigm.— Chapter 4 introduces the principles of deep learning and the CNNs used along this Thesis. Moreover, a review of the literature that discusses the use of neural-networks in the context of climate SD is also presented.

Part II consists on a single chapter which describes the methodological framework followed in most of the analysis presented in this Thesis. In particular, Chapter 5 describes the datasets, SD methods and validation metrics used.

Part III is formed by two central chapters that present the main results of this Thesis, which are based on six manuscripts published in prestigious international journals and conference proceedings. On the one hand, Chapter 6 assesses the suitability of CCNs for PP-SD and is based on Baño-Medina et al. (2020). Moreover, this chapter also includes an analysis of multi-site CNN topologies, —based on Baño-Medina and Gutiérrez (2019),— and an exploratory study of the internal functioning of DL models for SD —based on Baño-Medina and Gutiérrez (2018) and Baño-Medina (2020).— On the other hand, Chapter 7 studies the appropriateness of CNNs to downscale GCM simulations (both in historical and future scenarios). This chapter is based on Baño-Medina et al. (2021b) —which focuses on a single GCM— and Baño-Medina et al. (2021a) —which uses CNNs to downscale an ensemble of CMIP5 GCMs.—

Finally, in Part IV, Chapter 8 summarizes the main conclusions of the Thesis, discusses future lines of work and enumerates the main the achievements accomplished.

Part I

Introduction

CHAPTER 3

Statistical Downscaling

3.1 Climate Modeling

Numerical models are the main tool used nowadays to study the evolution of climate at different time-scales, from a few days into the future (weather forecasting) to several decades (climate change projections). These models solve numerically the equations that describe the dynamics of the climate system (energy and mass conservation, Navier-Stokes equations, etc.) over a discretized three-dimensional space formed by gridboxes. Computational limitations constrain the temporal and spatial resolution these models can achieve –doubling the spatial resolution would require ten times more of computational power to complete a simulation in the same time.— For instance, the Global Climate Models (GCMs) included in the Coupled Model Intercomparison Project (CMIP[1]) Phase 5 (CMIP5, Taylor et al. (2012)) present spatial resolutions in between 1° and 3°. These GCMs provide worldwide simulations of a large number of meteorological variables based on different natural (e.g. eruption of volcanoes) and anthropogenic forcings. Amongst the latter, it is of particular importance the observed and estimated concentration of Green House Gases (GHG) for historical and future periods, respectively. For the future, an ensemble of possible emission trajectories were designed (mostly for CMIP5) based on different socio-economic indicators. These trajectories are known as Representative Concentration Pathways (RCP, Van Vuuren et al. (2011)) and are labelled according to the radiative pressure that is expected for the year 2100[2]

[1] CMIP represents the most ambitious initiative for global climate modeling and has designed a wide catalog of model intercomparison projects (MIPs) targeting different time horizons and socio-economic sectors.

[2] For instance, RCP8.5 is an extreme emission scenario which assumes that GHG will continue to be emitted to the atmosphere without restriction, reaching a radiative pressure of 8.5 W/m^2 by 2100. This

Currently, 62 different GCMs from 29 different modeling groups have participated in CMIP5 (an illustrative subset of models is listed in Table 3.1). This large variety is key to characterize the inherent uncertainty associated to climate change simulations, which is derived not only by changes in model formulation/parametrizations and by the use of different RCPs, but also from the chaotic and non-linear nature of the climate system itself.

Name	Institution	HR	Reference
CanESM2	Canadian Centre for Climate Modelling and Analysis	$(2.81^o/2.79^o)$	Christian et al. (2010)
CNRM-CM5	Centre National de Recherches Météorologiques and Centre Européen de Recherche et de Formation Avancée	$(1.4^o/1.4^o)$	Voldoire et al. (2013)
MPI-ESM-MR	Max-Planck Institut für Meteorologie	$(1.87^o/1.87^o)$	Müller et al. (2018)
MPI-ESM-LR	Max-Planck Institut für Meteorologie	$(1.87^o/1.87^o)$	Müller et al. (2018)
NorESM1-M	Norwegian Climate Center	$(2.5^o/1.9^o)$	Bentsen et al. (2013)
GFDL-ESM2M	National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory	$(2.5^o/2.02^o)$	Dunne et al. (2013)
EC-EARTH	European-wide consortium	$(1.12^o/1.12^o)$	Doblas Reyes et al. (2018)
IPSL-CM5A-MR	Institut Pierre Simon Laplace Climate Modelling Center	$(2.5^o/1.27^o)$	Dufresne et al. (2013)

Table 3.1: An illustrative subset of the GCMs included in CMIP5, including the running institution, the lon/lat horizontal resolution (HR) and the reference manuscript for each model.

Nowadays, both public and private sectors elaborate mid- and long-term policies according to the available climate change scenarios provided by the different GCMs. For instance, CMIP simulations are analyzed in the periodic Assessment Reports (AR) elaborated by the Intergovernmental Panel on Climate Change (IPCC[3]), which ultimately are used to support national adaptation plans in many countries. Nevertheless, policy makers and stakeholders are often in need of high-resolution information which current GCMs are not able to provide (despite model resolution has increased considerably during the last decades due to the development of faster and more efficient computational infrastructures). These finer scale products are critical —especially in regions vulnerable to climate change— for efficient design and management of energy plants, irrigation schemes, civil and transport infrastructures, etc.

It is important to note that sub-grid processes which often drive the regional-to-local climate are misrepresented in GCMs. This can be partially alleviated by introducing parametrization schemes as part of the model formulation (McFarlane, 2011). Parametrizations are complex empirical functions which allow to adjust the variable of interest (e.g., precipitation) returned by the model towards more realistic values. However, many works have reported substantial model biases (as compared to observational records)

is the scenario considered in this Thesis.

[3] IPCC is the panel of worldwide experts founded by the United Nations (UN) in 1988 which aims to ease the communication between the scientific community and policymakers regarding the physical basis and potential impacts of climate change.

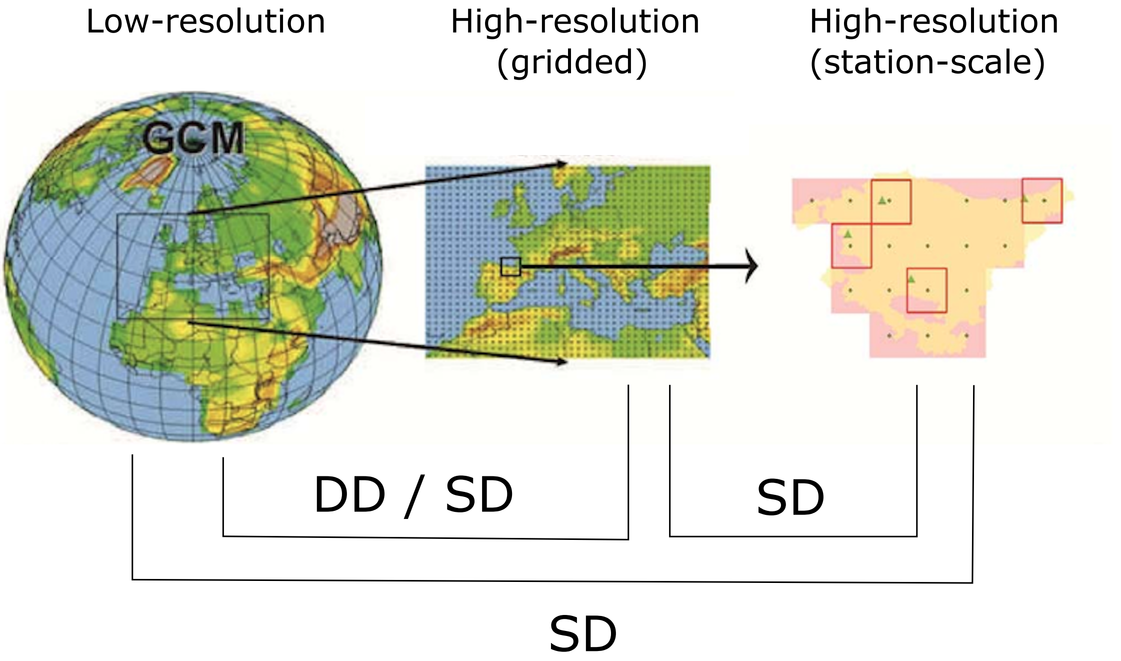


Figure 3.1: Diagram illustrating the spatial resolution of typical climate products; from the coarse outputs of a GCM (left) to the high-resolution gridded or station-scale data required by most of practical applications (middle and right, respectively). DD is used to bridge the gap between low- and high-resolution gridded products whilst SD can be employed to pass from the coarse to the local scale based either on gridded data or in station-wise measurements.

for particular variables and/or locations regardless of the use of tuned parametrizations (see, e.g., Kotlarski et al., 2014). For instance, this is the case of mountainous regions in which precipitation is often triggered by convective events related to the local orography, or coast borders where the land-sea interface determines the appearance of pronounced temperature gradients. Moreover, parametrizations do not modify the original resolution of the model towards finer grids, and therefore they are not enough to bridge the scale gap that hinders the use of GCMs in real-life problems (Demory et al., 2020).

In this context, downscaling emerged as a tool to improve the usability of the coarse resolution outputs provided by the GCMs (see Figure 3.1). In particular, two conceptually different approaches have been developed during the last decades to this aim: dynamical downscaling (DD) and statistical downscaling (SD).

DD is based on the use of Regional Climate Models (RCMs) which numerically solve the equations describing the dynamics of the climate system over a relatively small area (e.g., Europe), driven by GCM outputs at the boundaries (Rummukainen, 2010). As a consequence of running on finer grids, RCMs are typically better than GCMs at reproducing the observed climate in regions where it is mostly determined by local phenomena. Moreover, studies also point to an improvement of DD over GCMs in regions with a low dependence on local mechanisms (Sørland et al., 2018). Nevertheless, RCMs tend to inherit some of the limitations from the driving GCMs such as biases in the atmospheric fields (Christensen et al., 2008) and require powerful computational infrastructures to run.

Differently, SD establishes empirical relationships between the large-scale and a set of observational records at the local-scale (Maraun and Widmann, 2018). As a result of being calibrated directly with observations, SD models are expected to present very small biases (as compared to observations) over the training period. Moreover, SD is drastically cheaper than DD in terms of computational resources and requires much shorter times —e.g., the calibration of a typical SD model can take minutes or hours, depending on the particular technique used and the extension of the area of study.—- However, SD needs high quality, long enough observational records to establish robust links between the large- and the local- scale, which limits its applicability in many regions of the world. Also, SD does not take into account the physical principles linking the large- and the local-scale, which typically results in worse spatio-temporal coherence than RCMs. Furthermore, SD models can not capture small-scale dynamical changes, since they are not reflected in the large-scale predictors (Vrac et al., 2007b). Finally, the key limitation of SD is the stationary assumption, as the statistical models learnt over a given period (e.g. the recent past) are assumed to remain valid for other periods (e.g. the end of the century) in which the synoptic patterns may differ. This is especially relevant for the SD of long-term climate change scenarios (Gutiérrez et al., 2013).

Statistical Downscaling (SD)	**Dynamical Downscaling (DD)**
Unbiased predictions since it is calibrated directly with observations	Reduces the bias of GCMs but generates its own bias
Computationally cheap (∼ minutes/hours)	Computationally expensive and requires sophisticated supercomputers (∼ months)
It is not based on physical principles, and spatio-temporal consistency is not granted	It is based on physical principles, and therefore ensures inter-variable and spatio-temporal correlations

Table 3.2: A comparative summary between statistical and dynamical downscaling.

Several intercomparison studies between DD and SD have been carried out (see, e.g., Murphy, 1999; Haylock et al., 2006; Schmidli et al., 2007; Vaittinada Ayar et al., 2016), illustrating the mentioned issues (see Table 3.2 for a comparative summary). Nowadays, DD and SD are seen as complementary rather than mutually exclusive (see Vrac et al.

(2012) for an example). For instance, SD is often used to remove the systematic biases of RCMs (Casanueva et al., 2019). The next section is devoted to explain the different approaches and techniques available for SD, the focus of this Thesis.

3.2 *Statistical Downscaling*

This section provides a brief overview of the three different approaches available for SD: Perfect-Prognosis (PP), Model Output Statistics (MOS) and Weather Generators (WG). The interested reader is referred to Maraun and Widmann (2018) for further details.

PP-SD is based on the use of transfer functions or algorithms which allow to establish empirical links between a low- and a high-resolution dataset of observations. For the former, reanalyses are typically used, whilst for the latter, either high-resolution gridded data or station-scale records can be employed. Once the statistical model is fitted in these "perfect" conditions, it can be applied to derive high-resolution downscaled fields of the variable of interest using as inputs GCM predictors under different scenarios (e.g. historical and RCP simulations). Typical PP techniques include linear (Huth, 2002; Chandler and Wheater, 2002; Fealy and Sweeney, 2007; Hertig et al., 2013; Beecham et al., 2014) and non-linear (Vrac et al., 2007a; Huth et al., 2008; Chen et al., 2010; Quesada-Chacón et al., 2021; Olmo and Bettolli, 2021) regression methods, and analogs (Zorita and Von Storch, 1999; Walton et al., 2020).

MOS-SD aims to build statistical relationships which link the variable of interest (e.g. precipitation) simulated by a particular climate model (either a GCM or a RCM) with the corresponding observational record at a given site. This is often done by mapping some order-moments (e.g. the mean, the variance, and/or different percentiles) of the simulated distribution to the observed one. Once fitted over a determined period (e.g. the recent past), this link can be subsequently used to correct the outputs of the same GCM/RCM for other time periods (e.g. future decades). In climate change studies, MOS basically reduces to bias adjustment techniques, which have become increasingly popular during the last years. These range from simple additive or multiplicative scaling corrections (Durman et al., 2001; Iizumi et al., 2011; Casanueva et al., 2013) to more distributional-oriented ones (Piani et al., 2010; Lafon et al., 2013; Turco et al., 2017; Fauzi et al., 2020), such as quantile mapping (Panofsky et al., 1958). The latter performs a quantile-to-quantile adjustment of the probability functions of the variable of interest (see, e.g., Maraun et al., 2010; Teutschbein and Seibert, 2012; Maraun et al., 2017a; Manzanas et al., 2020b). MOS-SD essentially refines the climatological detail of the fields, but do not add significant value since the spatial and temporal structure of the downscaled series are largely inherited from the climate model. Therefore, MOS methods are most commonly used to post-process

RCM outputs (e.g. to correct their systematic biases) or to bridge the scale-gap when the resolution of the climate model is close to the target resolution.

Finally, WG (see Wilks and Wilby (1999) and Ailliot et al. (2015) for a review) is aimed at learning the distributional moments of the variable of interest at a given site from an observational record. For instance, it is common to estimate the parameters of an exponential function to model the observed wet-day distribution. Afterwards, a synthetic time-series which preserves the statistics of the observed climate can be generated by sampling from the estimated probability function. To downscale future projections, the calibrated parameters are perturbed —in a manner consistent with the climate change projected by the GCM/RCM of interest— to produce local series under different emission scenarios (Kilsby et al., 2007; Keller et al., 2017; Vesely et al., 2019). WGs are commonly built on a monthly basis (i.e., a WG per month, see Wilks (2010) for an illustrative case-study), or conditioned to certain atmospheric patterns. For instance, Bardossy and Plate (1992) and Fowler et al. (2000) have produced synthetic series of precipitation based on different weather types. It is of special interest to this Thesis the hybrid PP-WG approach in which large-scale reanalysis predictors are linked —based on some transfer function,— to the parameters of a selected probability function, describing the variable of interest at a given site. For instance, Williams (1998) and Cannon (2008) deploy linear and neural network models, respectively, to learn daily Bernoulli-Gamma distributions describing local precipitation, which are conditioned to the large-scale atmospheric situation (further details on this topic are given in section 4.4). This approach permits to overcome the miss-representation of the extremes of some PP-SD downscaled series —which appears when predictors lack from a sufficient informative power to explain the local variability,— by sampling from the estimated conditional distributions.

This Thesis focuses on PP downscaling. Therefore, the rest of the chapter is devoted to a deeper explanation of this particular approach.

3.3 Perfect-Prognosis Statistical Downscaling

As explained, SD is based on empirical functions that link the large-scale atmospheric situation to the local scale of the variable of interest (typically daily temperature or precipitation). Under the PP approach (see Figure 3.2), this link is inferred building on a particular statistical method —e.g., analogs, generalized linear models, etc.— considering observational datasets for both the predictors (reanalysis) and the predictand (e.g., station-scale records). The so-constructed model can be applied then to produce local downscaled projections using as inputs the large-scale variables simulated by the GCMs under different scenarios (for instance, future RCPs).

The following sections dig into the details of different methodological aspects that are key for PP-SD.

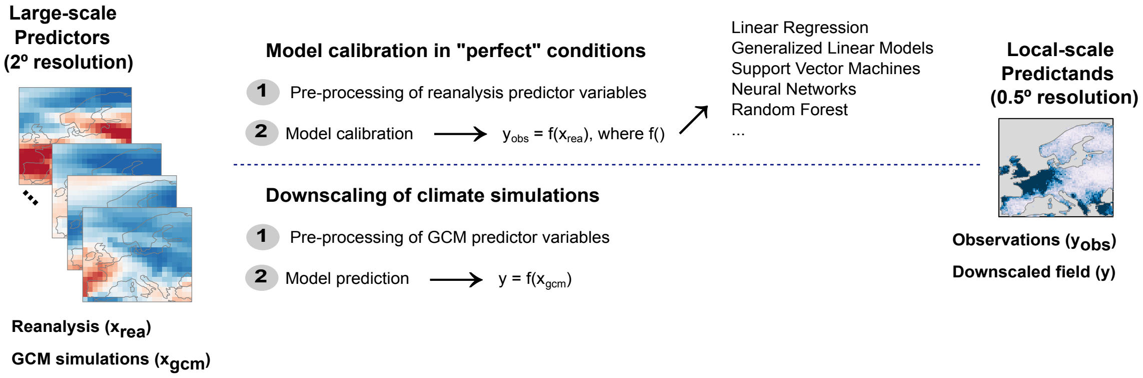

Large-scale Predictors (2º resolution)
Reanalysis (x_{rea})
GCM simulations (x_{gcm})
Model calibration in "perfect" conditions
1 Pre-processing of reanalysis predictor variables
2 Model calibration → y_{obs} = $f(x_{rea})$, where f()
Linear Regression
Generalized Linear Models
Support Vector Machines
Neural Networks
Random Forest
...
Downscaling of climate simulations
1 Pre-processing of GCM predictor variables
2 Model prediction → y = $f(x_{gcm})$
Local-scale Predictands (0.5º resolution)
Observations (y_{obs})
Downscaled field (y)

Figure 3.2: Diagram illustrating the different phases of a typical PP-SD experiment. First, large-scale reanalysis predictors (at a spatial resolution of 2° in this example), x_{rea}, are linked to a high-resolution gridded observational dataset, y_{obs} (at 0.5°) through an empirical functions, $f()$, which is learnt based on a given SD method (see section 3.3.3 for details). Subsequently, $f()$ can be applied to derive the high-resolution downscaled field of the variable of interest, y, using as inputs GCM predictor variables, x_{gcm}. Note that some processing of the predictors is usually performed (for instance, standardization is normally applied to avoid scale artifacts in the downscaled results).

3.3.1 Cross-Validation

When assessing the performance of any SD technique it is crucial to rely on a proper cross-validation scheme; otherwise, misleading conclusions may be obtained. Hold-out is the simplest approach to do so. It consists of dividing the entire observational datasets into independent train and test sets. Whilst the former is used for model calibration, the latter is used to assess the generalization capacity of the SD model, ensuring it does not incur into overfitting. One of the key advantages of hold-out cross-validation is that it allows for evaluating whether or not a given SD method presents "moderate" extrapolation capabilities. This can be done by wisely selecting train and test periods with different climatological properties. This kind of analysis is crucial if the SD model is planned to be used for downscaling of climate change scenarios, since it should be able to work reasonably well under new conditions which may have not occurred during the calibration phase.

Hold-out has evolved to a more sophisticated splitting of the data, called k-fold cross-validation (Markatou et al., 2005), which splits the data —either random or chronologically— in K folds. The SD model is tested on each of the K folds separately, using the remaining $K - 1$ folds for model calibration. By doing so, a complete prediction covering the full period can be recovered by appropriately concatenating the results from the K models.

Given the importance of the temporal structure in climate data, chronological partitioning is often used (Gutiérrez et al., 2019) for SD.

3.3.2 Key Assumptions

In a climate change downscaling context, the PP approach needs to fulfill the following key assumptions (see, e.g., Maraun and Widmann, 2018; Maraun et al., 2019):

- The predictors chosen have to be informative, i.e, they should explain a large fraction of the temporal variability of the predictand of interest. Moreover, they should carry the climate change signal —typically, this is ensured by including some temperature and/or humidity variables in the predictor set.— Otherwise, the downscaled projections might not reflect the changes that are expected in future climate conditions.

- The predictors chosen have to be realistically simulated by GCMs in both recent past (historical) and future (RCP) simulations. A minimum requirement in this sense would be to assure that GCM predictors present no biases with respect to their reanalysis counterparts for the calibration period; otherwise, the high-resolution products would inherit the biases from the GCM (Wilby and Wigley, 2000). To ensure compatibility among datasets, harmonization and/or standardization procedures can be applied prior to downscaling. For instance, Vrac and Ayar (2016) performed a bias correction of the monthly mean of the GCM predictors taking the reanalysis as reference, for a better fit of the seasonal cycle.

- The statistical models have to be flexible enough to model the complex interactions between the large- and the local-scale. On the one hand, very simple models (e.g., linear regression) might fail to reproduce the non-linear mechanisms that link predictors and predictand. On the other hand, overparameterized models may suffer from overfitting, which in turn might reduce their predictive skill in out-of-training conditions. Regardless of the model complexity, any technique should be able to show some extrapolation capability when driven by predictor spaces which had not been observed during model calibration.

Due to importance of these hypotheses to the PP-SD approach, these are analyzed along the Thesis, especially in Chapters 5, 6 and 7.

3.3.3 Techniques

The existing battery of statistical and machine learning techniques for PP downscaling is extensive, ranging from simple (generalized) linear models (Gutiérrez et al., 2019), quantile regression (Koenker and Hallock, 2001), self-organizing maps (Hewitson and Crane, 2002;

Hope, 2006), neural networks (Schoof and Pryor, 2001), support vector machines (Tripathi et al., 2006), random forests (Hutengs and Vohland, 2016) and analogs (Hewitson and Crane, 1996), among others.

Several intercomparison studies have demonstrated that there is not a single method which clearly outperform the others in terms of reproduction of the local variability and spatio-temporal consistency (Wilby et al., 1998; Chen et al., 2010; Sachindra et al., 2018; Yang et al., 2018; Gutiérrez et al., 2019). Due to their simplicity and general good performance, Generalized Linear Models (GLM) and analogs tend to be the preferred option among the downscaling community (Brandsma and Buishand, 1997; Chandler and Wheater, 2002; Abaurrea and Asín, 2005; Fealy and Sweeney, 2007; Hertig et al., 2013). We devote the rest of the section to explain the particularities of GLMs, which are considered the benchmark against the Convolutional Neural Networks (CNNs) developed in this Thesis are compared.

GLMs (Nelder and Wedderburn, 1972) establish an empirical link, $g^{-1}()$, parameterized by a set of coefficients ω which map a set of explanatory variables x to the expected value of the target variable, $E(y)$, which is assumed to follow a particular probability distribution belonging to the exponential family:

$$E(y) = g^{-1}(\omega x) \tag{3.1}$$

Well-known types of regression can be deduced from Eq.3.1. For instance, multiple linear regression is equivalent to assume a Gaussian distribution with an identity link in the GLM formulation. Logistic regression is also a special case of Eq.3.1, where a GLM with Bernoulli error distribution is trained using the *logit* canonical link. These two cases appear often in downscaling studies, principally to derive local temperature and precipitation occurrence, respectively (Gutiérrez et al., 2019). Note that even though they are linear by definition, GLMs can achieve a limited degree of non-linearity by considering sophisticated (e.g. logarithmic) links.

GLMs can be used either as a pure PP or as a hybrid PP-WG downscaling technique (Gutiérrez et al., 2019). In PP mode, GLMs predict the expectance of the conditional distribution being modeled. Differently, in a PP-WG context, the predictions are sampled out from the modeled distributions, which allows to increase the variance of the downscaled temporal series (see, e.g., Manzanas et al., 2020b).

3.3.4 Model Setup, Limitations and Challenges

PP-SD presents several challenges mostly related to the model setup. In particular, regarding the predictor set, one should ideally select a relevant and unredundant set of variables —confined within a meaningful geographical domain— which largely explain the local variability of the predictand of interest. For instance, whereas surface temperature is known to be highly determined by the large-scale atmospheric situation, precipitation is in many cases triggered by local processes (e.g., convection) which are not represented in the coarse-resolution predictors. Moreover, in climate change studies, circulation predictors (e.g., sea level pressure) should always be accompanied by thermodynamic variables such as temperature and/or humidity which account for the proper climate change signal (Maraun and Widmann, 2018). As a result, the predictor/domain selection is a complex task which is typically undertaken based on an exhaustive screening which should focus on different validation metrics (see, e.g., Gutiérrez et al., 2013; San-Martín et al., 2017). In addition, Manzanas et al. (2020a) proved —on a use-case over Malawi,— that the variability in the climate projections due to predictor selection is also an important source of uncertainty.

Previous studies suggest that temperature at 850 hPa (Huth, 1999, 2002; Maraun and Widmann, 2018; Gutiérrez et al., 2013; Gutiérrez et al., 2019) and geopotential height (Huth, 1999) are the most relevant predictors for surface temperature, and are often present in models aimed at downscaling this variable both from reanalysis and GCMs (Huth, 1999, 2002, 2004).

For precipitation, the choice of informative predictors is more difficult (Maraun and Widmann, 2018). Precipitation occurs when moist air ascends vertically leading to saturation and the formation of droplets. The presence or not of particles where to coalescence the droplet, local orography, convergence of air masses or local diabatic heating are some of the physical processes responsible for precipitation, which are only partially described by the synoptic-scale. As a result, fully informative predictors for this variable are typically lacking. Anyhow, there is general agreement regarding the usefulness as predictor of relative humidity, geopotential height and temperature at different vertical levels (Schmidli et al., 2007; San-Martín et al., 2017; Yang et al., 2018; Gutiérrez et al., 2019; Soares et al., 2019). Still, the observed variability of local precipitation is typically underestimated in the downscaled time series by deterministic PP methods (Enke and Spegat, 1997). PP-WG methods aim to overcome this issue by sampling out from the distributions conditioned to the large-scale atmospheric situation. However, this stochastic simulation procedure damages the spatio-temporal structure of the downscaled precipitation fields, making necessary to find an optimum trade-off between spatio-temporal representativeness and the adequate reproduction of extremes.

Despite this "a priori" knowledge, the choice of predictors is known to constitute an important source of uncertainty in SD for the generation of climate change scenarios (Huth, 2004; Manzanas et al., 2020b). In addition, due to the incapacity of state-of-the-art SD methods to efficiently handle high-dimensional input spaces, the available predictors need to undergo restrictive feature selection and/or reduction techniques in order to avoid overfitting. This implies a loss of information which ends by damaging the predictive skill of the resulting SD models. In this context, Deep Learning (DL) emerges as a powerful alternative for PP-SD, capable of handling high-dimensional input spaces in a meaningful way. For instance, the CNNs developed in this Thesis have the potential to extract relevant information from the complex spatio-temporal patterns present in the data, helping thus to overcome some of the explained limitations of traditional PP-SD techniques.

Moreover, any PP-SD technique has to deal with the "stationarity" assumption, which entails that the predictor-predictand links inferred in present climate conditions will remain valid under future climate change. In this regard, some of the strategies considered to-date to assess the extrapolation capability include the use of test periods with different climatic characteristics than the one used for calibration —for instance, by selecting the warmest years in a "perfect" conditions environment (see, e.g. Gutiérrez et al., 2013)— and the comparison of the downscaled future projections with those directly obtained from either GCM or RCM simulations, which are considered as "pseudo-observations" or "pseudo-reality"(Vrac et al., 2007b). In this Thesis we rely on these two approaches to dig into the "stationarity" assumption of the SD models proposed.

CHAPTER 4

Deep Learning

4.1 A Brief Historical Overview

The history of Neural Networks (NN) dates back to the 40's when the first neuron model was developed to differentiate between two categories (McCulloch and Pitts, 1943). In the next decade, the perceptron model (Rosenblatt, 1958) extended that first naive neuron model by implementing an optimization algorithm that would be the predecessor of the current backpropagation (Rumelhart et al., 1986). The perceptron was designed to learn a set of coefficients w that leverage a number of explanatory variables, x, allowing to estimate the outcome for a particular classification task, y —which was then casted to boolean based on a pre-defined threshold (step function).— Perceptrons evolved to *neurons*, which passed the resulting affine transformation to a non-linear activation function $f()$ —e.g., sigmoidal.— Basically, NNs consist of an arrangement of neurons in (hidden) layers to perform operations in a sequential and hierarchical manner (see Figure 4.1), resulting into a non-linear mapping between a set of explanatory (input layer) and response (output layer) variables. NNs were found to be useful for a variety of applications, including sequential modeling (Hochreiter and Schmidhuber (1997), Bengio et al. (1994)), distributed representation (Hinton et al., 1990) and computational neuroscience (Touretzky and Hinton, 1985). However, in the mid 90s, due to the lack of large data records and suitable computational infrastructures, NNs were found to easily incur in overfitting or become intractable. As a result, the use of NNs was limited to very shallow topologies/architectures (i.e., one or two hidden layers). These simple networks lacked from the ability to learn complex data patterns, and in parallel, kernel machines (Gunn et al., 1998) and graphical models (Koller and Friedman, 2009) outperformed shallow NNs for important tasks. As a consequence, investment and research towards neural-based topologies decreased considerably, remaining only a few isolated research centres truly devoted to this matter. In 2006, a special type of NN called deep belief network was efficiently trained (Hinton et al., 2006) followed by other neural configurations (Bengio et al. (2007), Ranzato et al. (2007)) and the field started to re-emerge. The definitive breakthrough came in 2012 when a deep convolutional network won the most challenging competition in object recognition (Krizhevsky et al., 2012), consolidating the back to scene of NNs.

The success behind the resurrection of NNs is closely related to different technological developments. On the one hand, the digitalization of society has facilitated data access and curation, eases the training of the network (the estimation of parameters becomes more robust). On the other hand, faster Computational Process Units (CPUs) and the adoption of Graphical Processing Units (GPUs) — commonly used for image processing— have improved the ability to store and process vast amounts of data and have allowed the development of rapid optimization procedures. Not only the hardware, but also new software has arrived in the form of *R* or *Python* libraries —e.g., *Theano* (Bergstra et al., 2010), *TensorFlow* (Abadi et al., 2016), *Keras* (Chollet et al., 2015) or *PyTorch* (Collobert et al., 2011), among others— which facilitate the design and coding of neural models, and also ease the access to distributed computing. Altogether, these new advances allowed that many NNs whose optimization had been impossible in the past became feasible, pushing towards more and more complex topologies. In particular, based on biological neural networks, the research community put the focus on the depth of the net, redefining thus the field to *deep learning* (DL). In addition to the above mentioned technical developments, advances in the topological elements of the network were also needed to successfully train these DL models. In particular, the appearance of Rectified Linear units (ReLu, Nair and Hinton (2010)) —which overcame the vanishing gradient problem (Hochreiter, 1998),— stochastic gradient descent (Bottou et al., 2018) and the development of sophisticated learning algorithms (e.g., Adam optimizer or RMSprop optimizer), among others — e.g., weights initialization (Glorot and Bengio, 2010), regularization techniques (Srivastava et al., 2014), normalization layers (Ioffe and Szegedy, 2015),— were responsible for the success of deep networks. For these reasons, DL refers not only to the growth in depth of the networks, but also to the technological developments that have made models with many layers tractable. DL has recently emerged as a powerful alternative to learn complex non-linear patterns from vast amounts of data (Goodfellow et al., 2016) and has outperformed other machine learning methods in a variety of commercial —e.g., image recognition (Krizhevsky et al., 2012)— and non-commercial —e.g., cancer detection (Amin et al., 2018)— applications.

But why have neural networks surpassed the wide variety of machine learning tech-

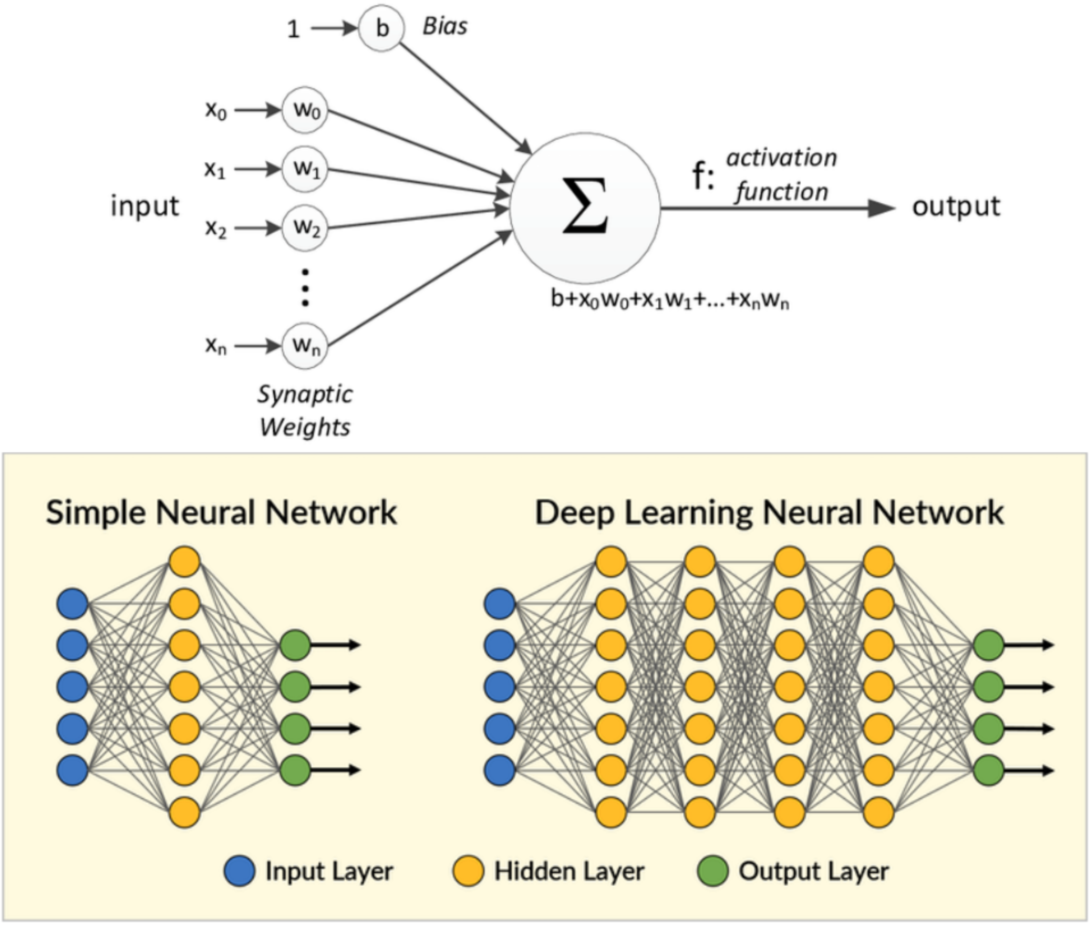


Figure 4.1: Diagram of a neuron model (top), together with a simple (bottom left) and deep (bottom right) neural network topologies. See sections 4.2.1 and 4.2.2 for details on the mathematical formulation of these particular topologies.

niques? Leaving aside bio-inspired reasons, similar to classical machine learning techniques, neural networks basically transform the input space into a hidden/latent one by performing some mathematical operation. However, unlike in other machine learning techniques in which the generation of this space relies on the kernel trick (e.g., support vector machines (Gunn et al., 1998)) or it is simply randomly produced (e.g., extreme learning machines (Huang et al., 2006)), NNs optimize this latent representation, which is shaped by the hidden layer dimensions. The key challenge in the design of a neural network resides in the choice of a suitable topology which allows to learn complex data patterns without incurring in overfitting. In this regard, the width (number of neurons) and the depth (number of hidden layers) of the network are key since they control its capacity to infer the potential non-linearities linking input and output variables. Indeed, if there is a sufficient number of neurons, a neural network with a single hidden layer could approximate in practice any function (Hornik (1991), Cybenko (1989)). Nevertheless, this particular architecture is usually not able to learn a function which generalizes to data samples not seen during model calibration. Differently, deep topologies have demonstrated better generalization properties than shallow ones due to their ability to extract useful knowledge from hierarchical structures that appear in the input features. Still, due to the large amount of parameters present in deep neural networks, some form of regularization is often needed. Despite the success of DL in many disciplines, the truth is that neural networks are often called "black-box" models since their complex topologies difficult the analysis of the function begin approximated. The interpretability of DL models is challenging and constitutes nowadays an active area of research which covers different approaches such as geometrical interpretations related to Riemannian spaces or visualization of the hidden structures (Daniely et al. (2016), Yosinski et al. (2015), Hauser and Ray (2017), Zintgraf et al. (2017)). This lack of interpretability is in fact one of the major drawbacks of NN in general (and DL in particular) to date and has hindered the adoption of these models in more scientific disciplines, including the climate science.

4.2 Principles of Neural Networks

This section describes the basics of neural networks. We start by introducing the perceptron model in section 4.2.1, followed by a brief explanation of the multi-layer perceptron in section 4.2.2. Convolutional networks and multi-task architectures —which are essential for the models developed in this Thesis,— are introduced in sections 4.2.3 and 4.2.4, respectively. In section 4.2.5, we describe the optimization procedure followed in NNs. To facilitate the reader the understanding of these topologies we accompany the explanations with diagrams of illustrative meteorological use-cases.

4.2.1 Neuron Model

Given a set of N input pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$ —where $x_j \in \Re^a$ and $y_j \in \Re^1$ are the explanatory and response variables, respectively,— an artificial neuron (see Figure 4.2) learns an affine transformation which is activated by a non-linear function (e.g., sigmoidal), such that each instance j of the variable y can be described as:

$$y_j = f\left(\sum_{i=1}^{a} x_{ji} w_i + b_i\right) \tag{4.1}$$

The election of the activation function, $f()$, depends on 1) the particular task being addressed —e.g. classification or regression— and 2) the place in which the neuron is

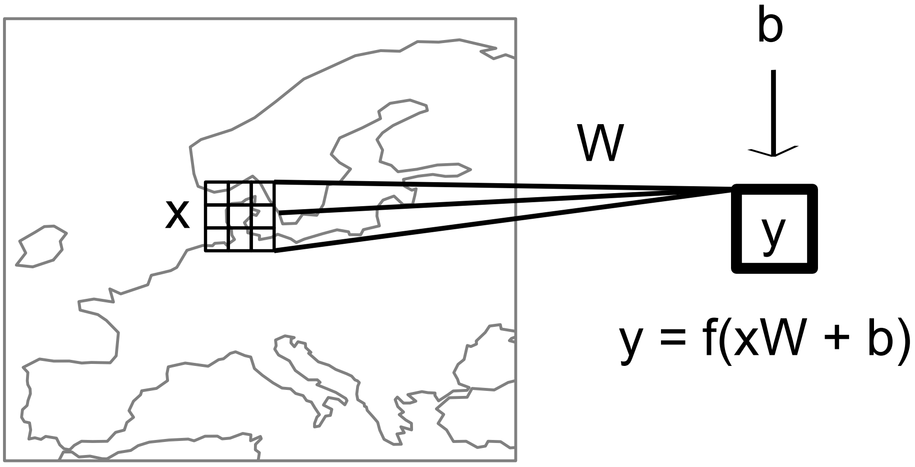


Figure 4.2: Diagram showing a neuron model based for a meteorological application. For a given sample, a row input vector of features, $x \in \Re^9$, is multiplied by a column vector of coefficients, $W \in \Re^9$, plus an independent term, b. The neuron, y, activates this affine transformation according to a given function $f()$. The set of coefficients $\omega = \{W, b\}$ are learned thanks to the gradient descent method and the backpropagation algorithm.

disposed within the network (see section 4.2.2). Sigmoidal functions are usually chosen for classification tasks since the output is bounded between 0 and 1, whereas the identity or exponential functions are preferred for regression problems. In fact, note that when the activation function is the identity, equation 4.1 is equivalent to a linear regression (or multiple linear regression if $a > 1$). Another special case occurs when the activation function is a Heaviside step function which reduces the model to its predecessor: the perceptron (Rosenblatt, 1958).

4.2.2 Dense Neural Networks

The degree of non-linearity that a single neuron can achieve is very limited regardless the activation function considered. To increase the model's flexibility, neurons are arranged in complex structures that define the topology of the network. The most common topology encountered in the literature is the feedforward neural network, in which neurons are disposed in layers connected in a sequential manner. Dense neural networks arrange the neurons into three or more layers (see Figure 4.3), with each particular neuron outputting an activated affine transformation based on all the neurons contained in the previous layer (i.e., fully-connected). This architecture dates back to the 70s and was developed as a natural extension of the perceptron model, reason why it is commonly known as Multi-Layer Perceptron (MLP).

In brief, a neural network of L layers learns a function $f^w()$ which is parameterised

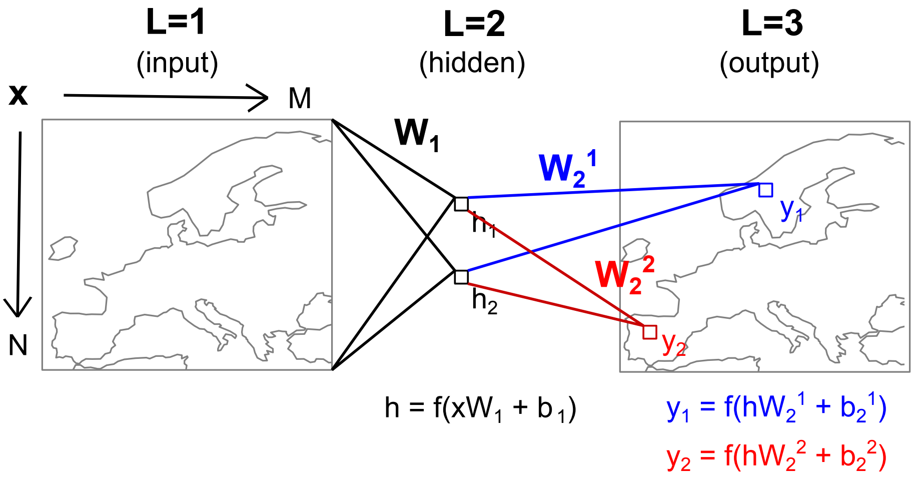


Figure 4.3: Diagram showing a 1-hidden layer dense NN in a meteorological application. In this case, the hidden space consists of 2 neurons which are the result of an affine transformation between an input vector of features, $x \in \Re^{MxN}$, and a matrix of coefficients $W_1 \in \Re^{(MxN)x2}$. Typically sigmoidal or ReLu activation functions are used as nonlinear operators for the hidden space. Finally, the desired variable, $y1$, is obtained based on the same procedure but relying on the 2-valued vector W_2^1, since there are only 2 neurons in the hidden layer. Multi-site topologies would consist of optimizing more than 1 site simultaneously, included as an additional neuron, y_2, in the output layer (see section 4.2.4 for a detailed explanation of these architectures). In this last case, the network would optimize the following set of coefficients: $\omega = \{W_1, W_2^1, W_2^2, B_1, b_2^1, b_2^2\}$.

by a set of coefficients, $\omega = \{W_1, W_2, ..., W_{L-1}\}$, mapping the input space to the output space. Hence,

$$y = f^{\omega}(x) \tag{4.2}$$

As mentioned in section 4.1, the width and depth of the network control the degree of non-linearity that can be achieved. The number of coefficients/parameters/weights that need to be adjusted by the network grows exponentially with both width and depth. Moreover, there is a lack of “a priori” knowledge with regards to the optimum size of the network in most applications. As a result, NNs are often grown in excess, which can easily lead to overfitting. Several approaches have been reported in the literature to avoid this problem, for instance data augmentation (Jaitly and Hinton, 2013), norm penalties (Tibshirani, 1996), bagging (Breiman, 1996), dropout (Srivastava et al., 2014), injection of artificial noise (Sietsma and Dow, 1991), multi-task learning (Ruder, 2017), early-stopping (Bishop, 1995) or parameter sharing (LeCun et al., 1995). Convolutional operations, which are the basis of the NNs developed in this Thesis, are a form of parameter sharing.

4.2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are a specific type of feedforward NNs in which adjacent layers are linked through a convolutional operation which allows for learning complex spatial structures (i.e. patterns) present in the data. This is due to the special configuration of the network's parameters, which are arranged in kernels that convolute over the dimensions of the input layer (typically 2-dimensional maps). CNNs (LeCun et al., 1995) were first introduced in the 90s for computer vision applications.

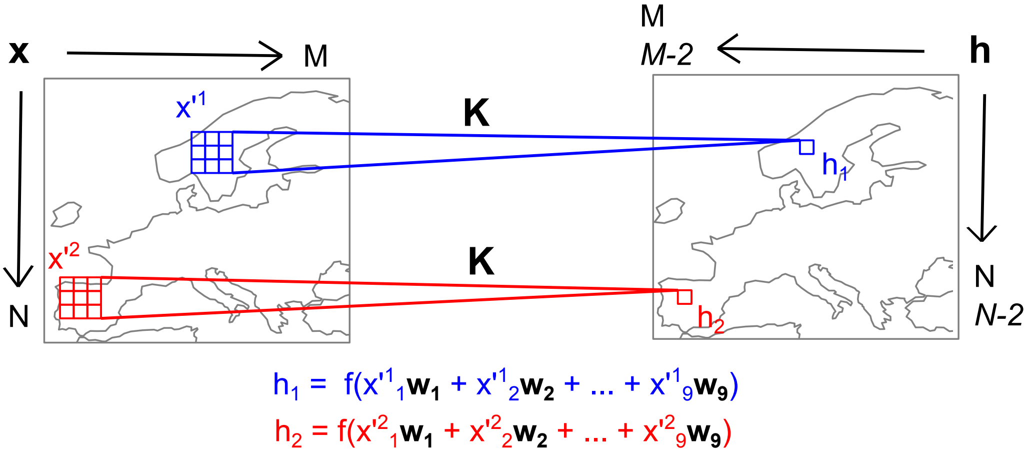

x
M
N
x'1
x'2
K
K
M
M-2
h
N
N-2
h1
h2
h1 = f(x'11w1 + x'12w2 + ... + x'19w9)
h2 = f(x'21w1 + x'22w2 + ... + x'29w9)

Figure 4.4: Diagram showing a convolutional layer designed in a meteorological application. The 2D input features, $x \in \Re^{MxN}$, are convoluted by a 2D kernel, $K \in \Re^{3x3x1}$. The affine transformation is based on patches of the input space, $x' \in \Re^{3x3x1}$, and the described kernel, $K = \{w_1, w_2, ..., w_9\}$, which is identical independently of the patch (i.e., parameter sharing). The result is a 2D filter map, $h \in \Re^{MxN}$ or $h \in \Re^{(M-2)x(N-2)}$ depending on whether padding is applied or not, respectively. The number of filter maps depends on the number of kernels (1 in this example).

Figure 4.4 illustrates the convolution operation in a meteorological application. Given a set of input features, $x \in \Re^{NxMxC}$ (e.g., latitude-longitude fields of different atmospheric variables) and a kernel, $K \in \Re^{K_1xK_2xC}$, convolution provides a scalar product of these 2 real-valued functions across the spatial dimensions. The neurons in an adjacent layer are the result of an activated affine transformation that involves a subset of the input neurons ($x' \in \Re^{K_1xK_2xC}$) and a common set of weights (i.e., kernel). Convolution generates a filter map, $h \in \Re^{N-K_1+1,M-K_2+1}$, which can be understood as the spatial representation of the feature learned by the kernel. Normally, a large number of kernels are used since many spatial features can be learned from the input space. To keep the spatial resolution of the outcome equal to that of the input space, padding can be applied by adding the necessary zero-valued rows and/or columns before the convolution operation. Actually, the dimensions of the output space could also depend on other parameters which are not of interest for this Thesis (e.g., stride). Note that when the dimensions of the kernel equal the dimensions of the input space, the CNN is equivalent to the dense layer described in section 4.2.2. In fact, the convolutional layer is equivalent to impose a strong prior over the weights in a dense layer. This prior says that the parameters are zero except for a valued receptive field representing a sub-domain —as defined by the kernel dimensions,— of the whole input/feature map, which in addition are equal to the weights of its neighbour, whose valued receptive field is shifted in space. This strong assumption makes CNN useful for applications where the input patterns present local (spatial) correlations, as is the case, for instance, of the atmospheric fields. The interested reader is referred to Goodfellow et al. (2016) in case more details are desired.

4.2.4 Multi-task Neural Networks

In multi-task NNs (Caruana, 1998) several tasks are simultaneously learned with the same model, sharing a latent space defined by the dimensionality of the hidden layers. For example, suppose a 2-task neural model for image classification in which the objective is to identify if a cat and/or a dog appears in a given picture. The output layer would be formed by two sigmoidal neurons, one for each animal, which would be activated when the animal in question is detected. The motivation behind these topologies relies on their implicit generalization property (Baxter, 1995). Intuitively, note that features that are relevant to identify cats may also be useful to identify dogs. This in turn would produce a similar effect than that of increasing the size of dataset, which would allow for a more robust estimation of the network's parameters and a better generalization power. Mathematically, predicting multiple tasks at a time can be viewed as imposing constraints to the single predictive tasks by forcing the latent space to generalize to all the tasks addressed in the model. Due to this regularization property, DL topologies are commonly designed in multi-task mode, especially in applications for which the data available is very limited (see Ruder (2017) for an overview). In section 4.4 we extend these ideas to neural-based models designed for climate downscaling.

4.2.5 Optimization of Neural Networks

But, how does the network optimize the parameters ω? In the late 80s neural networks were successfully trained thanks to the gradient descent method and the backpropagation algorithm (Rumelhart et al., 1986). Whilst gradient descent provides a mathematical formulation (see equation 4.3) to update the coefficients proportional to a learning rate, η, the backpropagation algorithm provides an easy and accessible way to repeatedly compute the partial derivatives $\frac{\partial E^{\omega}(x,y)}{\partial \omega}$ at every iteration. The number of iterations needed to train the network (e.g., until the coefficients reach a minimum in the surface error) are called epochs.

$$\omega' = \omega - \eta \frac{\partial E^{\omega}(x, y)}{\partial \omega} \tag{4.3}$$

Instead of computing the partial derivatives over the whole dataset, these can be estimated by limiting the calculation to a random subset of size m, namely "batch". This is commonly referred to as Stochastic Gradient Descent (SGD) (Bottou et al., 2018). Though the functioning of SGD constitutes nowadays an active area of research (Zhang et al., 2021), its use results into low computational requirements and regularization properties.

$$\omega' = \omega - \frac{\eta}{m} \Sigma_{j=1}^{m} \frac{\partial E^{\omega}(x, y)}{\partial \omega} \tag{4.4}$$

The choice of the loss function is case dependent (e.g., regression or classification problems). Among the wide variety of available loss functions, it is of special interest the negative log-likelihood, which allows for estimating the parameters of the distribution of interest by conducting some kind of maximum likelihood estimation. In fact, the most commonly used loss functions —mean squared error for regression and cross-entropy for classification tasks— can be deduced from the log-likelihood approach when the objective variables are Gaussian- or Bernoulli-like distributed, respectively (Goodfellow et al., 2016). The ability to infer any conditional probability distribution is very relevant for climate-related problems, which usually involve non-Gaussian variables (e.g., wind and precipitation) and is therefore one key aspect of the DL models developed in this Thesis. This will be explained in more detail in section 4.4.

4.3 Deep Learning for Climate Science

Along the last decades, the climate community has produced a huge amount of data from different sources, including in-situ measurements, radar images and model simulations, among others (Overpeck et al., 2011). While the number of in-situ observations is not expected to increase substantially in the next decades, the volume of satellite and model data is expected to reach 150 and 350 petabytes by 2030, respectively.

There is therefore growing interest within the climate research community towards developing machine learning and DL models which can take advantage of this large volume of data.

To date, a wide battery of statistical techniques are routinely applied with different aims in climate-related problems, which include 1) to replace physical components of numerical models with faster (and more accurate) statistical schemes, 2) to learn links between the different climate components when the physics are either unknown or very complex to model, and 3) to post-process climate products. For instance, some studies have attempted to emulate climate models (Meyer et al., 2021), or to learn sub-grid parameterizations (Chevallier et al. (1998), Seifert and Rasp (2020)). Others have reported successful use-cases for ML-based weather forecasting (Papale and Valentini (2003), Landschützer et al. (2013), Kühnlein et al. (2014)), or have employed these techniques for climate model evaluation (Nowack et al., 2020), or for the detection and attribution of anthropogenic climate change (Barnes et al., 2019); among others (e.g., data assimilation (Gilbert et al., 2010)). Also, ML has been traditionally used to downscale seasonal (Manzanas et al., 2020b) and long-term (Gutiérrez et al., 2019) climate simulations.

Despite promising results have been found in some of these works, traditional statistical techniques are insufficient to tackle the challenges (and opportunities) that arise from the ever-growing amount of climate data available. As a consequence, the climate community has paid attention to other promising machine learning techniques, in particular DL models (Monteleoni et al. (2013), Reichstein et al. (2019)), due to their recent success in computer vision applications. Despite certain similitude can be found between computer vision and climate —both have to deal with high-dimensional input spaces,— several challenges arise in the latter; among others the lack of interpretability and physical consistency and the high demand of computational resources (Reichstein et al., 2019).

Nevertheless, to date, some studies have already shown promising results regarding the usefulness of DL in several climate applications.

- **Emulation of climate models.** Several works have tested the suitability of DL topologies to emulate certain components of a climate model, or even full simple climate model formulations. For instance, Scher (2018) used an off-the-shelf CNN to emulate the evolution of four variables of a GCM. Lguensat et al. (2019) did the same to learn the dynamics of the upper ocean, as described by a quasi-geostrophic model. Since these DL models do not explicitly incorporate any physics, Beucler et al. (2019) forced some NNs to model the conservation of energy for the emulation of cloud processes. Following from the success of these works, the community has even coined the term Neural Earth System Modelling (NESYM, Irrgang et al. (2021)).

- **Parameterization of sub-grid processes.** NNs can be trained to learn parameterization schemes based on observational records and/or high-resolution climate model simulations, to improve the existing ones (Schneider et al., 2017). For instance, the radiation of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational model is parameterized with a shallow NN (Chevallier et al., 1998). Also, atmospheric convection —which is one of the key processes that limit predictability nowadays— was skillfully parameterized with DL topologies (Gentine et al. (2018) and Rasp et al. (2018)). Moreover, an alternative approach consists of directly emulating an existing parameterization scheme with NNs (Krasnopolsky and Fox-Rabinovitz, 2006).

- **Forecasting.** To date, a diverse number of DL topologies have been used for forecasting purposes at different time-scales. For instance, CNNs have showed potential to forecast atmospheric fields a few days into the future (Scher and Messori, 2019; Weyn et al., 2019; Scher and Messori, 2018). Also, Recurrent Neural Networks (RNN) and Long-Short-Term-Memory (LSTM) ones have been used for nowcasting of precipitation (Xingjian et al., 2015) and for air pollution forecasting (Chang et al., 2020). The interest in RNNs and LSTMs is due to their ability to learn the right temporal structure by incorporating lagged-information in the model (Hochreiter and Schmidhuber, 1997). The limits of DL in forecasting applications are yet unknown, and the community wonders if these models can fully replace the current operational systems which are currently based on physical principles (Dueben and Bauer, 2018). In this regard, it is particularly interesting the work presented by Rasp et al. (2020), who have recently released a benchmarking dataset intercomparing the performance of different DL topologies for medium-range weather forecasting. So far, a LSTM architecture has already achieved better forecast skill in lead times up to 12 hours, than the Weather Research and Forecasting (WRF) numerical weather prediction (NWP) model (Hewage et al., 2021), for the prediction of several surface atmospheric variables.

- **Extreme events detection.** Typically, both observations and numerical simulations contain “hidden” patterns which provide useful information about the state of the climate system. Nevertheless, identifying these patterns and predicting their evolution —which can be relevant for impact studies— is a challenging task. In this context, CNNs have been used for the detection of atmospheric rivers (Chapman et al., 2019) and extreme events (Liu et al., 2016), for hurricane tracking (Giffard-Roisin et al., 2020) and for the estimation of cyclone intensity (Pradhan et al., 2017).

Aside from these applications, it is of particular interest to this Thesis the DL models developed with downscaling purposes. We devote the next section for a detailed review of the state-of-the-art in this matter.

4.4 State-of-the-art in Deep Learning and Statistical Downscaling

NNs appeared in the 90s as a promising regression-based technique to downscale atmospheric fields (Wilby et al. (1998), Schoof and Pryor (2001)). They were designed to overcome the limitation of (generalized) linear models to learn non-linear links, which is especially relevant for downscaling of precipitation (Yuval and Hsieh, 2006). Moreover, NNs were found useful due to other properties soon. In particular, the hidden layers learnt latent representations which consist of shared relevant features useful to downscale at multiple sites. The potential of this latent space in multi-site architectures revealed successful to model spatial dependencies (MacKay (1997), Caruana (1998)), crucial in some sectorial applications such as hydrology (Salathé Jr, 2005). This contrasts with well-adopted approaches at the time, mostly based on multivariate linear regression (Uvo et al., 2001), which needed to be constrained with complex covariances matrices (Bürger, 1996) to include spatial information in the downscaling. Nonetheless, Cannon (2008) showed that, even with this latent structure, spatial constraints still had to be explicitly included in the model formulation —for instance in the form of loss functions— for an accurate reproduction of spatial fields.

The quality of any regression-based model mostly depends on whether the large-scale is able to explain the local-scale. Given a particular predictor configuration, (deterministic) regression fits the data by estimating the conditional mean of a predictive distribution. Consequently, the extremes are poorly represented unless the predictors are able to explain most of the local variance, which is not usually the case. The variance of the possible outcomes given a predictor configuration, mostly comes from noise measurements (e.g., calibration error in the observational instrumentation) or uncertainty in the model parameters —which can be improved by increasing the size of the dataset (Gal, 2016).— To account for these sources of uncertainty, NNs can be designed in a PP-WG setup to estimate conditional daily probability distributions by minimizing the negative log-likelihood of a certain predictive distribution. This approach is especially relevant to downscale precipitation, which is usually triggered by local processes not described by the coarse-resolution predictors. Differently, temperature is strongly linked to the large-scale configuration and therefore, downscaling models for this variable commonly minimize the mean squared error —which is equivalent to estimate the mean of a conditional Gaussian distribution given a certain predictor configuration (Goodfellow et al., 2016)— in PP mode.

To build a PP-WG model to downscale precipitation, one needs to select a proper parametric distribution which fits both the discrete (0: no rain, 1: rain) and the heavy-tailed nature of this variable. Several studies addressed this issue by modeling a variety of density functions, which include the Poisson-Gamma (Dunn, 2004), the Bernoulli-lognormal (Vandal et al., 2018a) and the Bernoulli-Gamma (Williams, 1998) distributions. This (stochastic) regression-based modelling permits to characterize the uncertainty of the predictions given a particular large-scale configuration, and therefore account for the possible extremes by sampling from the estimated conditional distributions. In this Thesis we follow Williams (1998) and Cannon (2008) to build PP-WG models, and thus infer Bernoulli-Gamma distributions to downscale precipitation. The mathematical formulation of this approach is described in equation 4.5

$$P(y|x;p,\alpha,\beta) = \begin{cases} 1-p & \text{if } y < 1 \\ \frac{p}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-\frac{y}{\beta}} & \text{if } y >= 1 \end{cases} \tag{4.5}$$

where $P(y|x;p,\alpha,\beta)$ is the daily distribution of rainfall, y, conditioned on a given daily predictor configuration, x, parameterized by the probability of rain, p, and the shape and scale parameters of a Gamma distribution, α and β[1]. Our NNs are trained to minimize the negative log-likelihood of this probability distribution (see equation 4.6). In addition, this loss function depends on the number of samples or batch size, N, the observed precipitation casted to a boolean object, y', and the Gamma function, $\Gamma()$ (see Williams (1998)).

$$L(\omega) = \frac{-1}{N}\Sigma_N (1-y')\log(1-p) + y'(\log p + (\alpha-1)\log y - \alpha\log\beta - \log\Gamma(\beta) - \frac{y}{\beta}) \tag{4.6}$$

The NNs based on this loss function can be reformulated as equation 4.7: given an input predictor set, the model (parameterized by ω) returns the three distributional parameters describing the conditional daily Bernoulli-Gamma distribution at a particular site.

$$[p,\alpha,\beta] = f^{\omega}(x) \tag{4.7}$$

Under the multi-task approach, these three parameters are estimated simultaneously at every site. Therefore, the output layer in this formulation consists of $i \times 3$ neurons, with i being the number of distributions (i.e., sites) inferred. Note that the inference is performed from independent daily Bernoulli-Gamma distributions (one per site) that share a common latent space rather than from a unique multivariate Bernoulli-Gamma distribution.

Overall, simple NN configurations proved successful to provide accurate downscaled predictions in “perfect” conditions due to their capacity to learn non-linear patterns among the large- and the local-scale. Moreover, they developed a mathematical formulation which produces probabilistic forecasts, crucial to reproduce the extremes. The literature addressing the suitability of NNs to downscale in the climate model space is very limited, with very few works digging into this topic. For instance, Quesada-Chacón et al. (2021) has recently presented a case-study in which single-hidden-layer NNs were used to downscale precipitation at three gauge stations in Central America until the end of the 21st century.

[1] Note that, for simplicity, we use $p(x) = p$, $\alpha(x) = \alpha$ and $\beta(x) = \beta$ in Eqs.4.5, 4.6, and 4.7.

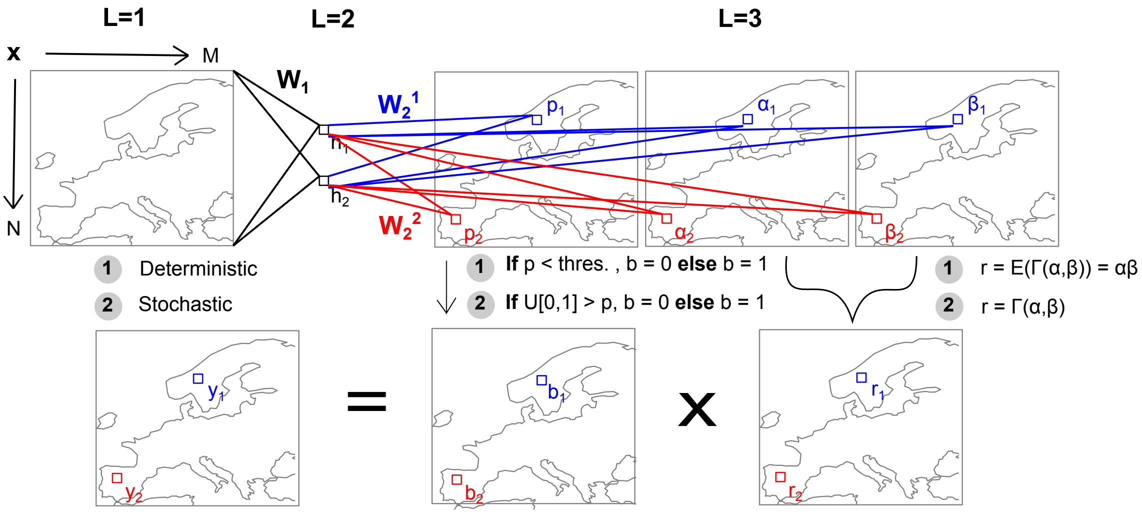


Figure 4.5: Diagram illustrating a 3-layer neural network that estimates conditional daily Bernoulli-Gamma distributions, as described in Williams (1998). The output layer consists of three output neurons, $y_i = \{p_i, \alpha_i, \beta_i\}$ if $i = 1$ (i.e., single-site mode). The complete predicted series is the result of a day-to-day multiplication, $y_i = b_i r_i$ between 1) a binary series, b_i, adjusted according to a threshold (in some studies this threshold depends on the observed frequency of rainy days (Gutiérrez et al., 2019)) and the expectance of the conditional Gamma distribution, $E(\Gamma)$, given by $r_i = \alpha_i \beta_i$, or 2) an independent sampling for both precipitation occurrence and amount series. Note that whereas 1) produce deterministic predictions, 2) gives place to stochastic ones.

The authors reported plausible downscaled projections for future emission scenarios, as compared to degraded RCM simulations for the same region.

Despite these merits in both "perfect" and climate model spaces, classical NNs still suffer from overfitting and predictor fields usually need to undergo subjective "human-guided" selection or dimensionality reduction techniques (see e.g., Olmo and Bettolli (2021)). The result is that the information coming from the large-scale predictor variables is only partially (e.g., restricted to very small domains) or poorly (e.g., arbitrary feature compression techniques which do not account for the predictor-predictand dependence) exploited. Moreover, the overparameterized structure of these models limits their use to very local case studies, stepping behind of RCMs in continental-sized downscaling.

Since the explosion of DL, several studies have investigated the suitability of these architectures for downscaling tasks with the idea of overcoming the above mentioned limitations. Broadly, these studies[2] can be grouped into the following categories (note that the search of literature has not been limited to climate downscaling):

[2] Please be aware that most of the these studies have been released during the development of this Thesis.

- **DL in synthetic case-studies.** Very naive toy examples have been developed to rapidly test off-the-shelf DL topologies for downscaling tasks. These case-studies basically upscale a gridded observational dataset to a coarser resolution (usually precipitation) and then use the low- and high-resolution fields as input-output pairs to build the model. For instance, Vandal et al. (2018b) built a CNN-based topology which estimated patches of local precipitation over the United States (US). This architecture was named DeepSD by the authors, and was lately modified to quantify the uncertainty of the precipitation estimates by assuming Bernoulli-lognormal predictive distributions (Vandal et al., 2018a). This topology was intercompared with other convolutional-based topologies in Kumar et al. (2021), yielding the best results for downscaling of daily summer monsoon precipitation over India. The authors also passed from this synthetic experiment to a more realistic case in which the calibrated DL model was used to downscale a state-of-the-art reanalysis. Wang et al. (2021) analyzed the suitability of deeper CNNs for the downscaling of daily temperature and precipitation over two areas in the US. To avoid degradation —i.e., a loss in accuracy in the predictions as a consequence of increasing the number of hidden layers— the authors included skip-connections in the network. This type of topology is called Deep Residual Network (DRN), since it aims to learn a mapping of the residual function (see He et al. (2016) for more details). Also, Sha et al. (2020a) and Sha et al. (2020b) aimed to estimate temperature and precipitation, respectively, based on U-NET topologies (Ronneberger et al. (2015), i.e., auto-encoder inspired) over the Western region of the US. Finally, Mu et al. (2020) included "a priori" meteorological knowledge in the DL models, describing multi-scale spatial correlations and the chaotic behavior of the atmosphere (Lorenz, 1963). This hybrid approach, which combines DL with physical principles, successfully reproduced some key spatial local dependencies. The models described herein, could be further exploited for more realistic applications in MOS mode. However, due to their dependence on predictor surface variables, we do not expect these topologies to work well in a PP setup.

- **DL in MOS setups.** On the one hand, inspired by certain computer vision applications which aim to reconstruct a high-resolution image from its low-resolution counterpart, a bunch of studies applying DL for SD tasks have recently appeared. These studies establish an empirical link between GCM coarse variables (e.g. precipitation in the historical scenario) and high-resolution observational fields —even though no great temporal correspondence between the two datasets is expected.— This link is then applied to downscale the future coarse simulations from the same

GCM under different emission scenarios. For instance, Liu et al. (2016) aimed to downscale monthly precipitation using a RSN fed with thirty-five stacked GCMs as input channels in the predictor field. A similar procedure was adopted in Rodrigues et al. (2018) to downscale daily precipitation based on a combination of CNNs and locally-connected networks. However, only Tran Anh et al. (2019) moved from historical simulations and applied a calibrated LSTM network to obtain station-scale projections from RCP simulations. A more plausible approach is based on unsupervised learning by leaning on Generative Adversarial Networks (GAN[3]). Chaudhuri and Robertson (2020) and François et al. (2021) deployed these topologies to improve the spatial consistency of annual and daily precipitation fields, respectively. On the other hand, DL has been also applied to RCMs. In particular, to link RCM evaluation runs —simulations nested to a reanalysis in nudging-mode— with a set of fine-grained observational records. This link can then be applied to the same RCM, both in historical and RCP scenarios. For instance, Steininger et al. (2020) produced a gridded 0.1° precipitation field by downscaling one RCM for the period 2000-2015. The DL topology used, which was based on CNNs, outperformed conventional MOS techniques for the same purpose.

- **DL in PP setups.** None of the above described studies fulfill the PP assumptions for the predictor set, since the majority build on surface variables, which are not well reproduced by GCMs. Indeed, very few DL topologies have been developed to date for PP downscaling —note that reanalysis data are always used in PP to learn the relationship between the large- and the local-scale, regardless the type of simulation to be downscaled: weather forecast, seasonal prediction or climate change projection.— The first attempts used convolutional (Vandal et al., 2017) and auto-encoder[4] (Vandal et al., 2019) topologies, with no clear benefits as compared to other well-established machine learning techniques. Nonetheless, several studies have successfully estimated local precipitation and temperature over different regions of the globe. On the one hand, Pan et al. (2019) used a combination of convolutional and dense layers which outperformed existing machine learning techniques over the United States (US). On the other hand, Sun and Lan (2021) intercompared a variety of CNN topologies to downscale gridded precipitation and temperature over

[3] GANs consist of a pair of mutually competing NNs with the overall objective of generating samples that preserve the coherence of the observed fields, for instance the inter-variable correlations (see Goodfellow et al. (2014) for more details).

[4] An auto-encoder is a particular NN topology in which the input and output spaces are equally-shaped. For instance, spatial atmospheric fields at different times are used as the input-output pairs in a weather forecasting application. The main characteristic of these topologies is that input features are compressed (encoder) by defining a low-dimensional hidden space. The dimensionality is then increased in a second stage of the network (decoder) until it reaches the original dimensions (see Goodfellow et al. (2016)).

China, with successful results for the former and no clear added value for the latter. On a sub-seasonal to seasonal scale, a conditional Generative Adversarial Network (CGAN, Mirza and Osindero (2014)) was used to downscale spatially consistent seasonal forecasts of temperature over the Iberian Peninsula, achieving satisfactory results when considering both reanalysis and (seasonal) model predictors (Gómez-Gonzalez et al., 2021). A similar study was performed by Miao et al. (2019), who aimed to adjust daily precipitation estimates of the subseasonal-to-seasonal ECMWF model over South China using a CNN-LSTM network. For station-scale downscaling, to our knowledge only Vaughan et al. (2021) has built a DL model based on convolutional conditional neural processes. This type of architecture has the advantage that permits to produce predictions to arbitrary off-the-grid locations not seen during the calibration period. Remarkably, none of the previous studies have passed from the reanalysis world ("perfect" conditions) to the GCMs world. To our knowledge, Stengel et al. (2020) presented the only case-study in which a PP-based DL model was applied to downscale a GCM scenario. Despite its simplicity, —they do not provide an extensive validation of their method neither in historical nor in RCP simulations,— as compared to CNNs optimized with point-based error loss functions, this study showed the potential of GANs to attain impressive levels of local spatial correlations for hourly wind fields. Finally, note that except Vaughan et al. (2021), all the DL models collected in this group perform a deterministic regression, leaving aside the uncertainty of the conditional local distributions.

- **DL to emulate RCMs.** As explained, RCMs are very expensive in terms of computational resources and just a few supercomputing centers around the world can run them. Statistical emulators based on DL topologies have recently emerged as a potential alternative to overcome this issue by trying to mimic the work done by a RCM. This can be done either by 1) using the GCM-RCM fields as input-output pairs to construct the DL model, or by 2) upscaling the circulation RCM variables to a coarser spatial resolution (predictors) and use the original high-resolution RCM fields as "pseudo-reality" (predictands). Note that 1) follows the philosophy of MOS-SD whilst 2) would be a form of PP-SD in the space of the RCM. On the one hand, the only MOS-inspired DL topology reported to date is the work done by Babaousmail et al. (2021), who stacked a set of GCMs as input channels to feed a CNN aimed at learning the mapping between these GCMs and an ensemble of high-resolution fields of monthly precipitation. In this setup, the target variables came from the RCA4 RCM, driven by each of the input GCMs. This study showed promising results to emulate the ensemble mean of the RCM simulations. However, improvements need still to be done regarding the emulation of each individual RCM. On the other hand, Serifi et al. (2021) degraded the original spatial resolution of temperature and precipitation from one RCM to derive their counterpart high-resolution fields over Central Europe with CNN-DRN topologies. Also, Doury et al. (2021) used coarsened large-scale variables from one RCM as predictors to emulate near-surface temperature over the Mediterranean with a U-NET inspired topology, with satisfactory results.

Despite many of these studies yield promising results regarding the applicability of DL for SD problems, the variety of DL topologies used, regions analyzed and GCM/RCMs —operating at different time-scales— considered make very difficult a fair, comprehensive intercomparison. Indeed, the use of DL for SD tasks is still an incipient field of research with many important questions to be answered yet. For instance, there is a clear lack of works focused on the study of the extrapolation capabilities of the different topologies, which may (or may not) justify their potential use for SD of climate change projections. This Thesis aims to fill part of these knowledge gaps by assessing if CNNs are able to outperform classical SD methods for downscaling of local temperature and precipitation in "perfect" conditions and by studying their suitability to generate high-resolution climate change projections over Europe, based on different emission scenarios.

Part II

Data and Methods

CHAPTER 5

Experimental Framework

This chapter describes the experimental framework followed in most of the analysis carried out for the elaboration of this Thesis. This framework has been designed in the European COST action VALUE, which is introduced in section 5.1. In section 5.2 we present the observational datasets, GCMs and RCMs used. The post-processing of the predictor variables used for SD is explained in section 5.3. Section 5.4 introduces the CNNs and the benchmarking GLMs considered for SD. Finally, in section 5.6 we present the metrics employed to validate the downscaled produced in both observational and climate model spaces.

5.1 The COST Action VALUE

The COST action VALUE (Maraun et al., 2015) was designed to provide an experimental framework to assess and intercompare different SD techniques in the context of climate change research. This European initiative gathers climatologists, stakeholders, impact modellers and statisticians to foster collaborations, ease the transfer of knowledge and improve the quality of research in downscaling.

Overall, VALUE aims to answer the following questions:

- Can we gather the downscaling community into a single collaborative initiative to promote a better understanding of the regional climate over Europe?

- Can we provide a common framework to comprehensively analyze the advantages and limitations of (well-established) SD techniques, including PP-, MOS- and WG-like methods?

	Objective	Exp.	Predictand Data	Specific Objectives
Exp. 1	- Provide a framework to intercompare and validate SD methods in “perfect” conditions, i.e., using observed datasets for both predictor and predictand variables.	**1a**	Station data	Perform a 5-fold cross-validation (1979-2008) to measure the downscaling skill. Predictors are taken from reanalysis data which is supposed to represent accurately the state of the atmosphere.
		1b	Gridded data	Same as Exp.1a but for E-OBS (Cornes et al. 2018), to asses observational differences between the grid box-level and the station-scale (Klein Tank et al. 2009).
		1c	Nested station data	Evaluate the spatial aspects of the downscaled series (e.g., pairwise cross-correlation, decorrelation length, variogram range).
		1d	Station data	Same as Exp.1a but focusing on the sub-daily scale.
Exp. 2	- Provide a framework to evaluate the SD models developed in Experiment 1 to downscale GCMs (both historical and future emission scenarios).	**2a**	Station data	Validation of both downscaling performance and errors inherited by the GCMs, at station-scale.
		2b	Gridded data	Same as Exp.2a but for gridded data. This enable the comparison with RCM projections, as it avoids the representativeness problem.

Table 5.1: Summary of the VALUE experiments (see Maraun et al. (2015) for more information), including their main objective and the target resolution of interest (i.e., station-scale or gridded data). Yellow indicates the experiments which are directly related to this Thesis. We refer the reader to this website http://www.value-cost.eu/ for more details on the VALUE experiments.

- Can SD models extrapolate to climate change conditions?
- Can we provide stakeholders and practitioners with user-friendly portals which ease the access to the downscaled products generated?

To address these questions, VALUE has defined a set of experiments which are summarized in Table 5.1. To date, only Experiment 1, the largest-to-date intercomparison of SD methods in "perfect" conditions —to which 27 European institutions contributed,— has been officially concluded. This experiment has led to a collection of publications which evaluate the suitability of the many different (well- established) SD methods to reproduce the 1) extremes (Hertig et al., 2019), 2) marginal (Gutiérrez et al., 2019), 3) temporal (Maraun et al., 2017b) and 4) spatial properties (Widmann et al., 2019) of the observed records. Moreover, a synthesis of the full experimental design, objectives and challenges foreseen in VALUE can be found in Maraun et al. (2019).

Recently, VALUE has evolved to EURO-CORDEX Empirical Statistical Downscaling (EURO-CORDEX ESD, Jacob et al. (2020)), a branch of EURO-CORDEX[1] that focuses on SD. Most of the work developed in this Thesis aligns with the goals of VALUE and EURO-CORDEX ESD and takes advantage of the data and experimental frameworks developed in these two initiatives. This has allowed us to test the suitability of CNNs both in "perfect" conditions and in the climate model space at a continental-sized level.

5.2 Data Used

In this section we first introduce the observational datasets considered to build and assess the performance of our SD models (both CNNs and GLMs) in "perfect" conditions. Then, we also introduce the GCMs and RCMs which have been selected to test the suitability of CNNs for downscaling of climate change projections.

As in VALUE's Experiment 1 (see Table 5.1), our CNNs and GLMs were build in "perfect" conditions based on ERA-Interim (Dee et al., 2011) predictors on a 2° regular grid. Note that the original spatial resolution of this reanalysis is 0.75°. However, we have degraded it to a coarser 2° grid for better compatibility with the GCMs listed in Table 3.1, whose resolution range in between 1° and 3°. The set of possible predictor variables considered for Experiment 1 is listed in Table 5.2. Based on previous literature dealing with SD over Europe (Huth, 2002, 2005; Gutiérrez et al., 2013; San-Martín et al., 2017; Gutiérrez et al., 2019), this set includes circulation and thermodynamic variables at different altitudes (500, 700, 850 and 1000 hPa).

[1] EURO-CORDEX is the European community of the Coordinated Region Downscaling EXperiment (CORDEX), a global initiative that aims to develop high-resolution climate change projections worldwide building on RCMs.

Variable (code)	Units	Height (hPa)					Predictor	Predictand
		Surface	1000	850	700	500		
Zonal wind velocity (ua)	m/s		6 7.1	6 7	6 7	6 7	✓	
Meridional wind velocity (va)	m/s		6 7.1	6 7	6 7	6 7	✓	
Air temperature (ta)	$^{\circ}C$		6 7.1	6 7	6 7	6 7	✓	
Specific humidity (hus)	$kgkg^{-1}$		6 7.1	6 7	6 7	6 7	✓	
Geopotential (z)	m^2/s^2		6 7.1	6 7	6 7	6 7	✓	
Sea level pressure (slp)	Pa	7.2					✓	
Precipitation (pr)	mm/day	6 7						✓
Near-surface air temperature (tas)	$^{\circ}C$	6 7						✓

Table 5.2: List of variables used in this Thesis. Since we do not lean on the same set of (predictor) variables throughout the Thesis, we indicate with numbers the section in which they participated. We use a ✓to designate whether any variable is used as predictor or as predictand in the models developed. All variables have a daily temporal resolution.

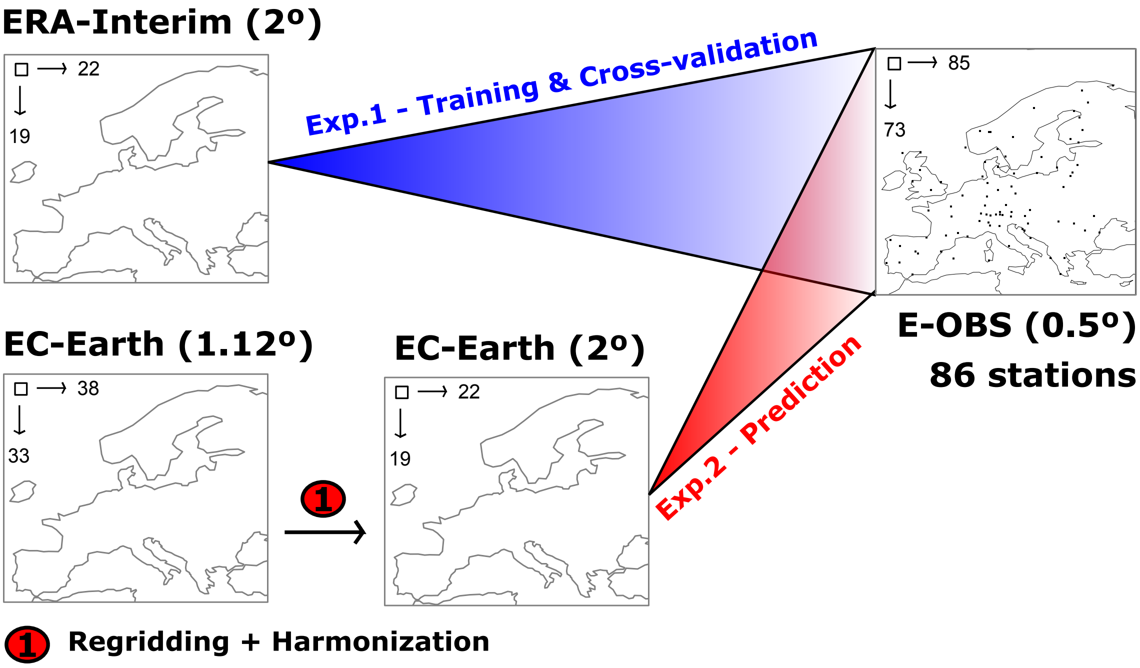


Figure 5.1: Schematic representation illustrating the grids used for the predictors (left) and the predictands (right) considered in the Experiments 1 (blue) and 2 (red) of VALUE. For instance, a 19 by 22 latitude-longitude grid over the domain of study is considered for ERA-Interim reanalysis. Note that the map on the right shows both the E-OBS grid and the 86 stations addressed in VALUE's Experiment 1a.

VALUE proposed to downscale daily precipitation and temperature fields both at a station-scale (Experiment 1a) and over high-resolution grids (Experiments 1b, 2). For the former, 86 stations covering the different climates of the continent were selected. For the latter, the E-OBS (version 14, Cornes et al. (2018)) dataset, which provides daily temperature and precipitation on a 0.5° regular grid, was considered. Figure 5.1 shows an schematic representation of the spatial structure of the predictor and predictand fields used in this Thesis.

VALUE's Experiment 1, (Gutiérrez et al., 2019) covered the period 1979-2008, which was split into 5 chronological folds (1979-1984, 1985-1990, 1991-1996, 1997-2002, 2003-2008) for cross-validation purposes. However, since one of the goals of this Thesis is to study the extrapolation capability of the proposed SD methods, we used for Chapter 6 a hold-out approach in which the total period of study was split into independent train (1979-2002) and test (2003-2008) sets. Figure 5.2 shows some climatology statistics of local temperature (top panel) and precipitation (bottom panel) for these two periods: the mean, and the 2^{nd} (P02) and 98^{th} (P98) percentiles for temperature; and the daily rainfall, the frequency of rainy days ($\geq$ 1mm/day, R01) and the P98 for precipitation. The second and fourth rows show the differences between the test and train (taken as reference) period for these statistics. On the one hand, warmer conditions (of about 1°) are found for the test period for the three temperature metrics displayed —except for the P02 in some regions of Southern Europe.— On the other hand, precipitation statistics do not seem to vary significantly between the train and the test periods and the notable differences found in the North of Portugal, Southern Greece and Lithuania are due to deficiencies in E-OBS, —changes or interruptions in the national station networks used to construct the dataset— rather than real climatology variations. This figure proves that the hold-out approach followed provides a reasonable scenario to evaluate the ability of our SD methods to extrapolate to unseen conditions during the calibration phase (e.g., a warmer climate).

In VALUE's Experiment 2, the statistical models built in "perfect" conditions during Experiment 1 are subsequently applied to GCM predictors. In particular, to the 12^{th} run of EC-Earth (**?**), since this GCM is known to satisfactorily reproduce key large-scale patterns observed over Europe, including storm tracks (Lee, 2015). As already explained in 3.3, this is crucial since the GCM predictors used for SD of climate change projections should realistically resemble their counterpart variables in the reanalysis. Based on this idea, the SD models built in Chapter 6 are applied to EC-Earth's predictors for the historical (1979-2008) and RCP8.5 (2071-2100) scenarios in section 7.1. Moreover, section 7.2 extends this analysis to the subset of CMIP5 models listed in Table 3.1. For compatibility, all these GCMs are re-gridded by means of nearest interpolation to the ERA-Interim's 2° grid. Unlike in section 7.1, in which just the far-future (2071-2100) is considered, in section 7.2 we downscale the periods 1979-2005 and 2006-2100 for the historical and RCP8.5 scenarios, respectively. Furthermore, in section 7.2, this ensemble of SD-based projections is compared against a representative subset of RCMs from EURO-CORDEX, listed in Table 5.3. These RCMs, which are used as "pseudo-reality" (Vrac et al., 2007b), have been

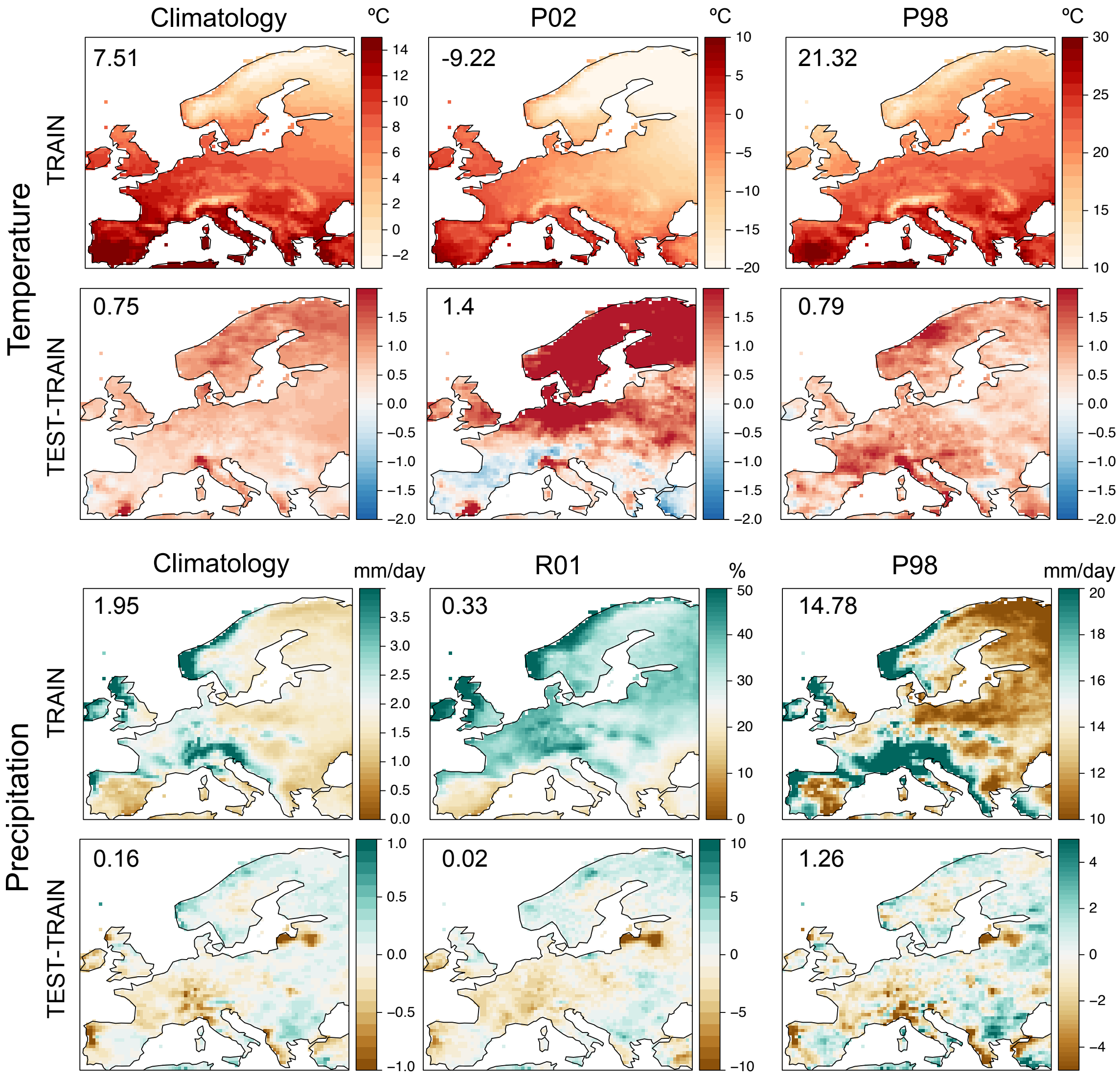


Figure 5.2: Top panel, top row: E-OBS climatology for the mean, the P02 and the P98 values of temperature in the train period (1979-2002). Top panel, bottom row: Mean difference between the test (2002-2008) and train period —the latter is taken as reference— for the statistics shown in the top row. Bottom panel: As the top panel, but for precipitation. In this case, the mean, the frequency of rainy days (R01) and the P98 are shown. In all cases, the numbers within the maps indicate the spatially averaged values.

selected based on their driving GCMs to match those used as inputs in the SD models. For direct comparison against the SD-based projections, all the EURO-CORDEX RCMs were re-gridded by nearest interpolation from its original spatial resolution of 0.44°[2] to the E-OBS's grid, at 0.5°.

[2] They belong to EURO-CORDEX 44, a specific experiment which aims to produce climate projections over Europe with a spatial resolution of 0.44°

GCM	Run	RCM	Institution
CanESM2	1^{st}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
CNRM-CM5	1^{st}	CLMcom-CCLM5-0-6	Climate Limited-area Modelling Community
CNRM-CM5	1^{st}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
MPI-ESM-LR	1^{st}	CLMcom-CCLM4-8-17	Climate Limited-area Modelling Community
MPI-ESM-LR	1^{st}	MPI-CSC-REMO2009	Max Planck Institute for Meteorology
NorESM1-M	1^{st}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
GFDL-ESM2M	1^{st}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
EC-EARTH	12^{th}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
EC-EARTH	12^{th}	CLMcom-CCLM5-0-6	Climate Limited-area Modelling Community
IPSL-CM5A-MR	1^{st}	SMHI-RCA4	Swedish Meteorological and Hydrological Institute, Rossby Centre
IPSL-CM5A-MR	1^{st}	IPSL-INERIS-WRF331F	Institut Pierre-Simon Laplace

Table 5.3: List of RCMs considered in section 7.2 (third column). The first column indicate the GCM to which each RCM is nested. The second and fourth columns show the particular GCM run and the institution responsible for running the RCM simulations, respectively.

VALUE has made publicly available both ERA-Interim's and EC-Earth's predictors proposed for Experiments 1 and 2, which are used in Chapters 6 and 7 for model calibration and prediction, respectively. These data can be downloaded as *netCDF* files from `http://www.meteo.unican.es/tds5/catalogs/value.html`. E-OBS (version 14) is available at the European Climate Assessment Dataset (ECA&D) website: `https://www.ecad.eu/download/ensembles/download.php`.

5.3 Harmonization of Global Climate Model Predictors

To ensure that GCM and reanalysis predictor fields are reasonably similar —note that this is one of key requirements that should be fulfilled in PP-SD (see section 3.3.2),— recent studies have demonstrated that GCM predictors need to undergo some form of post-processing, even in GCMs which have proved robust to reproduce key large-scale atmospheric processes (Cheng et al., 2008; Vrac and Ayar, 2016; San-Martín et al., 2017; Nikulin et al., 2018; Manzanas et al., 2020b).

In this Thesis we follow the approach adopted in Vrac and Ayar (2016), in which GCM predictors are subjected to harmonization and standardization procedures. Harmonization consists of adjusting the GCM (x_{GCM}) monthly means towards the corresponding reanalysis values (x_{REA}) at a gridbox level. This is done to overcome the possible misrepresentation of the seasonal cycle in the GCM. Equations 5.1 and 5.2 formulate mathematically the mentioned harmonization for the historical and RCP8.5 scenarios, h and f, for a given predictor variable j and month $i = \{1, 2, ..., 12\}$. Additionally, in order to avoid undesired artifacts related to possible magnitude mismatches among different variables, standardized values (at the gridbox level) are considered. Each reanalysis/GCM predictor is standardized based on its own mean and standard deviation over a reference period:

1979-2002 in Chapter 6[3] and 1979-2008 in Chapter 7.

$$x'^{i}_{j,GCM_h} = x^{i}_{j,GCM_h} - \bar{x}^{i}_{GCM_h} + \bar{x}^{i}_{REA} \tag{5.1}$$

$$x'^{i}_{j,GCM_f} = x^{i}_{j,GCM_f} - \bar{x}^{i}_{GCM_h} + \bar{x}^{i}_{REA} \tag{5.2}$$

To assess whether or not the PP requirement of having similar predictors in the reanalysis and in the GCM (see Brands et al. (2011a) Gutiérrez et al. (2013) for an example) is fulfilled, we rely in this Thesis on the Kolmogorov-Smirnov test (KS), which measures the maximum distance between two cumulative density functions. The test was applied, independently for each predictor variable in Table 5.2, to compare EC-Earth and ERA-Interim over the period 1979-2009. With illustrative purposes, Figure 5.3 shows the results obtained for ta1000 and hus700 (for brevity, results for the rest of variables are skipped here). Red crosses in the left (right) panel identify those gridpoints where the null hypothesis of the test —reanalysis and GCM distributions are indistinguishable— can be rejected at a 95% confidence level when the EC-Earth is (is not) harmonized.

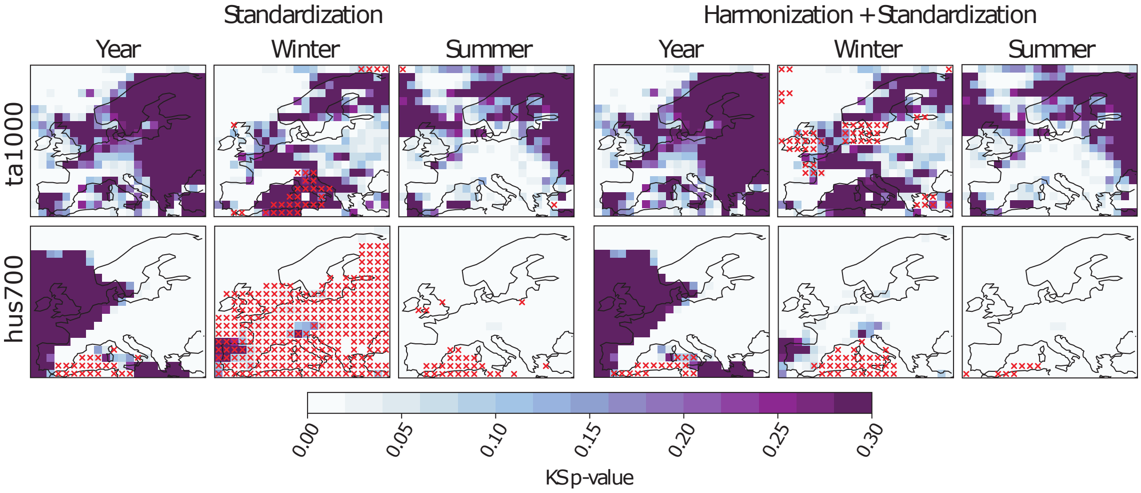


Figure 5.3: P-value from the Kolmogorov-Smirnov (KS) test applied to quantify the similarity of raw EC-Earth and ERA-Interim distributions for temperature at 1000 hPa and specific humidity at 700 hPa (top and bottom row, respectively), considering the entire year, winter and summer (in columns) during the period 1979-2008. Red crosses in the left (right) panel identify those gridboxes where the null hypothesis of the test —both distributions are indistinguishable— can be rejected at a 95% confidence level when the EC-Earth is (is not) harmonized. See the text for details.

Both ta1000 and hus700 present in general low p-values (below the significance level of

[3] This chapter focuses on the results of the performance of the different SD methods considered in “perfect” conditions, i.e., using reanalysis predictors. Therefore, only standardization is applied in this case.

0.05), reflecting that EC-Earth and ERA-Interim raw distributions are significantly different over many regions. This is partially explained by the systematic biases that are usually exhibited by GCMs. However, the situation is substantially improved once standardization —or harmonization plus standardization— is carried out (see the red crosses in the corresponding panels). In particular, if EC-Earth and ERA-Interim distributions are compared over the entire year (left column) both ta1000 and hus700 fullfil the PP hypothesis tested practically over the entire domain (with a few exceptions in the Mediterranean for the case of hus700). Differently, when the comparison is undertaken for winter and summer (middle and right column, respectively), different results are found depending on whether or not harmonization is applied, which suggests the importance of this procedure for PP-SD. Indeed, for other predictor variables —especially wind velocity components in Southern Europe and specific humidity at other height levels— harmonization is crucial to make reanalysis and GCM predictors compatible (not shown). Therefore, both standardization plus harmonization are applied to all the GCMs considered for SD in this Thesis.

5.4 Deep Learning Models

As explained in section 4.1, NNs are usually seen as "black-box" models, which leads to distrust feelings among the climate community (Reichstein et al., 2019). For downscaling, a proper an extensive comparison of DL topologies aiming to shed light on the use of these techniques has not been addressed yet. To partially overcome this shortcoming, we propose in this Thesis different configurations of CNNs of increasing levels of complexity, which are listed in Table.5.4. This allows us to marginalize the role that each element of the network plays for SD tasks. In Chapter 6 we present a comprehensive evaluation of the performance of these CNNs in "perfect" conditions. Based on the results from this analysis, we select the best-performing models to downscale GCM scenarios in Chapter 7.

Model	**Topology**	**Rationale**
CNN-LM	`inp`-**50**-**25**-**1**-`out`	Using convolutions to obtain meaningful spatial features
CNN1	`inp`-**50**-**25**-**1**-`out`	Testing the added value of non-linearity
CNN10	`inp`-**50**-**25**-**10**-`out`	Increasing model complexity from 1 to 10 feature maps
CNNdense	`inp`-**50**-**25**-**10**-50-50-`out`	Combining the spatial patterns with dense layers
CNN-PR	`inp`-**10**-**25**-**50**-`out`	Using standard topologies from pattern recognition

Table 5.4: Topology of the CNNs developed and intercompared in this Thesis. `inp` and `out` stand for the input and output layers, respectively. The numbers indicate the amount of neurons or filter maps (boldfaced numbers) conforming every hidden layer. The symbols '-' and '-' denote the dense and convolutional connections, respectively. In addition, a brief rationale describing the purpose of each topology is also given.

Figure 5.4 shows a schematic representation of the CNNs developed in this Thesis. The particular network here illustrated, which is designed for precipitation downscaling, consists of a 3D (latitude-longitude-variable) input layer. Similarly to the RGB channels in computer vision models, the spatial fields of the different predictor variables (20 in this example) are stacked along the third dimension. The architecture continues with a three convolutional layer stage (deeper architectures were tested with no added value). Data flow through the hidden layers —which present linear (CNN-LM) or nonlinear ReLU (CNN1, CNN10, CNN-PR and CNNdense) activation functions,— transforming the input space into meaningful spatial patterns. Finally, the last hidden layer is fully-connected to the output layer. Consequently, most of the parameters are located in this part of the network which centralizes a large fraction of the learning power. The CNNs proposed in this Thesis follow the distributional estimation approach presented in Williams (1998) and explained in section 4.4. The output layer produces estimates for the daily parameters of a Bernoulli-Gamma (Gaussian) probability function for the downscaling of precipitation (temperature) based on three (two) output neurons per predictand site. We use linear activation functions on the output neurons describing the parameters of the probability distributions proposed —either Bernoulli-Gamma for precipitation or Gaussian for temperature,— except for the probability of rain, p, for which we use a sigmoidal function. The output layer in Figure 5.4 consists of three vectors of size 3258, associated to the land gridboxes in E-OBS showing no missing values. These vectorized fields are reshaped to the latitude-longitude domain of study (73x85 gridboxes) after model calibration, during the prediction phase.

For this Thesis we developed different CNNs of increasing levels of complexity. The simplest one is referred to as CNN-LM. This network presents linear activation functions in the hidden layers. Similar to the computation of Empirical Orthogonal Functions (EOFs), setting linear activation functions at every layer leads to a projection of the data into an alternative space which is the result of a linear combination of the input dimensions. We use this model as “control” for the presence of non-linearities in the predictor-predictand link. Next, we built the non-linear version of CNN-LM, labelled as CNN1, which is able to learn more complex functions (if necessary). According to the experience gained with DL models in other disciplines, limiting the last layer to one feature map (CNN-LM and CNN1) is quite restrictive. Therefore, we increased the number of filter maps to ten (CNN10). Since the last hidden layer is fully-connected to the output layer in our convolutional networks, we needed to find an adequate compromise between model complexity and the ability to take advantage from meaningful, complex data patterns. Building on this idea, we designed a topology which placed two fully-connected layers of fifty neurons each right after the convolutional stage (named CNNdense). This network allows for diminishing the number of parameters in the last layer. Besides, computer

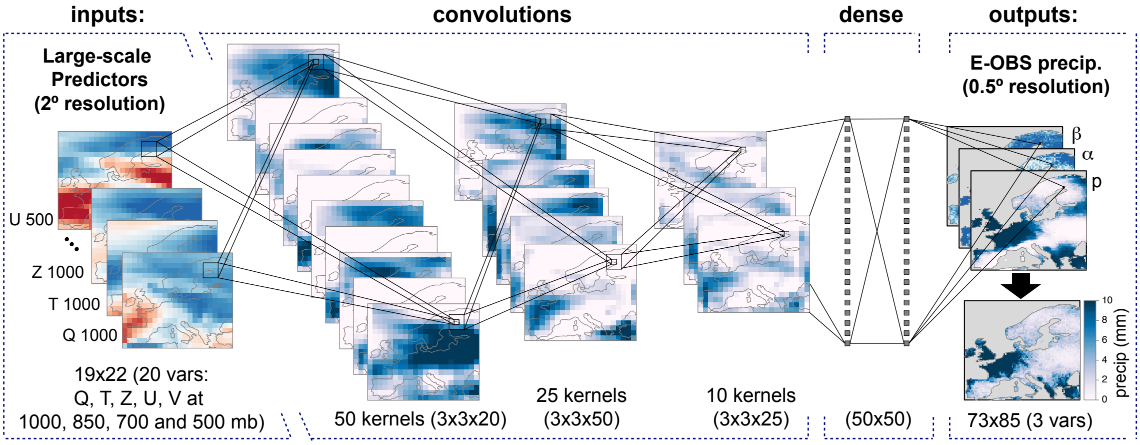


Figure 5.4: Schematic representation of the CNNs developed in this Thesis. This network includes a block of three convolutional layers with 50, 25 and 10 feature maps —produced based on a kernel of dimensions $3 \times 3 \times \#channels$ (different kernel sizes were tested with no added value),— respectively, followed by two fully-connected (dense) layers with 50 neurons each. This illustrative network is designed to downscale daily precipitation, and therefore the output layer —formed by three neurons at each predictand site— estimates the daily parameters of a conditional Bernoulli-Gamma distribution (p, α and β), see the text for details. Based on these parameters, either deterministic or stochastic daily downscaled precipitation values can be produced (see section 4.4).

vision models typically present this type of topology in which the spatial patterns learned are mixed in the dense connections. Finally, inspired by computer vision applications, we also proposed the CNN-PR model, which contains an increasing number of filter maps. This configuration allows for learning a larger amount of complex spatial patterns. With regards to the dimensionality of all these CNNs (listed in Table 5.4), we considered padded and non-padded versions of each topology. Padding adds artificial zeroes in the borders of the 2D feature maps to preserve the original dimensions of the input space after the convolutional operation. If padding is not applied, the dimensions of the output space are reduced by a magnitude that depends on the kernel's size (Goodfellow et al., 2016). We found that keeping the original latitude-longitude dimensions (i.e., padding) led to better results for the networks with low-dimensional spaces in the last hidden layer (CNN-LM and CNN1). Therefore, all the results shown in section 6.1 correspond to the best version (with or without padding) of each of the CNNs listed in Table 5.4.

5.5 Generalized Linear Models

Along the Thesis, we used as benchmark for the CNNs described in the previous section, three different implementations of the GLM technique (see section 3.3.3).

Several reasons supported this choice: 1) GLMs are regression-based models, which allows for a fair comparison with CNNs, 2) GLMs are well-established within the climate community and have been extensively used for SD tasks, and 3) despite their simplicity, GLMs ranked among the best performing methods in VALUE's Experiment 1 (Gutiérrez et al., 2019)). In particular, we focused in this Thesis on the GLMs developed in Bedia et al. (2020), a recent extension of the VALUE's Experiment 1. These GLMs use local predictor information at the 'n' closest gridpoints to each predictand site. Panels a and b in Figure 5.5 illustrate the case of $n = 1$ (GLM1) and $n = 4$ (GLM4), respectively. Increasing 'n' allows to assess the influence of larger spatial local patterns. In this same line, we also considered the GLMPC method appearing in Gutiérrez et al. (2019), which uses as predictor the leading principal components (PCs, Preisendorfer and Mobley (1988)) instead of local fields. In particular, this model reduces the input space by projecting the predictor set over the PCs that explain the 95% of the total variance over each PRUDENCE region (Christensen and Christensen, 2007). Panel c shows these regions, which roughly correspond to different climate regimes, in squares of different colors. Note that, differently to the DL models presented in section 5.4, which are able to automatically handle high-dimensional input spaces in an efficient way, tedious and human-guided feature selection or feature reduction techniques (i.e., PC analysis) have to be applied to generate a meaningful predictor set for the case of GLMs. Indeed, the full set of predictor variables listed in Table 5.2 was not even tested for these models since they are known to overfit under high-dimensional input spaces (see section 3.3.4).

Note also that, actually, three different GLM-like models were needed at each predictand site (thus moving from multi-site downscaling in CNNs to single-site downscaling in GLMs) for the three configurations shown in Figure 5.5: a logistic regression, a Gamma regression with logarithmic link and a Gaussian regression. The first two models address the downscaling of precipitation; in particular its binary (0: no rain, 1: rain) and continuous (rainfall amount in wet days) aspects, respectively. Similar to the downscaling performed with CNNs (see Figure 4.5 in section 4.4), the estimated parameters —p in the logistic regression, and α and β in the Gamma model— can be used to generate either deterministic predictions or stochastic ones. The third of the models, the Gaussian regression one, directly returns the estimated local temperature.

5.6 Evaluation Metrics

Assessing the performance of any SD method is not trivial. Several aspects of the predictions should be evaluated, including marginal, temporal and spatial properties. To date, downscaling methods are tested in different case studies, each using its own

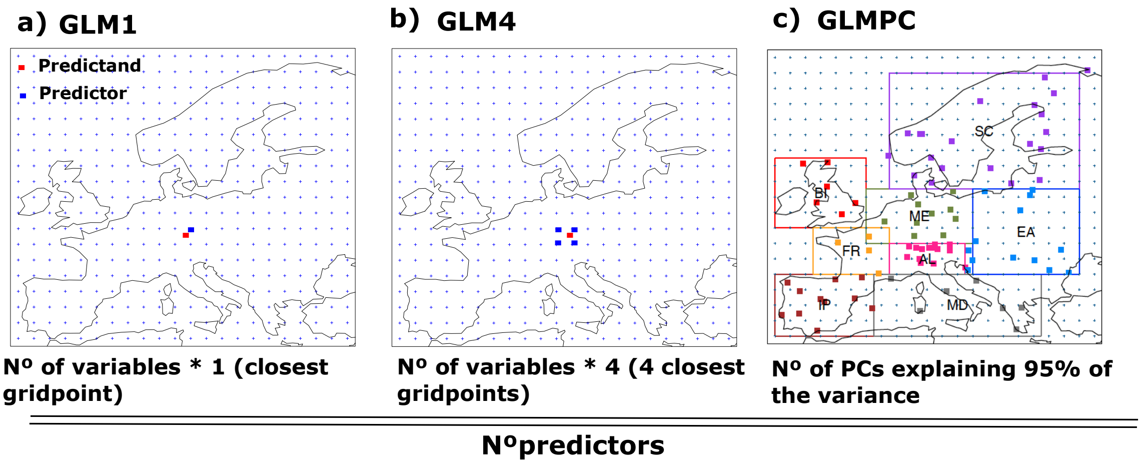

a) GLM1
Predictand
Predictor
b) GLM4
c) GLMPC
SC
BI
ME
EA
FR
AL
IP
MD
Nº of variables * 1 (closest gridpoint)
Nº of variables * 4 (4 closest gridpoints)
Nº of PCs explaining 95% of the variance
Nºpredictors

Figure 5.5: Schematic representation showing the different predictor configurations used in GLM1, GLM4 and GLMPC methods (from left to right). Blue squares in Panels *a* and *b* indicate the gridboxes in which local predictor information is used to build the statistical model for a particular predictand location, marked with a red square. Panel *c* displays the eight PRUDENCE regions, delimited by squares of different colors. The 86 stations considered in VALUE's Experiment 1a are also shown in different colors (according to the PRUDENCE regions they belong to). In all cases, the 2° regular grid in which predictors are available is shown with blue dots.

validation metrics, what makes difficult a meaningful assessment of their performance and prevent from a fair intercomparison. To alleviate this issue, VALUE developed a comprehensive list of indices and measures (available at the VALUE validation portal: `http://www.value-cost.eu/validationportal`) which allows to properly evaluate the most relevant forecast aspects. Moreover, these metrics were implemented in the *R* package *VALUE* (`https://github.com/SantanderMetGroup/VALUE`), which facilitates research reproducibility. We present in this section the subset of VALUE metrics used along this Thesis.

On the one hand, Table 5.5 shows the indices and metrics used for the validation in "perfect" conditions of the temperature (tas) and precipitation (pr) downscaled fields presented in Chapter 6. For temperature, we validate the biases of our predictions (i.e. their mean errors with respect to the observations for a common period, expressed in °C) for the mean, 2^{nd} and 98^{th} percentiles. In addition, we also consider the ratio between predicted and observed standard deviations (the closer to 1, the better). For precipitation, we consider the biases relative to the observed value (expressed in %) for the mean and the P98. Moreover, to validate the accuracy of the downscaled series, we also compute for both temperature and precipitation the Root Mean Squared Error (RMSE), which measures the average error. For the particular case of precipitation, the RMSE is conditioned on the observed wet days (≥ 1 mm). Furthermore, to assess the quality of the temporal structure of the predictions, we also measure their correlation with the corresponding observations. For temperature, the Pearson correlation coefficient —which measures the degree of linear dependence between the predicted and the observed series and is therefore appropriate for Gaussian variables— is used. Differently, for precipitation, we employ the Spearman correlation, which is based on ranks and is therefore more adequate for non-gaussian variables. For completeness, we also calculate the biases of the lag-1 autocorrelation (i.e., correlation of the series shifted by 1 day) and a number of annual maximum indices, including the warm (WAMS), cold (CAMS), wet (WetAMS) and dry (DryAMS) spells. For precipitation, we additionally include the bias of the relative amplitude of the annual cycle. Finally, to properly validate the probabilistic prediction of precipitation occurrence, we also consider the ROC skill score (ROCSS, e.g., Manzanas et al. (2014)), which is based on the area under the ROC curve (see Kharin and Zwiers (2003), for details).

Description	**Units**	**Variable**	**Perfect score**
Bias (for the mean)	$^\circ C$, %	tas, pr	0
Bias (for the 2nd percentile, P02)	$^\circ C$	tas	0
Bias (for the 98th percentile, P98)	$^\circ C$, %	tas, pr	0
Root Mean Square Error (RMSE)	$^\circ C, mm/day$	tas, pr	0
Ratio of standard deviations (Std ratio)	-	tas	1
Pearson correlation	-	tas	1
Spearman correlation	-	pr	1
ROC Skill Score (ROCSS)	-	pr	1
Bias (warm annual max spell, WAMSl)	days	tas	0
Bias (cold annual max spell, CAMS)	days	tas	0
Bias (wet annual max spell, WetAMS)	days	pr	0
Bias (dry annual max spell, DryAMS)	days	pr	0
Bias (lag 1 autocorrelation, AC1)	-	tas	0
Bias (relative amplitude of the annual cycle)	-	pr	0

Table 5.5: Subset of VALUE metrics used for the validation in “perfect” conditions of the downscaled fields (see Chapter 6). The symbol ‘-’ indicates that the corresponding metric is dimensionless.

On the other hand, Table 5.6 lists the metrics used in Chapter 7 to asses the suitability of the different SD models used for downscaling in the climate model space. For temperature, we consider the mean, and the 2^{nd} (P02) and 98^{th} (P98) percentiles of the distributions, whilst for precipitation we additionally include the frequency of wet days (R01), the mean precipitation amount in wet days (Simple Daily Intensity Index: SDII) and the 98^{th} percentile of the wet-day distribution (P98Wet). All these metrics allow both for a robust validation of the downscaled fields in the historical scenario and for a good evaluation of their plausibility in the RCP8.5 (by comparing the statistics of the downscaled projections with those from an ensemble of GCMs and RCMs). Note that in section 7.1 we compute the relative biases of the precipitation statistics in Table 5.6, whilst in section 7.2 we consider absolute biases.

Code	Description	Units	Variable
R01	Frequency of wet (≥ 1 mm/day) days	%	pr
SDII	Simple daily intensity index	mm/day	pr
P98Wet	98th percentile of the wet (≥ 1 mm/day) days distribution	mm/day	pr
P02	2nd percentile	°C	tas
Mean	Mean	°C	tas
P98	98th percentile	°C	tas

Table 5.6: List of metrics used to asses the suitability of the different SD methods considered for downscaling in the climate model space (see Chapter 7).

Part III

Main Results

CHAPTER 6

Downscaling in "Perfect" Conditions

This chapter focuses on the performance of CNNs for SD in "perfect" conditions —i.e., using reanalysis data as predictors.— On the one hand, in section 6.1 we intercompare a set of CNNs of increasing levels of complexity with classical GLMs. These benchmark methods ranked among the best ones in VALUE's Experiment 1a, which represents the largest-to-date downscaling intercomparison study in "perfect" conditions over Europe (Gutiérrez et al., 2019; Bedia et al., 2020). On the other hand, we analyze the "black-box" nature of NNs (section 6.2), one of the key factors that limit their use in the climate science. In particular, we 1) assess the benefits of multi-site topologies, as compared to single-site ones (section 6.2.1) and 2) shed some light about the modeling of the predictor-predictand link in the CNNs considered in this Thesis by producing and studying a set of saliency maps[1] (section 6.2.2).

The first part of this Chapter (section 6.1) is based on the manuscript entitled *"Configuration and intercomparison of deep learning neural models for statistical downscaling"*, published in *Geoscientific Model Development*. The second part (section 6.2) is based on a series of papers published in the proceedings of the Climate Informatics (CI) international conferences held in 2018, 2019 and 2020. Two of these works, *"Deep convolutional networks for feature selection in statistical downscaling"* and *"Understanding deep learning decisions in statistical downscaling models"* deal with the interpretability of CNNs. The third one, *"The importance of inductive bias in convolutional models for statistical*

[1] The term *saliency map* refers to any transformation of the information contained in an input image to another meaningful space which facilitates its interpretability. Saliency maps have been widely utilized in previous works to better understand the functioning of DL applications (Simonyan et al., 2014; Zhou et al., 2016; Zintgraf et al., 2017; Montavon et al., 2018; Larraondo et al., 2019; Reimers et al., 2019; Toms et al., 2021).

downscaling”, analyzes the benefits of multi-site topologies, as compared to single-site ones.

6.1 Performance Intercomparison for Different Deep Learning Models

This section discusses the performance of the different DL models proposed in this Thesis (see section 5.4) for SD in “perfect” conditions. We use as benchmark methods the GLM1 and GLM4 described in section 5.5 —we do not include GLMPC in this analysis since it was found to provide worse results than the two local GLMs in San-Martín et al. (2017) and Bedia et al. (2020).— This section is framed within VALUE’s Experiment 1b (see Chapter 5). Therefore, we produce daily precipitation and temperature over the 0.5° E-OBS grid for the period 1979-2008 building on ERA-Interim predictors (see Table 5.2). This study is limited to a hold-out approach (train: 1979-2002, test: 2003-2008) to measure the extrapolation capability of the CNNs and GLMs considered (see Figure 5.2 for a an assessment of the climatological differences between these two periods). Note that this aspect is crucial to provide insight into the applicability of these methods to climate change studies. To validate our downscaled results in “perfect” conditions we rely on the subset of VALUE metrics listed in Table.5.5.

Fig 6.1 shows the validation of the downscaling of temperature, in terms of nine different metrics. In each panel (one per metric), the results for the seven methods intercompared are shown by means of boxplots which represent the spread of the targeted metric along the entire E-OBS grid. The dark gray boxplot corresponds to the CNN10 model, which provides overall the best results for this variable.

Accuracy is evaluated in terms of the RMSE and the Pearson correlation (panels *a* and *b*, respectively). For the latter we have removed the seasonal cycle of the series, to avoid overestimated values. Even though they only differ in the number of neighbouring points included in the predictor field, GLM4 outperforms GLM1 for these two metrics. These results are consistent with previous literature suggesting that the use of spatial rather than local predictor information helps to gain prediction accuracy (Gutiérrez et al., 2013). In this regard, it is interesting to see that the CNN-LM model —which uses linear convolutional operations to exploit the spatial structure of the entire predictor domain— yields indeed better RMSE and correlation values than GLM1, but comparable to those from GLM4. This indicates that 1) a subset of the 4 closest gridboxes to a given site seems to be enough to gather most of informative power required to reproduce the local variability of temperature, and 2) the CNN-LM is able to deal with continental-scale predictor fields without leading to overfitting. Note that the latter is a clear benefit with respect to traditional SD methods since it avoids the use of “human-guided” selection

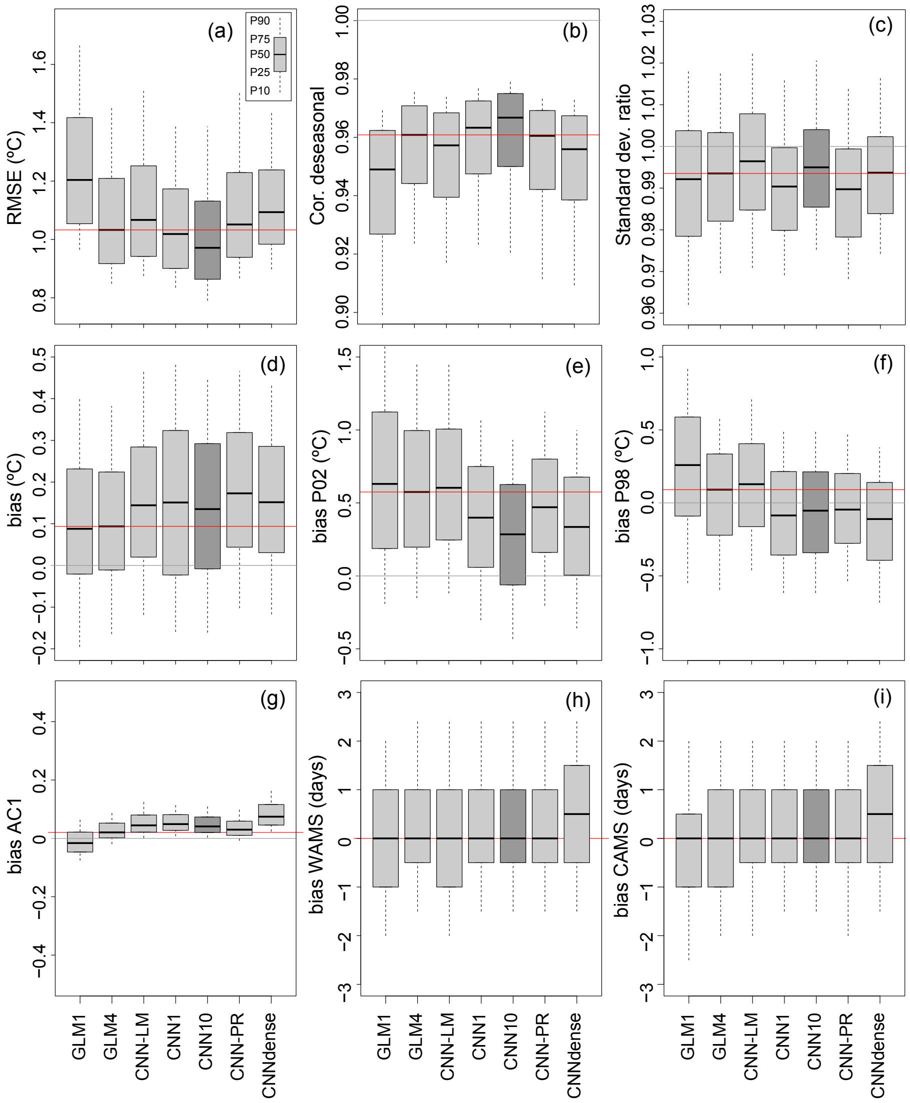


Figure 6.1: Validation of the downscaling of the temperature, in terms of nine different metrics. In each panel (one per metric), the results for the seven methods intercompared are shown by means of boxplots which represent the spread of the targeted metric along the entire E-OBS grid (the boxes/whiskers cover 25–75th/10-90th percentile range). For comparison purposes, the horizontal red line marks the median value for the benchmarking GLM4. The horizontal gray lines indicates the 'perfect' value for each metric. The dark gray boxplot identifies the overall best-performing method, in this case the CNN10.

procedures of filtering techniques to retain only the most explicate predictors. The CNN1 and CNN10 models achieve lower (higher) RMSE (Spearman correlation) values than the GLM4. However, these improvements are not excessively high, which suggests that non-linear models add actually little value for the downscaling of temperature. Similar results were obtained in previous studies over Europe testing the use of non-linear regression for this variable (Huth et al., 2008). The computer vision inspired topologies, CNN-PR and CNNdense, present worse results than GLM4 in terms of RMSE and correlation. For the CNN-PR, a high amount of filter maps in the last hidden layer may have overparameterized the net, whilst the CNNdense may have lost local-connectivity since the spatial patterns are fully-mixed in the dense layers.

To assess the performance of our predictions in terms of distributional similarity with E-OBS, we considered the ratio of standard deviations and the biases for the mean, P02 and P98 (panels *c*, *d*, *e* and *f* respectively). Overall, all the methods intercompared yield similar results. There is a general slight positive bias for the mean (0.1-0.2°C) that tends to be higher for P02 (0.3-0.5°C). For P98, positive and negative errors (in between -0.2 and 0.3°C) are found. These results indicate that all methods present moderate extrapolation capabilities (recall the test period is warmer than the train one).

Regarding temporal aspects, the biases for the AC1, the WAMS and the CAMS are shown in panels *g*, *h* and *i*, respectively. With the exception of the CNNdense model, which yields positive biases for all these metrics, very slight differences are found among the rest of models, all of them exhibiting zero-centred biases (note however that the spatial variability of these errors in considerable). Overall, no method outperforms clearly the others in terms of these temporal metrics. It has to be noted that neither the linear nor the CNN models have been specifically designed to preserve the temporal structure of the data, and improvements in this aspect might be achieved by recurrent connections or LSTM networks.

To gain spatial detail about the results presented in Figure 6.1, Figure 6.2 shows maps for a subset of the analyzed metrics. For simplicity, only the GLM1 and GLM4 are displayed, together with the best-performing CNN: the CNN10. The RMSE and the correlation exhibit similar spatial patterns regardless of the model considered. In particular, the highest (lowest) RMSE (correlation) values are found in Scandinavia, the Balkans and some regions over the Iberian Peninsula[2], for which the CNN10 yields slightly better results than the two GLMs. Moreover, GLM1 leads occasionally to patchy (discontinuous) spatial patterns, unlike the GLM4 which produces smoother fields as a result of incorporating predictor information representative of a wider area.

In terms of distributional similarity, all models exhibit similar spatial patterns for the ratio of standard deviations and the biases for the mean, P02 and P98. In particular, low mean errors appear in areas where the E-OBS dataset present inconsistencies (the Eastern Balkans and Southern Iberia). For the extremes, P02 and P98, a gradient of negative-

[2] As already pointed out in Sec5.2, the anomalous results found for Southern Iberia could likely be related to issues in the E-OBS dataset.

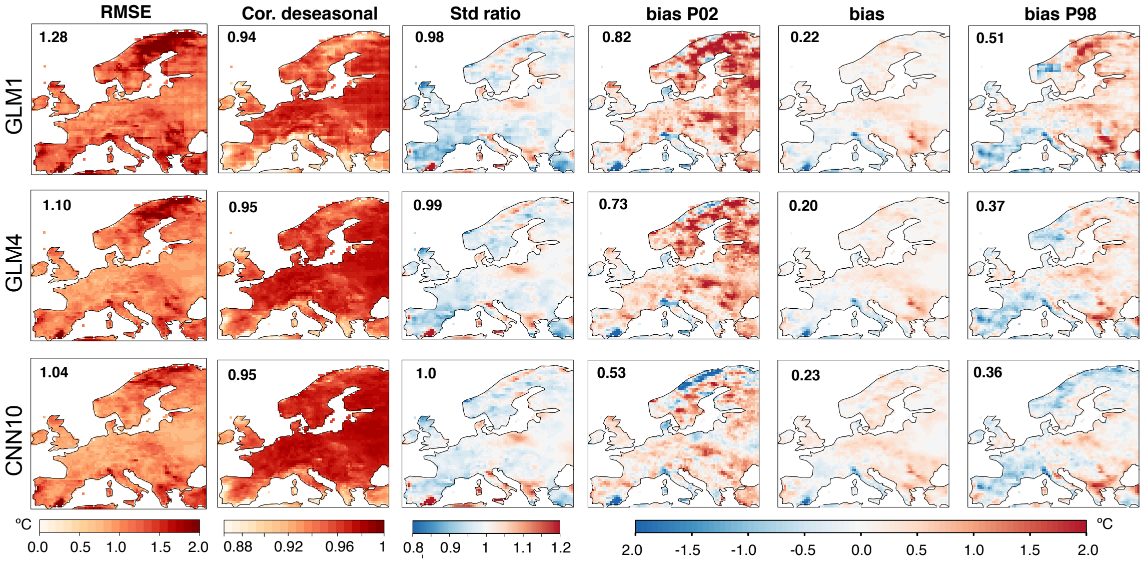


Figure 6.2: Spatial results for a subset of the validation metrics considered for the downscaling of temperature (in columns), for the two benchmarking GLMs (top and middle row) and the best-performing method, the CNN10 (bottom row). The numbers within the maps indicate the spatial mean absolute values (to avoid error compensation).

positive biases crosses Europe from west to east. Noticeable, the CNN10 model reduces the magnitude of these biases, especially for P02.

Similar to Figure 6.1 for temperature, Figure 6.3 shows the validation results obtained for the downscaling of precipitation. In this case, the dark gray boxplot corresponds to the CNN1 model, which provides overall the best results for this variable. Panels *a*, *b* and *c* show the ROCSS, the RMSE (conditioned on the observed wet days) and the Spearman correlation, respectively. There is a considerable improvement of GLM4 over GLM1 for the ROCSS and the correlation. However, both models exhibit similar RMSE values. These results indicate that precipitation occurrence (described by the ROCSS) is better predicted when making use of a wider area in the predictor field, which in turn improves also the correlation values attained. Nevertheless, predictor information at the closest gridbox seems to be sufficient to predict rainfall amount, which is usually not much affected by the processes that occur in far regions. This idea is supported by the results obtained for the CNN-LM. This model, which is designed to exploit the spatial structure of the entire predictor domain, yields very similar results to GLM4 for ROCSS, correlation and RMSE. Nevertheless, except for the CNN-PR —which behaves similarly to the GLMs,— all the (non-linear) CNN models tested achieve better validation scores for the these three metrics, in particular, the CNN1. Differently to the case of temperature (whose dependence on the synoptic situation is nearly linear), these results suggest that CNNs would outperform

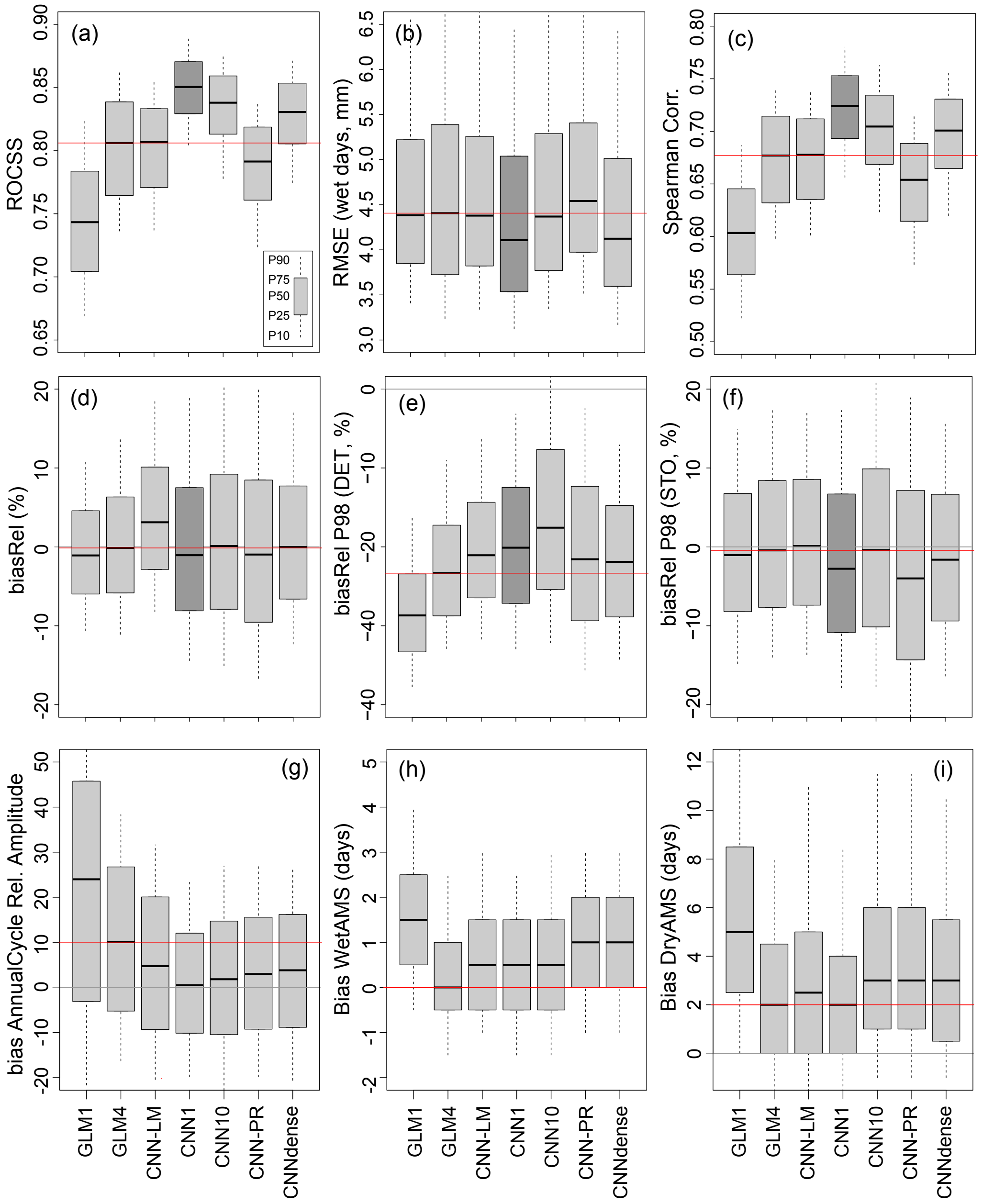


Figure 6.3: Equivalent to Figure 6.1 but for the validation of the downscaling of precipitation. In this case, the labels 'DET' and 'STO' for the relative bias of the P98 refer to the deterministic and stochastic versions of the different methods intercompared, respectively.

classical GLMs for downscaling of precipitation due to their ability to properly model the non-linearities linking this variable with the large-scale predictors.

To validate the marginal aspects of the downscaled precipitation, panels *d*, *e* and *f* show the relative biases for the mean, P02 and P98, respectively. For the latter, results for both deterministic and stochastic predictions are provided. For the mean, all models show zero-centred values (note however the spatial variability of these errors, represented by the spread of the boxplots). As expected, the deterministic predictions underestimate the P98 whilst stochastic ones show small, zero-centred biases. Presumably, the underestimation of extreme precipitation in the deterministic series might be a consequence of a lack of informativeness in the predictor set, which is typical for this variable.

Regarding temporal aspects, panels *g*, *h* and *i* show the biases for the relative amplitude of the annual cycle, the WetAMS and the DryAMS. For these three metrics, all models present positive biases, especially the GLM1. We argue that other type of NNs which explicitly take into account the temporal structure of the data such as LSTMs may help to improve these results.

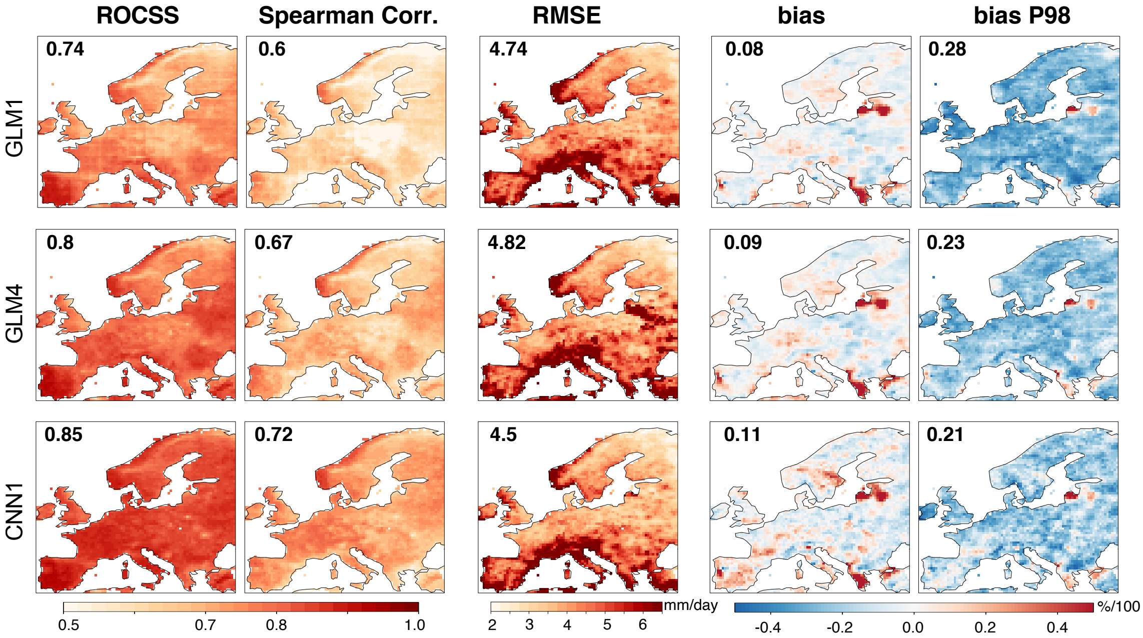


Figure 6.4: Equivalent to Figure 6.2 but for a subset of the validation metrics considered for the downscaling of precipitation. In this case, CNN1 is given in the bottom row for being the best-performing model for this variable. The numbers within the maps indicate the spatial mean absolute values (to avoid error compensation).

For further insight into the spatial distribution of the results shown in Figure 6.3, Figure 6.4 provides maps for some the analyzed metrics, in particular the ROCSS, Spearman correlation, RMSE, and the relative biases for the mean and the P98 (from left to right). In rows, the GLM1 and GLM4 are shown along with the best-performing CNN: the CNN1. The latter outperforms the two GLMs for the accuracy metrics (ROCSS, correlation and RMSE) whilst present comparable biases. The highest ROCSS and correlations are found over France and the Mediterranean arch, which may be explained by the type of mechanisms driving precipitation in these regions (mostly Atlantic fronts), which are well-represented by the large-scale atmospheric configuration. Likewise, the highest RMSE values are located in mountainous regions (e.g. the Alps), where precipitation is often related to convective processes and other local phenomena not reflected in the predictor set. Still, CNN1 exhibits lower RMSE values than the GLMs do, especially over Central and Eastern Europe.

Notice that the anomalous results found over Northeastern Iberia and the Baltic states for the bias of P98 are likely due to the issues identified in the E-OBS dataset. Nonetheless, particularly bad results are also found over the Greek peninsula (especially for the mean bias), for which we do not envisage a clear explanation.

The last column correspond to the bias for the P98, as obtained from deterministic predictions. As expected, extreme precipitation is underestimated by all methods. However, this issue can be overcome by stochastic predictions, although at the cost of losing part of the spatio-temporal consistency. To shed light on the benefits and shortcomings of both deterministic and stochastic predictions for rainfall amount, Figure 6.5 shows the results obtained for the CNN1 model when 1) directly predicting from the expected value of the conditional daily distribution learnt and 2) sampling out a new value from its parameters —in both cases, deterministic predictions of precipitation occurrence are considered.— In particular, panel *a* (*c*) shows the Spearman correlation over the entire time-series (conditioned to the observed wet days). Panels *b* and *d* correspond to the RMSE and the ratio of standard deviations. Similar correlations are found for both deterministic and deterministic-stochastic prediction when the complete time-series is assessed since its temporal structure is determined to a great extent by the binary occurrence. However, if the validation is restricted to the days in which observed precipitation was above 1 mm, the deterministic-stochastic prediction exhibit lower correlations than deterministic ones over the entire continent. Moreover, the introduction of a stochastic component leads to larger RMSE values (as compared to the deterministic predictions). Nonetheless, as shown by the ratio of standard deviations, it is needed in order to achieve a realistic variability in the predictions (note that this metric is clearly underestimated by purely deterministic predictions).

Overall, our results show that CNNs yield better validation metrics than the benchmarking GLMs in "perfect" conditions. In particular, CNN1 and CNN10 were shown to be the best-performing models for the downscaling of temperature and precipitation, respectively. This is due to 1) the ability of CNNs to learn complex and non-linear patterns from data — which has been found to be particularly relevant for the case of precipitation— and 2) their ability to efficiently handle high-dimensional input spaces in a multi-site

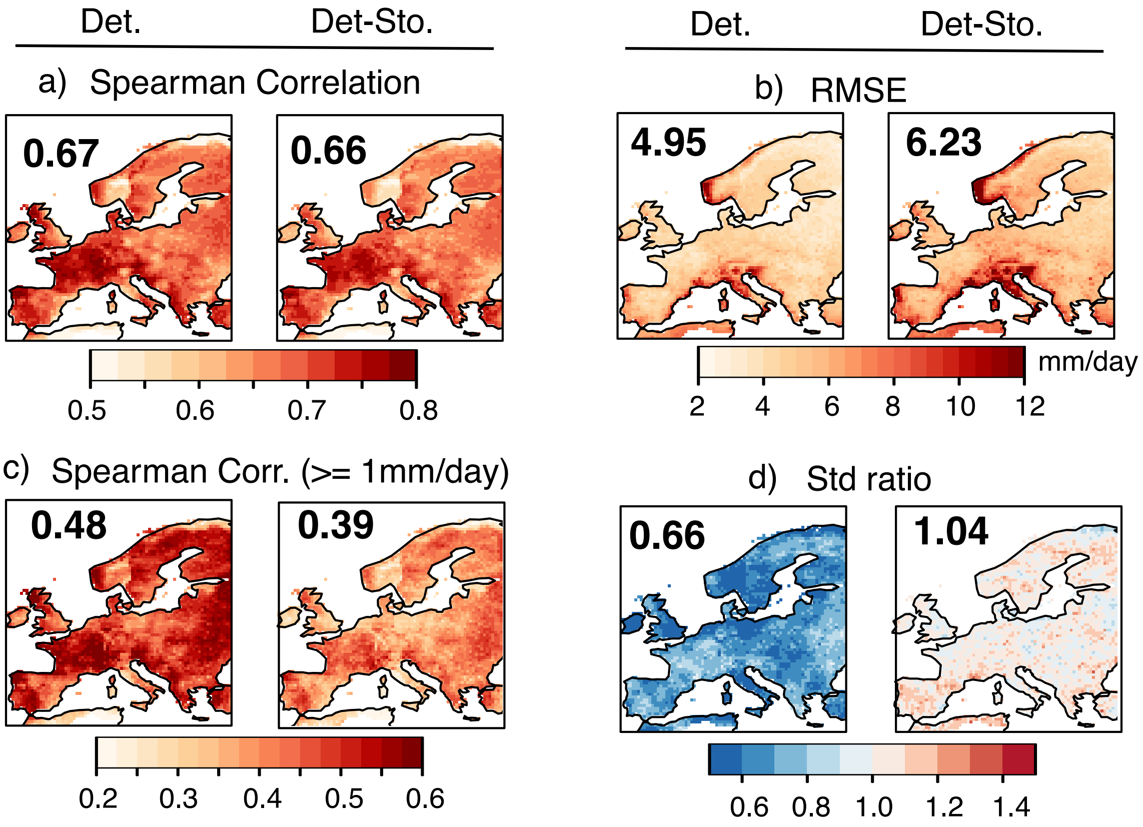


Figure 6.5: Comparison between deterministic (Det.) and deterministic-stochastic (Det-Sto.) predictions from the CNN1 model in terms of the Spearman correlation for the full series (a), the RMSE (b), the Spearman correlation conditioned to the observed wet days (c) and the ratio of standard deviations (d) —see the text for details about these metrics.— The numbers within the maps indicate the spatially averaged values.

configuration. The latter constitutes a clear advantage over traditional SD models since tedious and human-guided dimensionality reduction techniques —which may entail a loss of relevant information for the downscaling— are no longer needed. However, the reasons explaining this success for the CNNs are still unknown and we provide dig into this matter in the next section.

6.2 *Unveiling the “Black-box” Nature of Deep Learning Models*

This section delves into the functioning of DL topologies for climate downscaling and try to explain the success of the CNNs analyzed in the preceding section. In particular, in section 6.2.1 we explore the implicit regularization that occurs in multi-site convolutional networks. Moreover, with the idea of detecting the most relevant input features for a given climate downscaling application, we analyze in depth the modelling of the predictor-predictand link in our CNNs in section 6.2.2.

6.2.1 The Implicit Regularization of Multi-site Topologies

To better understand the reasons that may be behind the ability of CNNs to efficiently treat high-dimensional input spaces without leading to overfitting, we compare in this section single- and multi-site versions of the CNN1 model (labelled as CNN1-SS and CNN1-MS, respectively) for downscaling of precipitation. During single-site mode, each site is downscaled with independent statistical models (CNNs in our case), whilst multi-site topologies downscale simultaneously the entire predictand field in a single statistical model (see section 4.2.4). To enrich the analysis, we also include in this experiment the GLM4 model, which is limited to the PRUDENCE region encompassing the Iberian Peninsula (see Figure 5.5c). In contrast to section 6.1, which built on a hold-out approach, we use here the 5-folds defined in VALUE’s Experiment 1 for cross-validation purposes: 1979-1984, 1985-1990, 1991-1996, 1997-2002, 2003-2008.

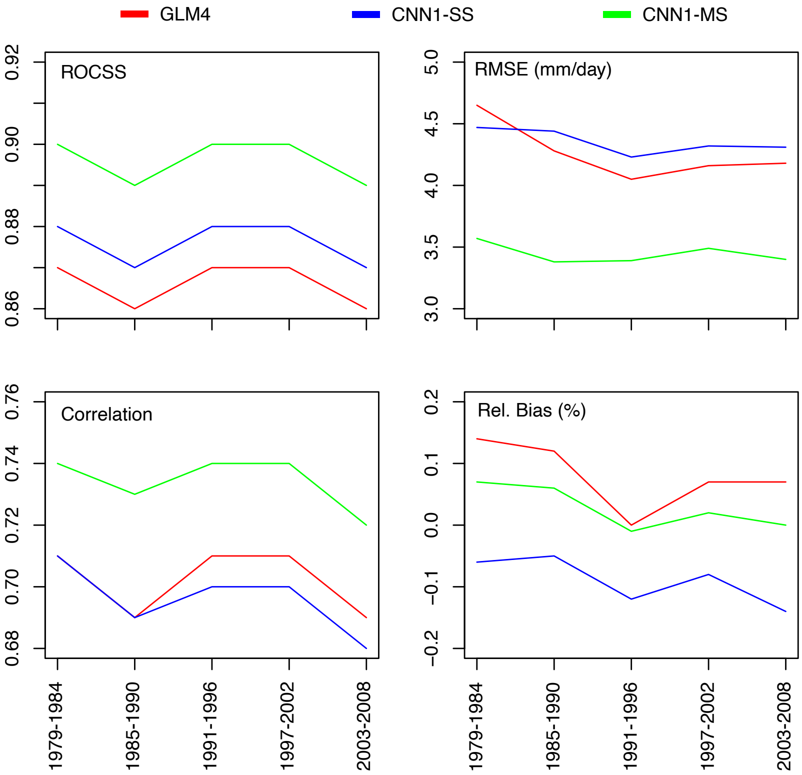


Figure 6.6: Cross-validated results in each of the five folds considered for the single- and multi-site versions of the CNN1 model (CNN1-SS and CNN1-MS, respectively) and the GLM4, in terms of the ROCSS, the RMSE, the Spearman correlation and the relative bias for the mean.

Figure 6.6 shows the results obtained in each fold for the CNN1-SS, the CNN1-MS and the GLM4 methods in terms of the ROCSS, the RMSE, the Spearman correlation and the relative bias for the mean, which have been spatially averaged across all predictand sites within the region of interest. In agreement with the results from the preceding section (Figure 6.3), very high ROCSS (with values close to 0.9) are found for the CNN1-MS, which outperforms the CNN1-SS and especially the GLM4. Likewise, the multi-site version of CNN1 exhibit better correlations than the rest of the models, with the CNN1-SS yielding even slightly lower values than the GLM4 for the period 1991-2008. A similar behaviour is also found for the RMSE, with the CNN1-MS yielding better results than both CNN1-SS and GLM4. Moreover, it has to be noticed that about a 5% of the predictand sites which exhibited very large RMSE values (above 10 mm/day) in the CNN1-SS had to be excluded from this analysis. This suggests that CNN1-MS regularizes the network by training simultaneously to all the target sites, therefore avoiding the potential instabilities that may appear in single-site topologies. The relative biases for the mean are also better for the CNN1-MS than for the GLM4(CNN1-SS), which present positive(negative) values.

This experiment indicate that, whereas single-site CNNs are prone to overfitting in certain locations, the equivalent multi-site topologies perform an implicitly regularization which allows the network to treat simultaneously the high-dimensional predictor space, avoiding overfitting and leading to improved forecast accuracy.

6.2.2 Automatic Feature Selection

This section digs into the “black-box” nature of DL topologies by providing insight into the predictor-predictand link in a climate downscaling application. This is done in two different ways. In the first part of the section we show spatial representations of the last filter map in the CNN1 model and compare it against the different predictor fields for a particular day. Also, we show the coefficients which link this last filter map to two illustrative locations with different climate regimes: Madrid and Helsinki. In the second part of the section we extend this naive study to other locations distributed along Europe, and base the interpretation of the predictor-predictand link on Prediction Difference Analysis (PDA[3], (Zintgraf et al., 2017)).

Following the methodological framework described in Chapter 5, we downscale precipitation occurrence over the 86 stations considered in the VALUE’s Experiment 1a building on ERA-Interim 2° large-scale variables. To assess the influence of low- and high-dimensional input spaces, we build two different versions of the CNN1 model which differ on the predictor setup used (see Figure 6.7). In particular, whereas CNN1(20) was

[3] PDA is a mathematical formulation that permits to marginalize the influence that each predictor variable has on the model outputs.

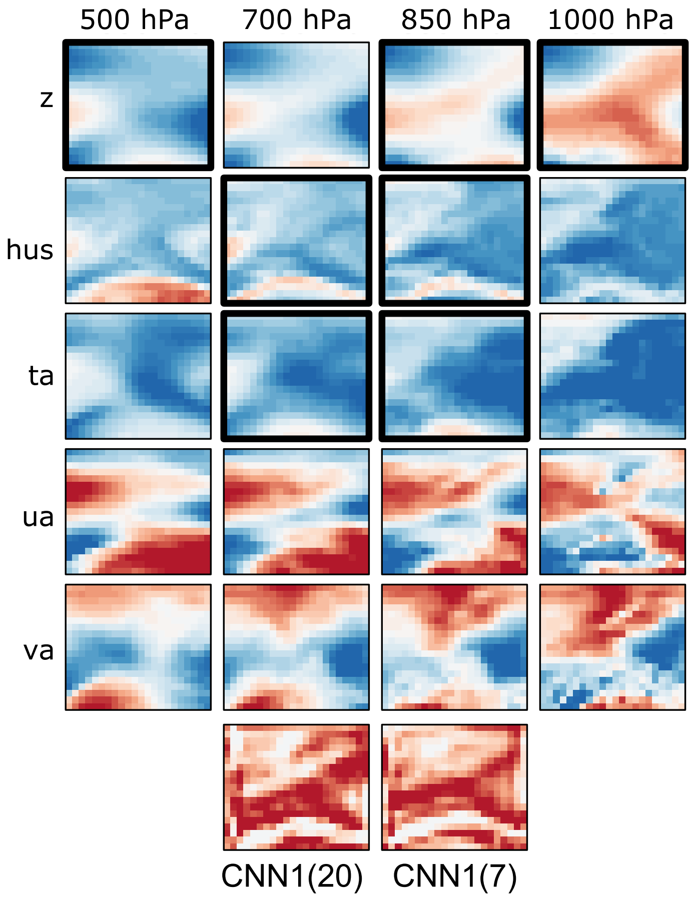


Figure 6.7: The twenty predictors listed in Table 5.2 (rows 1-5) for a given day —randomly selected— plus the resulting features in the last convolution layer (bottom row) for the CNN1(7) and CNN1(20) models. Whereas CNN1(20) was trained with the full predictor set, only the seven variables marked with a black frame were considered in CNN1(7).

trained with the full predictor set, only the seven variables marked with a black frame —coinciding with those used in VALUE's experiment 1 to build regression-based models (Gutiérrez et al., 2019),— were considered in CNN1(7).

Figure 6.7 shows the atmospheric situation as described by the twenty predictor variables listed in Table 5.2 for a given day (selected randomly), plus the last filter map in the two versions of the CNN1 model —CNN1(7), CNN1(20).— The goal is to search for similarities between these predictor fields and the last feature map, which allows for visualizing the implicit predictor selection that occurred within the hidden layers of the network. Visual inspection reveals that humidity at intermediate height levels (700 and 850 hPa) resembles the bottom part of the pattern found in the last feature map for both CNN1(20) and CNN1(7), which suggest that these are the variables used by the network to downscale precipitation in this area. The same occurs for geopotential height at 850 and 1000 hPa over the central and top parts of the studied domain. The robustness of the model to high-dimensional spaces is proved, since both CNN(20) and CNN(7) present very similar patterns in their last feature map. The main difference between these two maps appears in the top-right corner of the domain, which seems to be dominated by zonal wind velocity —not present in the CNN1(7) predictor configuration.— This analysis shows how variables which are not relevant (or play a minor role) for downscaling, such as meridional wind velocity, are mostly neglected by the network without altering the predictive skill. The results derived from this study are consistent with previous works which suggest that humidity and geopotential are the most useful variables to explain the local variability of precipitation in Europe (Timbal and McAvaney (2001), San-Martín et al. (2017)).

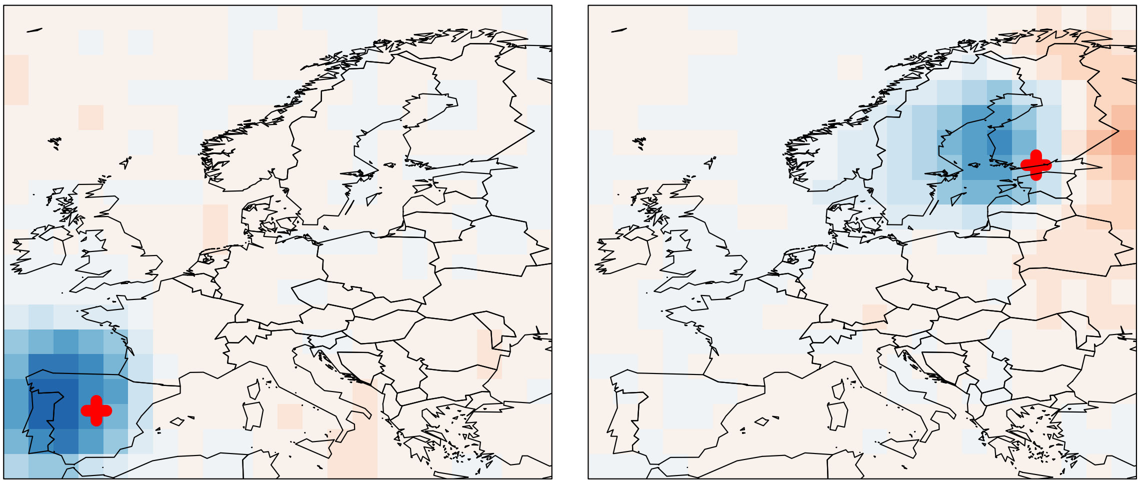


Figure 6.8: Weights connecting the last convolution layer to the output neurons in Madrid (left) and Helsinki (right). A 5x5 spatial moving average is applied to represent the effect of kernels. Blue/red colors indicate positive/negative weights.

Figure 6.8 shows the coefficients (i.e. the weights) that link the last filter map to two illustrative locations which were selected for presenting very different climates: Madrid and Helsinki. Since we have applied padding to the convolutional layers in the CNN1 model, the last feature map has the same latitude-longitude dimensions than the input variables, which allows to visualize the importance of the predictor set for downscaling across the entire domain[4]. Noticeably, the largest weights are found over an area of approximately 5x5 gridboxes surrounding the location of interest, with (quasi) zeroed-values elsewhere. This indicates that the network automatically neglects the regions which are not of interest for downscaling, taking advantage thus of site-dependent windows of information.

[4] In computer vision, this type of representation is referred to as saliency maps.

To gain further insight into the conclusions drawn from this naive analysis, it is next extended by considering a more sophisticated way to produce saliency maps, namely Prediction Difference Analysis (PDA). PDA allows to directly estimate the relevance of each predictor variable for downscaling —to a certain extent, this aspect was qualitatively assessed in Figure 6.8 but without quantitatively evaluate the degree of participation of each predictor variable.— Moreover, here we focus not only on the prediction of precipitation occurrence but also in rainfall amount.

Following from the mathematical formulation of a NN described in Eq.4.2, PDA estimates the relevance of an input feature, x_j, by measuring how the predicted parameters, (i.e., $y = \{p, \alpha, \beta\}$ for precipitation or $y = \{\mu, \sigma^2\}$ for temperature) change when it is unknown, (i.e., y' = $\{p', \alpha', \beta'\}$ for precipitation or y' = $\{\mu', \sigma^{2\prime}\}$ for temperature). This can be done by marginalizing the j feature:

$$y' = f^{\omega}(x_{\setminus j}) = \sum_{x_j} P(x_j|x_{\setminus j}) f^{\omega}(x_j, x_{\setminus j}) \tag{6.1}$$

Where $x_{\setminus j}$ refers to the complete set of input features except x_j. As Zintgraf et al. (2017), we adjust the probability function $P(x_j|x_{\setminus j})$ with a conditional multivariate normal distribution from which M predictor configurations are sampled. These are feed-forwarded through the network, being the output y' the average of these M realizations.

Approximating $P(x_j|x_{\setminus j})$ is usually not feasible so these authors simplify the term by conditioning only on a surrounding region of size LxL, described by $\hat{x}_{\setminus j}$. Moreover, a multivariate analysis can be carried out by removing jointly a set of z features grouping a patch of size KxK. Hence,

$$y' = f^{\omega}(x_{\setminus z}) = \sum_{x_z} P(x_z|\hat{x}_{\setminus z}) f^{\omega}(x_z, x_{\setminus z}) \tag{6.2}$$

Figure 6.9 shows an schematic representation of the PDA technique. For a given predictand site, given N daily samples and the full set of predictor variables listed in Table 5.2, we obtain saliency maps of dimension Nx19x22x20 (time-latitude-longitude-variable). For every sample i, each pixel in these maps represents the Activation Difference (AD) of the expectance of the predictive distributions (i.e., Bernoulli-Gamma and Gaussian for precipitation and temperature models, respectively):

$$AD_i = \mu_i' - \mu_i \tag{6.3}$$

$$AD_i = p_i'\alpha_i'\beta_i' - p_i\alpha_i\beta_i \tag{6.4}$$

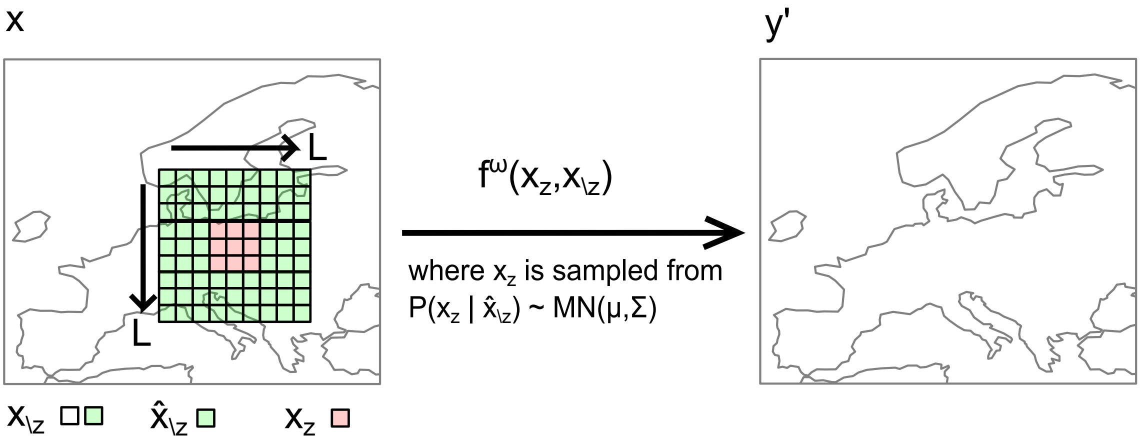

x
y'
L
L
$f^{\omega}(x_z, x_{\setminus z})$
where x_z is sampled from
$P(x_z \mid \hat{x}_{\setminus z}) \sim MN(\mu, \Sigma)$
$x_{\setminus z}$ ⬜🟩
$\hat{x}_{\setminus z}$ 🟩
x_z 🟥

Figure 6.9: Diagram illustrating the marginalization of x_z, of size KxK (red), conditioned on its surrounding region of size LxL, $\hat{x}_{\setminus z}$ (green), given an input feature channel x —note this particular case corresponds to one single feature map, $K = 3$ and $L = 9$.— The input features, x_z and $x_{\setminus z}$, are feed-forwarded to the neural network f^{ω}, where x_z is drawn from the Multivariate Normal (MN) distribution $P(x_z|\hat{x}_{\setminus z})$, described by μ (a z-variate vector of means) and Σ (the covariance matrix), which are the estimated conditional parameters. This process is repeated by successively centering x_z in the rest of predictor gridboxes.

With illustrative purposes, we focus here on four locations —actually their closest gridboxes in the E-OBS grids— corresponding to different climates regimes: Paris, Rome, Cophenaguen and a point in the Alps. The objective is to gain interpretability about the internal functioning of CNN1 and CNN10, which were found to be the best-performing models for downscaling in “perfect” conditions of precipitation and temperature, respectively (see section 6.1). Building on this idea, we applied the PDA technique on the conditional daily predictions learnt with these models for the year 2008 —we limited the study to a single year of the test set due to computational limitations.— For each predictand variable, the results are 365 saliency maps of dimension 19x22x20x4 (latitude-longitude-variable-site) providing the spatial distribution of the AD. We used $K = 3$, $L = 11$, and $M = 30$ for the PDA technique, which proved to be a good trade-off between computational requirements and representativeness in the saliency maps.

Figure 6.10 shows the saliency maps obtained for each target location, averaged across the time dimension —to avoid error compensation, absolute AD values are considered.— For each location, results for the twenty predictor variables available is presented. In agreement with the results found in the first part of this section, the specific humidity at 1000 hPa is found to play a key role for precipitation downscaling, finding strong signals in their surrounding areas for all locations except in the Alps. Nevertheless, different importance patterns emerge for other predictor variables at different sites. For instance, the wind velocities appear to be quite informative to precipitation in the Alps, —especially at 700 and 850 hPa— whilst having only relative influence over Paris and Rome. Also, the geopotential at 1000 hPa presents (low) high AD values for (Alps and Cophenaguen) Paris and Rome. Finally, Cophenaguen is the only location where air temperature at 1000 hPa is detected to have a strong influence for the downscaling of local precipitation.

To our knowledge, a systematic study about the importance of the different predictor variables typically used for SD of precipitation has not been been undertaken yet. Indeed, just a few works addressing this topic can be found in the literature. For instance, Soares et al. (2019) evaluated the importance of some large-scale mechanisms (e.g., the North Atlantic Oscillation) for downscaling of precipitation over Europe, and Yang et al. (2018) used a step-wise algorithm and partial correlations to search for relevant variables for the same task over China. Other studies perform an exhaustive screening of predictors and domains to find the best model setup (see San-Martín et al. (2017) for an example in Spain). Despite this lack of studies makes difficult to properly contextualize the results from Figure 6.10, the high AD values found for humidity are consistent with previous literature (San-Martín et al., 2017) and physical principles. Moreover, the spatial extent of the importance patterns revealed by Figure 6.10 are also in agreement with previous works which suggest that an area of 5x5 gridboxes centered around the location of interest is enough to retain most of the informative power that is needed to reproduce the observed local variability (Timbal and McAvaney, 2001; Timbal et al., 2003; Gutiérrez et al., 2004; Brands et al., 2011b; Gutiérrez et al., 2013; San-Martín et al., 2017). Furthermore, this conclusion is also consistent with the results described in section 6.1 and with the saliency maps displayed in Figure 6.8.

Figure 6.11 is the equivalent to Figure 6.10 but for temperature. Unlike for precipitation, all predictor variables except air temperature at 1000 hPa present very low AD values, suggesting that most of the informativeness required to downscale local surface temperature is already provided by the large-scale near-surface temperature —note that even air temperatures at other vertical levels exhibit a nearly negligible influence.— This result is in agreement with previous studies (Huth, 1999, 2002, 2004).

Overall, the analyses presented in this section allow to better understand the internal functioning of the CNNs used for climate downscaling in this Thesis.

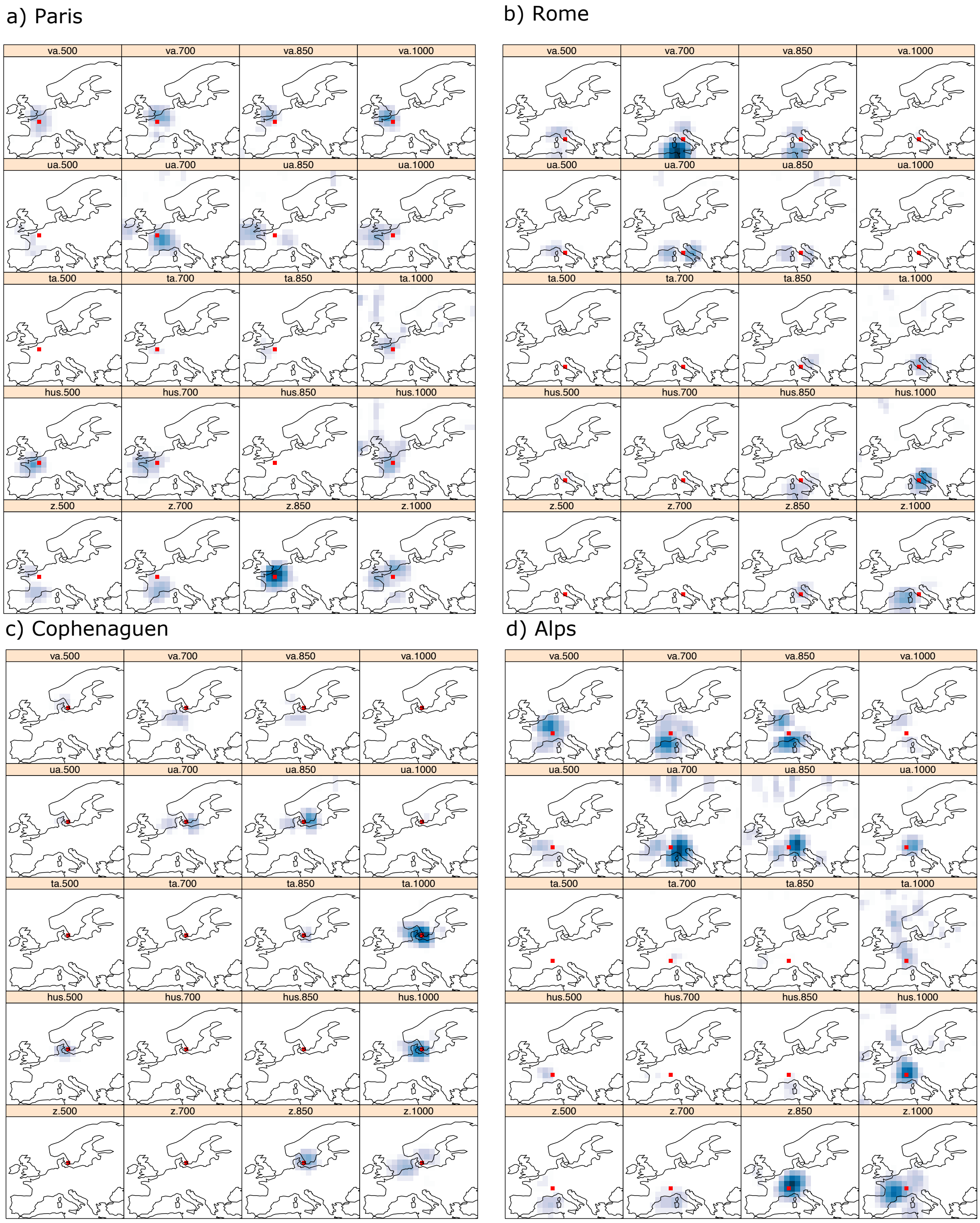
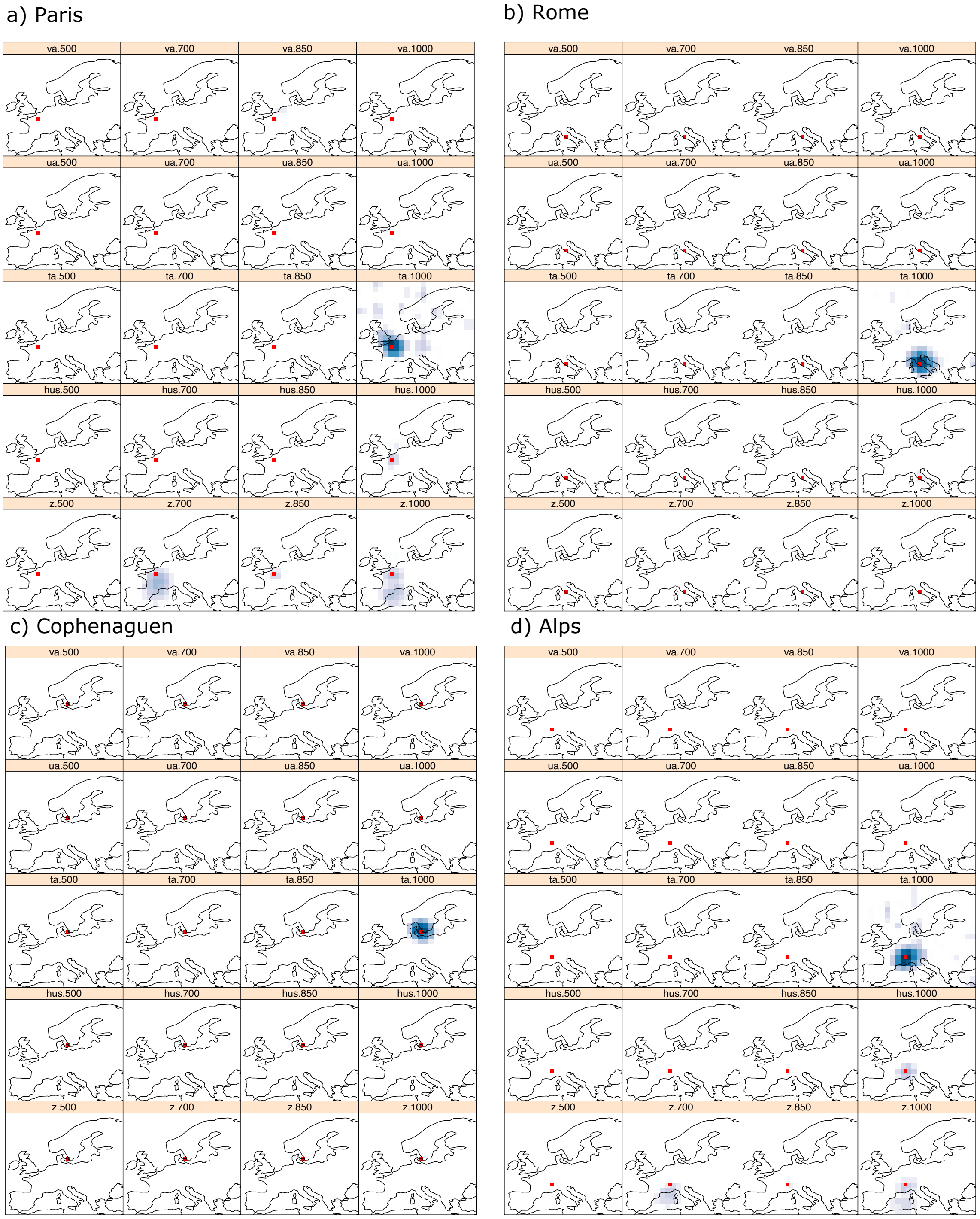


Figure 6.10: Saliency maps obtained for the CNN1 model at four illustrative locations: Paris, Rome, Cophenaguen and Madrid. For each location, results for the meridional (va) and zonal (ua) wind velocities, air temperature (ta), specific humidity (hus) and geopotential (z) at 500, 700, 850 and 1000 hPa are given. To avoid error compensation, the maps show the absolute AD values, averaged across the time dimension (for the year 2008).


a) Paris

b) Rome

c) Cophenaguen

d) Alps

va.500 va.700 va.850 va.1000
ua.500 ua.700 ua.850 ua.1000
ta.500 ta.700 ta.850 ta.1000
hus.500 hus.700 hus.850 hus.1000
z.500 z.700 z.850 z.1000

Figure 6.11: As Figure 6.10 but for temperature.

CHAPTER 7

Downscaling from Global Climate Models

This chapter is formed by two sections which allow to comprehensively assess the suitability of SD methods —with special focus on the best-performing CNNs found in the “perfect” conditions experiment (Chapter 6)— to downscale GCM simulations.

First, section 7.1 focuses on the use of CNNs (and GLMs) to downscale the 12^{th} run of the EC-Earth model to the target E-OBS resolution (0.5°), producing daily precipitation and temperature fields over Europe for both historical (1979-2008) and RCP8.5 (far-future: 2071-2100) scenarios. For the former, the downscaled fields are directly validated against E-OBS. However, for RCP85, since no observational reference exists for the future, the raw outputs from the EC-Earth —interpolated to the target 0.5° grid— are used as “pseudo-reality” to “validate” the downscaled projections. This approach, which has been widely used in the literature (Vrac et al., 2007b; Gutiérrez et al., 2013; San-Martín et al., 2017; Quesada-Chacón et al., 2021), builds on the idea that significant deviations from the driving GCM could be an indicator of the implausibility of the SD-based projections (unless it is justified by process understanding).

Second, section 7.2 extends section 7.1 by using our CNNs to downscale the subset of CMIP5 models listed in Table 5.4 to produce *DeepESD*, an ensemble of high-resolution projections of daily precipitation and temperature over Europe for the entire 21^{st} century (2005-2100). The objective of this dataset —which represents the first of its kind at a continental scale,— is to be used as reference by the community for the research of crucial aspects of climate statistical downscaling (e.g., stationarity assumption), with views to further explore the use of this type of projections in real impact studies. In this case, in addition to the driving GCMs, a subset of RCMs from EURO-CORDEX are also used as “pseudo-reality” to “validate” the plausibility of the downscaled projections. The reason for this is that, as compared to their driving GCMs, RCMs may significantly alter the climate change signal at the regional-to-local level, since they explicitly resolve small-scale mechanisms which are not taken into account by GCMs (Giorgi and Gutowski Jr, 2015; Giorgi et al., 2016; Sørland et al., 2018). Nevertheless, deciding whether or not RCMs are able to provide more plausible climate change scenarios than GCMs is nowadays a hot research topic. In addition to the increase in spatial resolution, other factors such as inconsistencies among the parameterization schemes used (Pinto et al., 2018), the absence of ocean coupling in the RCM formulation (Gaertner et al., 2018; Akhtar et al., 2018), and the inclusion (or not) of time-varying anthropogenic aerosols, may also lead to large differences between the RCMs and their driving GCMs (Gutiérrez et al., 2020; Boé et al., 2020). All these aspects make difficult the election of either GCMs or RCMs as the "pseudo-reality" our SD-based projections should be compared to. For this reason, the plausibility of *DeepESD* is studied here based on both GCMs and RCMs.

The first part of this Chapter (section 7.1) is based on the manuscript entitled *"On the suitability of deep convolutional neural networks for downscaling climate change projections"*, published in *Climate Dynamics*. The second part (section 7.2) is based on another paper, *"DeepESD: An ensemble of regional climate change projections over Europe based on deep learning downscaling"*, which is currently under review in *Nature Scientific Data*.

7.1 Assessing the Suitability of Deep Learning Models to Downscale Historical and Future Climate Simulations

In this section we explore the suitability of CNNs to downscale the historical (1979-2008) and RCP8.5 (2071-2100) scenarios from a single GCM. We frame the study in EURO-CORDEX ESD, and therefore we focus on the 12^{th} run of the EC-Earth. To do this, we build on the best-performing CNNs found for downscaling in "perfect" conditions (section 6.1): CNN10 for temperature and CNN1 for precipitation. For completeness, the GLMs introduced in section 5.5 (GLM1, GLM4 and GLMPC) are also considered. All these SD methods are first trained based on the ERA-Interim predictors used in Chapter 6 during 1979-2008. Then, they are applied to the EC-Earth predictors —note that the distributional similarity between ERA-Interim and EC-Earth predictors was assessed in section 5.3.— The resulting downscaled fields are validated in terms of different metrics depending on the scenario of interest, either the historical (focusing on 1979-2008) or the RCP8.5 (focusing on the far-future: 2071-2100). In particular, for the historical scenario we compute the (relative) biases for the indicators listed in Table 5.6, taking as reference the E-OBS (precipitation) temperature. With respect to the RCP8.5 scenario, we focus on the projected "delta" changes (i.e. the mean difference between the RCP8.5 and the historical fields) provided by the different SD methods for the same indicators. Based on previous literature (Vrac et al., 2007b; Gutiérrez et al., 2013; San-Martín et al., 2017; Quesada-Chacón et al., 2021), these changes are compared with those obtained from the EC-Earth's raw simulations, which are considered as "pseudo-reality".

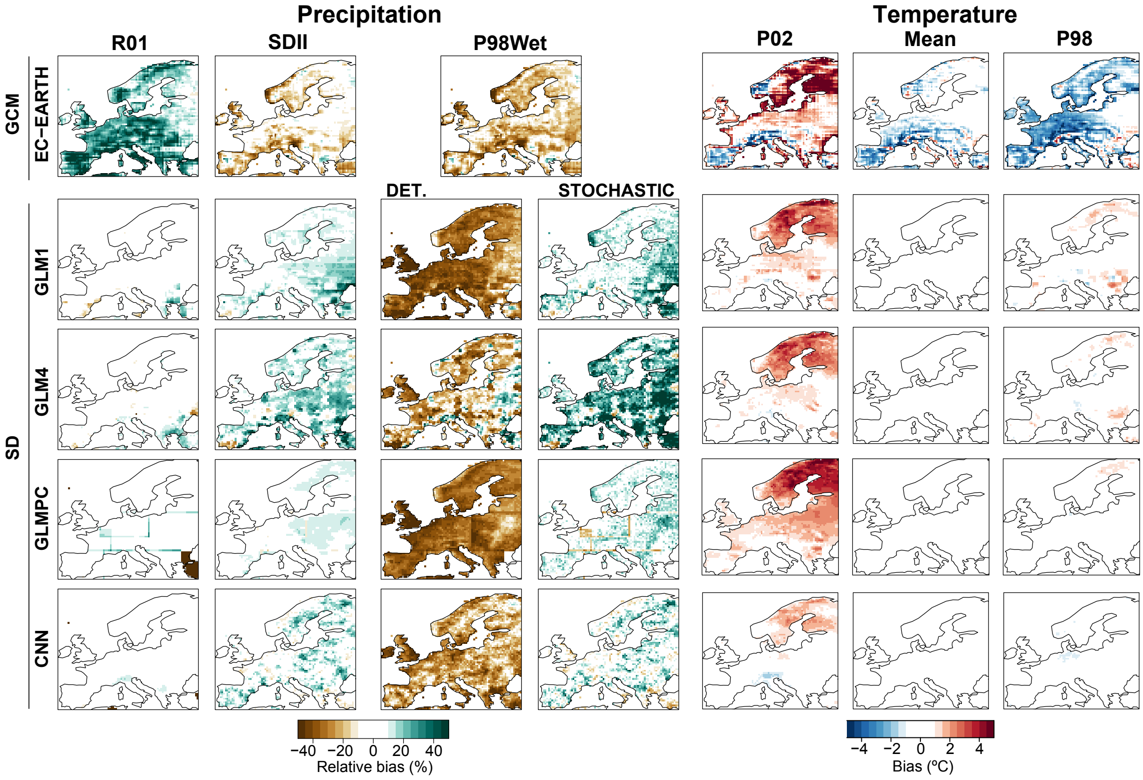


Figure 7.1: (Relative) biases of the downscaled fields in the historical scenario for (R01, SDII and P98Wet) P02, the mean and P98, which are computed taking E-OBS (precipitation) temperature as reference. In rows (from top to bottom), results for the EC-Earth, GLM1, GLM4, GLMPC and CNN (CNN1 for precipitation and CNN10 for temperature) are given. For P98Wet, deterministic and stochastic predictions are considered.

Figure 7.1 shows the (relative) biases of the downscaled fields in the historical scenario for (R01, SDII and P98Wet) P02, the mean and P98, which are computed taking E-OBS (precipitation) temperature as reference. In rows (from top to bottom), results for the EC-Earth, GLM1, GLM4, GLMPC and CNN (recall, CNN1 for precipitation and CNN10 for temperature) are given.

The EC-Earth shows systematic errors across vast regions of the continent for all the metrics considered. For precipitation, the R01 is overestimated —especially in Southern and Western Europe— whilst the SDII (P98Wet) presents small (high) positive biases over large areas, especially in mountainous regions. The overestimation of the R01 can be explained by the so-called "drizzle effect"[1] (see Dai (2006) and references therein). Likewise, the underestimation of rainfall amount (as described by SDII and P98) could be due to the miss-representation in the model of certain atmospheric phenomena which trigger local precipitation (e.g., orographic convection). For temperature, a gradient of negative-positive biases are found for P02 from South to North. Differently, negative biases are generally found for the mean and the P98, especially in Central Europe and the Iberian Peninsula. Again, these errors might be attributed to the limited spatial resolution of EC-Earth, which have as a result the misrepresentation of important local features such as complex orography, land-sea contrasts, etc. (Manzanas et al., 2018)

Differently to the EC-Earth's raw simulations, the GLMs and CNNs exhibit in general lower biases for all the metrics considered. However, before analyzing the performance of these SD methods for downscaling the EC-Earth's historical scenario, it is important to recall that all of them showed nearly negligible biases across the entire Europe in the "perfect" conditions experiment (Chapter 6). Having this in mind, the biases found in Figure 7.1 for the GLMs and CNNs might presumably be caused by violations of the PP assumption of having GCM predictors which are reasonably similar to their counterpart variables in the reanalysis.

For R01, all SD methods yield satisfactory results in the historical scenario, with minor positive biases for GLM1 and GLM4 over the Balkans.

Moreover, for the SDII, these two local-based GLMs present positive biases over large regions. As already argued, these errors are presumably due to some inconsistency between ERA-Interim and EC-Earth predictors over those regions. However, the GLMPC and the CNN (especially the former) exhibit very small biases for these two metrics since the "problematic" predictors may be not represented in the leading PCs, or directly neglected in the hidden layers of the CNN. For P98Wet, we show the resulting biases both for deterministic and stochastic versions of the SD models considered. In agreement with the results from Figure 6.4, deterministic predictions underestimate clearly this indicator. Nonetheless, stochastic predictions allow to better reproduce the local extremes. In particular, the local-based GLMs (especially GLM4) exhibit strong (positive) biases for P98Wet while the GLMPC and the CNN present small (positive) biases in some scattered regions across the continent.

With regards to temperature, Figure 7.1 shows nearly null biases for the mean and the P98 for all the SD models considered. Nevertheless, in agreement with the results from Figure 6.2 positive biases are found for P02 over Eastern Europe and Scandinavia. These errors are less pronounced for the CNN than for the GLMs. Contrary to what

[1] Overly frequent drizzle is a persistent problem in climate models which arises from the use of convective parametrization schemes that tend to trigger precipitation too easily.

happened for local GLMs (in particular for GLM4) in the case of precipitation, the errors found for P02 in temperature for GLM1 and GLM4 are lower than for GLMPC, and only slightly larger than for CNN (which provides the best results). Since the predictor set considered is the same for downscaling of both temperature and precipitation, we argue that the predictor variables responsible for the biases found in the SDII and P98Wet (stochastic) maps —in the precipitation panel— would not be driving local temperature. Indeed, recall from section 6.2.2 that most of the informative power required to explain local surface temperature comes solely from large-scale near-surface temperature.

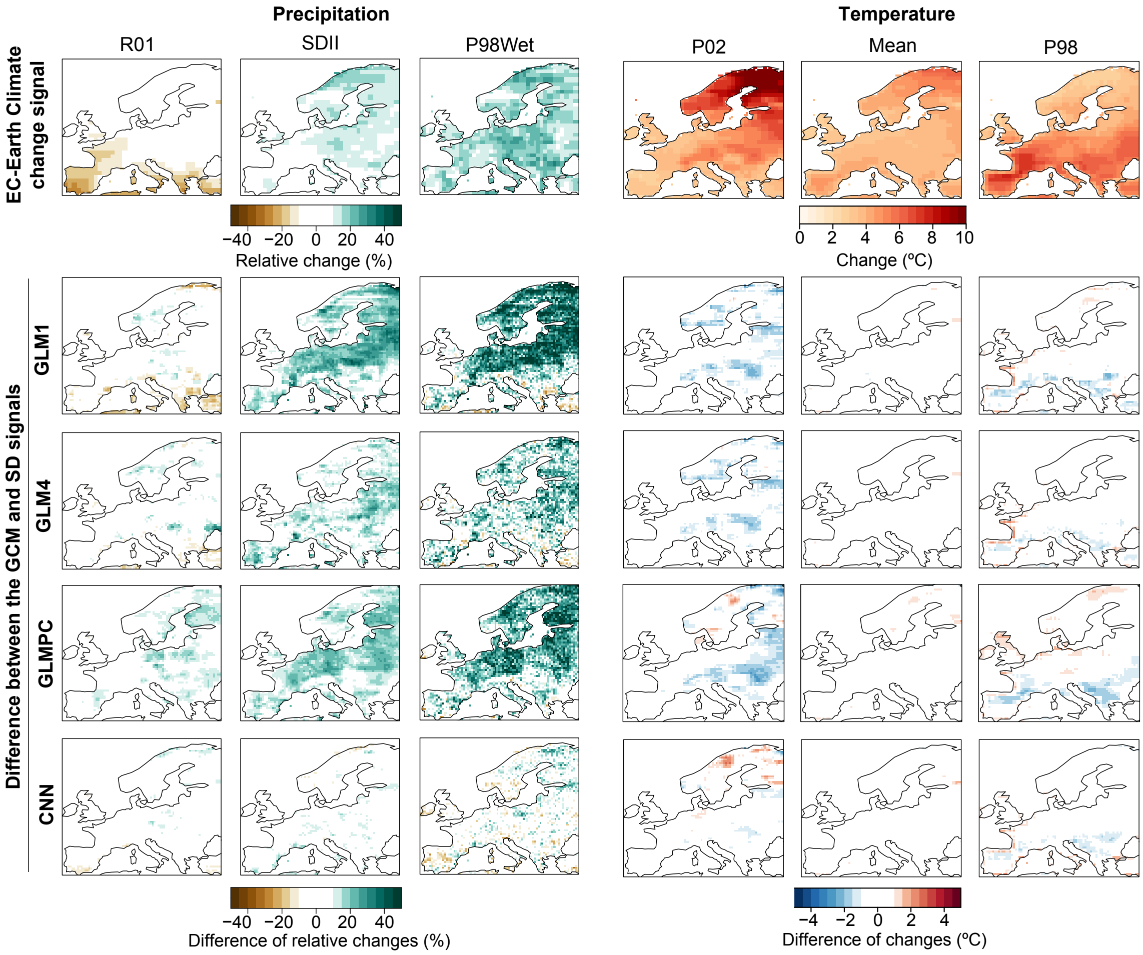


Figure 7.2: First row: (Relative) "delta" changes projected by the EC-Earth's raw outputs —interpolated to the target 0.5° grid— for (R01, SDII and P98Wet) P02, the mean and P98. These changes are computed based on the mean difference between the RCP8.5 (2071-2100) and the historical (1979-2008) fields. Rows 2-5: Difference between the changes projected by GLM1, GLM4, GLMPC and CNN (CNN1 for precipitation and CNN10 for temperature) and those shown in the first row for EC-Earth —which are considered as "pseudo-reality".—

To "validate" the plausibility of the downscaled projections in the RCP8.5 scenario we compute the "delta" changes derived from the different SD methods and compare them with those directly obtained from the EC-Earth's raw outputs —interpolated to our target 0.5° grid.— In this line, the first row of Figure 7.2[2] shows the (relative) "delta" changes obtained from EC-Earth for (R01, SDII and P98Wet) P02, the mean and P98. Rows 2-5 correspond to the GLM1, GLM4, GLMPC and CNN (recall, CNN1 for precipitation and CNN10 for temperature), and display the difference between the changes projected by these methods and those from EC-Earth (which are considered as "pseudo-reality"), shown in the first row. Based on previous works (Vrac et al., 2007b; Gutiérrez et al., 2013; San-Martín et al., 2017; Quesada-Chacón et al., 2021), we argue that large (small) differences between the first row and rows 2-5 might be an indicator of bad (good) extrapolation capability for the corresponding SD method.

For R01, EC-Earth exhibits almost no changes except for the Mediterranean arch, where a decrease (in between -10% and -20%) in the number of rainy days is projected. Nonetheless, the SDII and the P98Wet (especially the latter) are expected to increase considerably (by about 20-40%) over most parts of the continent. This indicates that a fewer number of rainy days would bring larger rainfall amounts in the far-future (with respect to the historical period). For the case of temperature, positive changes (i.e. increases) are projected for all the indicators considered. In particular, a warming of about 2-5°C is expected for mean temperature across the entire continent, reaching 6-10°C for the case of P02 over Scandinavia. Similarly, increases of about 5-7°C are projected for P98 over France and Southern Europe. These results are in agreement with the literature (Giorgi and Lionello, 2008; Terray and Boé, 2013; Collins et al., 2013).

For R01, the changes projected by the SD methods show very low differences with respect to those obtained from EC-Earth. A similar situation is also found for the P02, the mean and the P98 for the case of temperature. Moreover, the minor deviations found with respect to the EC-Earth's signals (in between -15 and 20% for R01; -2 and 2°C for the P02 and P98) appear mostly for the two GLMs, particularly in some regions over Central and Northeastern Europe. Furthermore, for SDII and P98Wet, large positive differences are also encountered for the two GLMs over vast parts of the continent. Nevertheless, these differences are much lower for CNN, and restricted to small scattered locations. The "overestimation" of future precipitation for the GLMs —regardless of the predictor configuration used— has also been reported in other studies (see San-Martín et al. (2017) for an example in Spain). The latter cannot be justified by any known physical mechanism,

[2] The colorbars used in this figure are designed to ease the visualization of the differences between the "delta" changes provided by the different SD methods and those given by EC-Earth's raw outputs. Note that providing a detailed description of the projected climate change signals is not the scope of this study.

and could be attributed to a lack of extrapolation power for future climates based on the (linear) predictor-predictand link learned in present-climate conditions. In this regard, building on the non-linearities learnt by the network, the CNN model projects a picture of change for local precipitation which is broadly consistent with the one given by the EC-Earth. For the case of temperature, we had already seen that the predictor-predictand link is mostly linear (see section 6.1) and therefore the CNN and the GLMs yield similar "delta" changes. This result is consistent with previous works which have demonstrated that regression-based methods provide good extrapolation capabilities for the downscaling of anomalously warm temperatures which have not been seen during the calibration period (Gutiérrez et al., 2013).

Overall, the SD methods analyzed allow to reduce the systematic biases exhibited by the raw simulations from the EC-Earth in the historical period. Nevertheless, some of the variables here included in the predictor set seem to violate the PP assumption of being well simulated by the GCM, which leads to notable biases for SDII and P98Wet in local GLMs (GLM1 and GLM4). This undesired effect diminishes when the predictor space is adequately manipulated, making use of either PCs (GLMPC) or convolutional operations (CNN10 for temperature and CNN1 for precipitation). Moreover, as compared to GLMs, when the CNNs are used to downscale the RCP8.5 scenario (for the far future 2071-2100), they lead to patterns of change which are notoriously more compatible with those projected by the EC-Earth's raw outputs —considered as "pseudo-reality"— for all the metrics analyzed, and especially for those related to rainfall amount. Therefore, this section demonstrates that CNNs provide a good alternative (in particular better than GLMs) to downscale climate change scenarios, particularly for precipitation.

7.2 Building an Ensemble of Regional Climate Change Projections for Europe Based on Deep Learning

In this section we extend the analysis performed in the preceding one —which focuses exclusively on EC-Earth— by using our CNNs to downscale the subset of CMIP5 models listed in Table 3.1. As a result, we produce *DeepESD*, the first ensemble of high-resolution (0.5°) climate change projections for the 21st century (1975-2100) over the entire Europe based on DL models —CNN10 for temperature and CNN1 for precipitation.— *DeepESD* is publicly available from the Earth System Grid Federation (ESGF) node at the University of Cantabria[3].

To produce *DeepESD*, the CNN10 and CNN1 models were trained based on ERA-

[3] https://data.meteo.unican.es/thredds/catalog/esgcet/collections/CORDEX-DeepESD-EE/catalog.html

Interim and E-OBS (as predictor and predictand datasets, respectively) during the period 1979-2008. Once trained, the two CNNs were applied to downscale the historical (1975-2005) and RCP8.5 (2006-2100) scenarios of the eight CMIP5 GCMs shown in Table 3.1. A previous study by Brands et al. (2013) has demonstrated that CMIP5 exhibits good skill to reproduce the key large-scale circulation and thermodynamics over Europe —as represented by reanalysis data— once the seasonal mean is removed from the time-series. Therefore, for this experiment, we post-process all the GCM predictor variables considered according to the two-step process (harmonization+standardization) described in section 5.3). The only difference with respect to Chapter 5 is that here we have removed from the predictor set all the variables at 1000 hPa since this vertical level was not available for all the GCMs listed in Table 3.1.

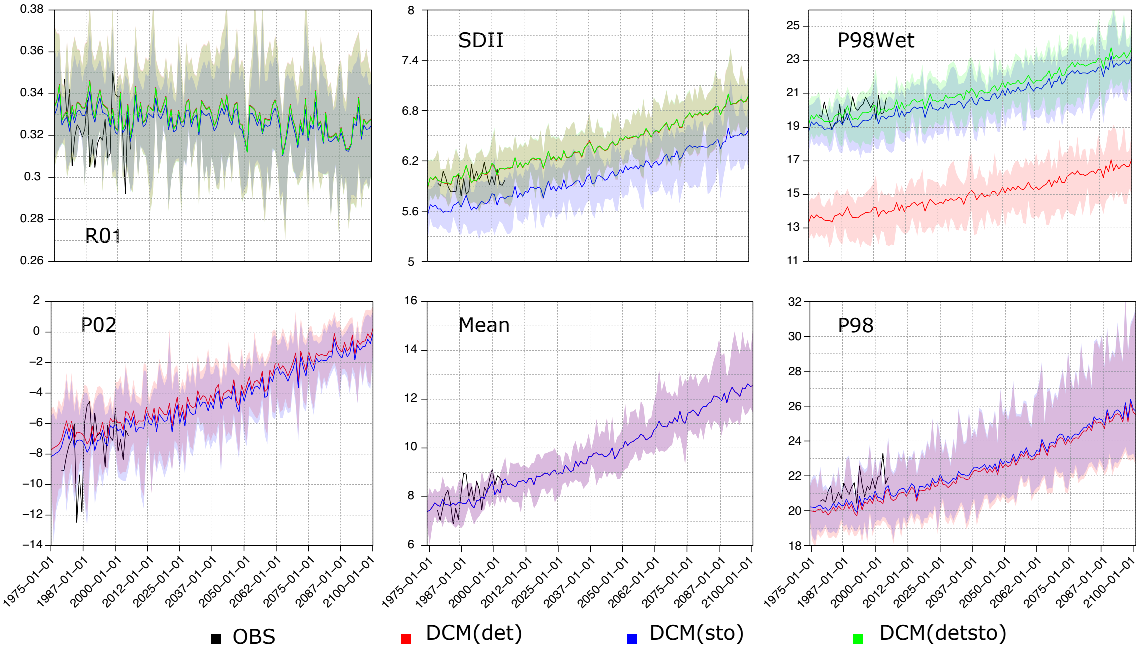


Figure 7.3: Inter-annual time-series —spatially averaged over the entire Europe— for R01, SDII, P98Wet (precipitation) and P02, the mean and P98 (temperature), as obtained from deterministic, stochastic and deterministic-stochastic projections —see the colors in the legend— based on our CNNs (CNN1 for precipitation and CNN10 for temperature; see the text for details). In all cases, solid lines correspond to the multi-model ensemble mean whilst the shadows encompass the eight downscaled GCMs. For comparison purposes, the black lines show the results obtained for E-OBS.

To better understand the trade-off between the gain in prediction accuracy attained by deterministic methods and the loss of spatio-temporal structure of stochastic ones, we start by producing both deterministic and stochastic projections for CNN10 and CNN1 as described in section 4.4. Figure 7.3 allows to examine the differences that emerge in the downscaled fields when using 1) deterministic projections, 2) stochastic projections and 3) for the particular case of precipitation, a combination of deterministic projections for rainfall occurrence and stochastic ones for rainfall amount. In particular, this is done here by looking at the inter-annual time-series for R01, the SDII and the P98Wet (the P02, the mean and the P98) for precipitation (temperature). In all cases, interannual time-series for the spatially averaged —across the whole Europe— indicators are shown. Whilst the solid lines correspond to the multi-model ensemble mean, the shadows encompass the eight downscaled GCMs. For comparison purposes, the E-OBS reference values for the period 1979-2005 are also given.

Beyond the emerging future trends (which will be later analyzed in more detail), it is important to note that deterministic-stochastic (stochastic) projection yield nearly unbiased results for the historical period for all the precipitation (temperature) indicators considered.

For temperature, deterministic and stochastic implementations provide very similar results in all cases. Moreover, all the indicators analyzed are projected to increase, which is consistent with previous studies assessing the future warming signals over Europe (Giorgi et al., 2016; Sørland et al., 2018; Boé et al., 2020).

For the case of precipitation, stochastic projections underestimate slightly (as compared to E-OBS for the 1979-2005 period) the SDII and P98Wet. This is explained by the small fraction of wet days which are incorrectly given as dry ones when precipitation occurrence is stochastized. Nevertheless, and more importantly, deterministic projections underestimate clearly the P98Wet, which is in agreement with the results from Figures 6.4 and 7.1. As argued in section 6.1, this is presumably due to a lack of informativeness in the large-scale predictors considered to explain the variability of local precipitation.

Taking into account that SDII and P98Wet are two key indicators, the previous results suggest the importance of counting on stochastic projections of rainfall amount, which can be easily produced by sampling out from the daily conditional Gamma distributions learnt by the CNN —for rainfall occurrence, deterministic values should be considered.— However, since the sampling is done at a gridbox-level from univariate distributions, introducing this stochastic component implies a certain loss of spatio-temporal structure in the downscaled fields. This situation is illustrated by Figure 7.4, which shows, for a given day (21-February-1975), the precipitation (top row) and temperature (bottom row) fields provided by the CanESM2's raw outputs (left column), together with the corresponding downscaled fields. In particular, the middle (right) column corresponds to the deterministic (stochastic) implementation of CNN1 and CNN10 (for precipitation and temperature, respectively). Note that for the precipitation fields, the rainfall occurrence

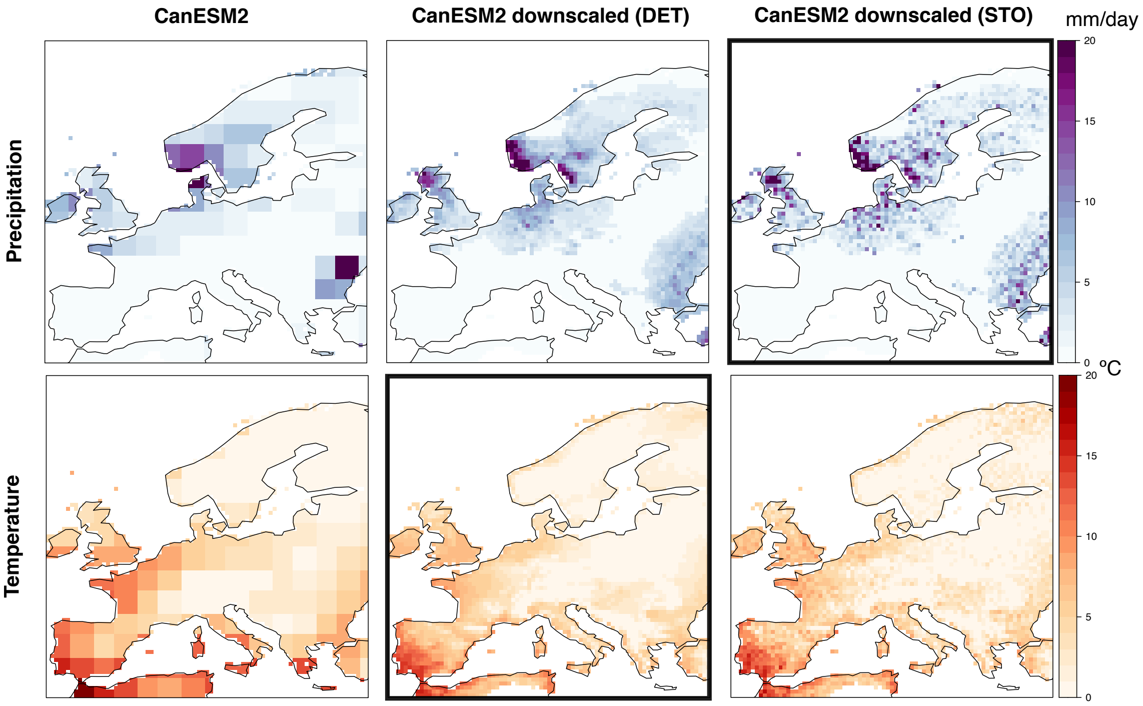


Figure 7.4: Precipitation (top row) and temperature (bottom row) fields for 21-February-1975, as given by the CanESM2's raw outputs (left column), together with the corresponding downscaled values. In particular, the middle (right) column corresponds to the deterministic (stochastic) implementation of CNN1 and CNN10 (for precipitation and temperature, respectively). Note that for the precipitation fields, the rainfall occurrence has been computed in a deterministic manner for both middle and right columns. The black frames identify the implementation finally used to produce *DeepESD*.

has been computed in a deterministic manner for both middle and right columns.

Beyond the obvious gain in spatial detail achieved by the CNN models, this figure shows that deterministic projections yields smooth spatial patterns which keep a good correspondence with the driving GCM. However, stochastic projections give place to more discontinuous spatial patterns whilst provide higher values than deterministic ones for some particular gridboxes (this is especially evident for precipitation).

Based on the results from Figures 7.3 and 7.4, and taking into account that extreme events are expected to considerably impact a large number of socio-economic activities in a changing climate, we decided to use the deterministic-stochastic implementation of the CNN1 method to produce the high-resolution precipitation fields delivered with *DeepESD*, even at cost of losing some spatio-temporal consistency at the daily scale. For temperature, however, the deterministic version of the CNN10 was considered to build *DeepESD*.

The procedure to validate *DeepESD* is similar to the one adopted in section 7.1. For

1975-2005 (historical scenario), the downscaled fields are directly compared against E-OBS. Differently, for 2005-2100 (RCP8.5) the plausibility of our high-resolution fields is assessed by comparing them against the driving CMIP5 GCMs —which are interpolated to the target 0.5° grid based on conservative remapping (Jones, 1999),— which are considered as "pseudo-reality". Moreover, a subset of RCMs from EURO-CORDEX 44 (see Table 5.3), re-gridded from their original spatial resolution (0.44°) to the target 0.5° grid by means of nearest interpolation— is also considered as "pseudo-reality" for the same purpose. As explained at the beginning of the chapter, deciding whether or not RCMs are able to provide more plausible climate change scenarios than GCMs is nowadays a hot research topic. Note therefore that it is important to consider both to assess the plausibility of the future projections delivered with *DeepESD*.

The top (bottom) panel in Figure 7.5 shows the results obtained for precipitation (temperature), as given by the multi-model ensemble mean for the GCMs, the RCMs and for *DeepESD* (in columns from left to right) for the period 1975-2005. In particular, the first row displays the corresponding climatologies whereas the second one shows the absolute biases with respect to the observational reference, E-OBS. Obviously, both the RCMs and *DeepESD* provide much finer spatial details than the GCMs, which yield smoother climatology maps for the two target variables. For precipitation, RCMs show a spatial structure which is directly linked to the orography, finding the highest rainfalls in mountainous regions such as the Pyrenees or the Alps. However, orography-driven precipitation is rather nonexistent in the GCMs and softened in *DeepESD*. For temperature, a general North-to-South positive gradient —driven by solar radiation— is found, with regional-to-local variations due mostly to the orography, which are more pronounced in the RCMs and in *DeepESD*.

In terms of biases with respect to E-OBS, the physical models (i.e., GCMs and RCMs) provide clearly worse results than *DeepESD*, which exhibits nearly negligible errors for both precipitation and temperature over the entire continent. This suggests that the PP assumption of having predictors which are well reproduced by the GCM —as compared to reanalysis— is overall not violated for the predictors/GCMs considered to produce *DeepESD*. In this case, both GCMs and RCMs —which show a similar pattern of biases— overestimate mean daily precipitation by about 1.5-2 mm in most regions of the continent (with the exception of the Western coast of Scandinavia, some parts of the United Kingdom and Croatia). Presumably, this is due to the drizzle effect, which is known to increase the number of rainy days in the climate models (more details in Figure 7.8). Differently, for temperature, whilst the GCMs alternate positive and negative biases, the RCMs systematically underestimate this variable across the whole Europe, especially in the Iberian Peninsula and Scandinavia. The results found are consistent with previous

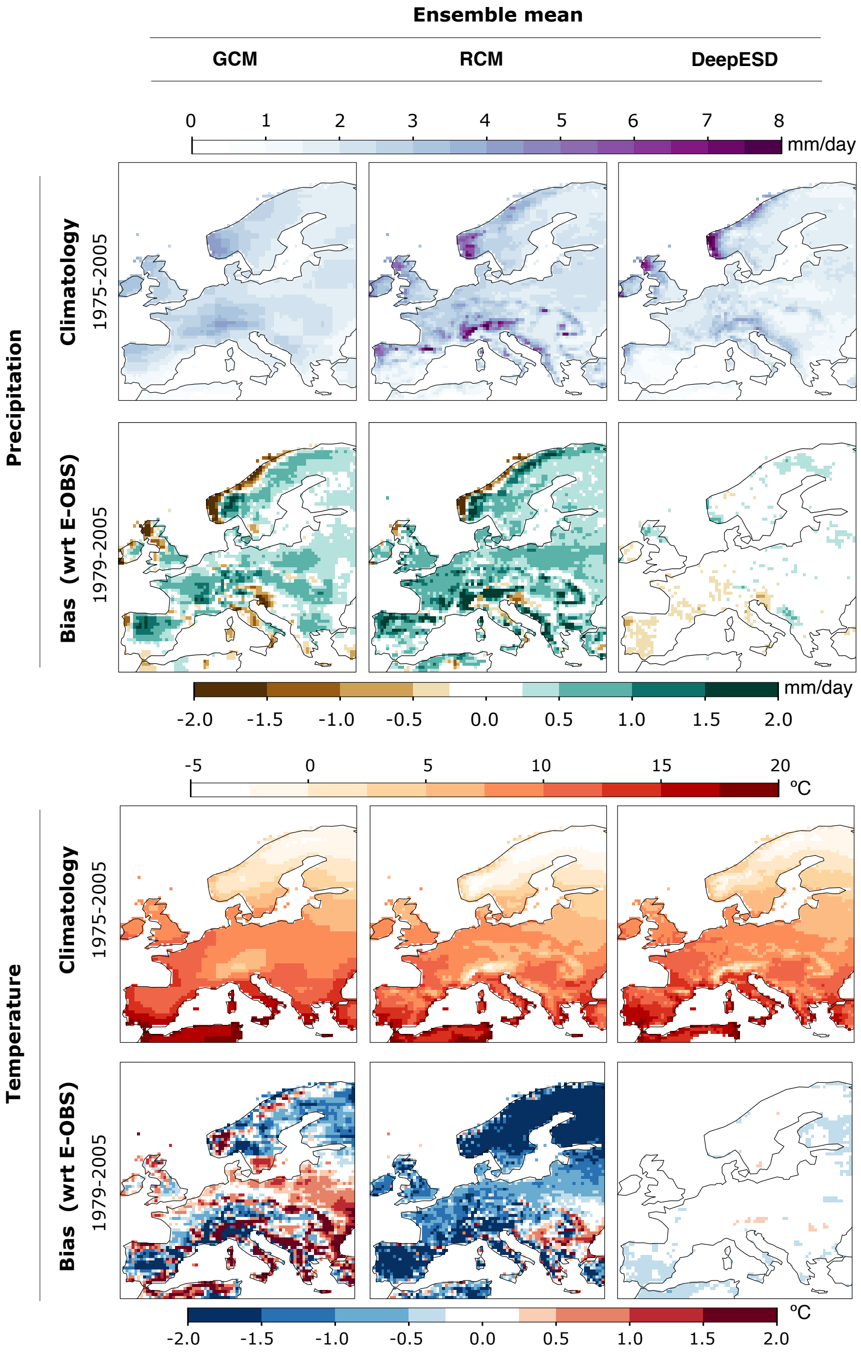


Figure 7.5: Top (bottom) panel: The first row shows the mean climatology for precipitation (temperature), as given by the multi-model ensemble mean for the GCMs, the RCMs and for *DeepESD* for the period 1975-2005. The second row displays the absolute biases of these climatologies with respect to the observational reference, E-OBS.

studies which have analyzed the ability of GCMs and RCMs to reproduce the European climate (Christensen et al., 2008; Jacob et al., 2020). Although bias-corrected versions of both GCMs and RCMs over Europe have been developed (Dosio, 2016), previous works have shown that this type of post-processing can potentially alter the climate change signals projected by the climate models (Casanueva et al., 2019, 2020). Therefore, since GCMs and RCMs are here used as "pseudo-reality" to assess the plausibility of the future projections delivered with *DeepESD*, we rely on the raw models outputs with the idea of preserving the original change signals from the models.

As opposite to GCMs and RCMs, *DeepESD* has proved to produce unbiased high-resolution fields of daily precipitation and temperature over the entire continent for the period 1975-2005 (under the historical scenario). To assess how plausible the future projections (up to 2100, based on the RCP8.5 scenario) delivered with *DeepESD* are, the left (right) panel in Figure 7.6 shows the climate change signals obtained for precipitation (temperature) from the multi-model ensemble mean of GCMs, RCMs and *DeepESD* —in columns— for the near (2006-2040), mid (2041-2070) and far (2071-2100) future —in rows.— In all cases, absolute differences with respect to 1975-2005 (see the first row in Figure 7.5) are shown. Recall that the idea here is to use the changes projected by the GCMs and the RCMs as "pseudo-reality".

The projected pattern of changes for precipitation gets intensified as we move from the near- to the far-future. In fact, notable changes do not appear until the mid-future, with an increase (decrease) of about 0.25-0.5 mm/day in Northern Scandinavia (Western region of the Iberian Peninsula). By the end of the century, a latitudinal dipole is found, with the Northern (Southern) part of the continent receiving increased (decreased) rainfall. Giorgi and Lionello (2008) justifies this decrease in precipitation in the Mediterranean arch due to a northward shift of the Atlantic storm tracks, which at the same time results into higher rainfall in the Northern European regions (Hanssen-Bauer et al., 2005). Despite the three ensembles show similar spatial mean patterns, some differences exist. In particular, as compared to GCMs and RCMs, *DeepESD* projects stronger increases over certain parts of Scandinavia and Central and Eastern Europe, and weaker decreases over the Iberian Peninsula for the far-future. For temperature, a generalized increase —which is intensified as we move from the near future (0.5-2°C) to the far-future (3-6°C)— is projected all over the continent. Whilst the strongest warming is expected to occur over Scandinavia and the Mediterranean basin, the smallest warming would be found over the British Islands and Central Europe. In addition, regional-to-local signals of change are projected by the RCMs and *DeepESD* in mountainous regions, especially the Alps and the Balkans. Moreover, as compared to the GCMs, RCMs and *DeepESD* project a smaller warming (with differences of about 1-1.5°C over Central and Eastern Europe for the far-future). This difference between GCMs and RCMs has been reported in previous studies over

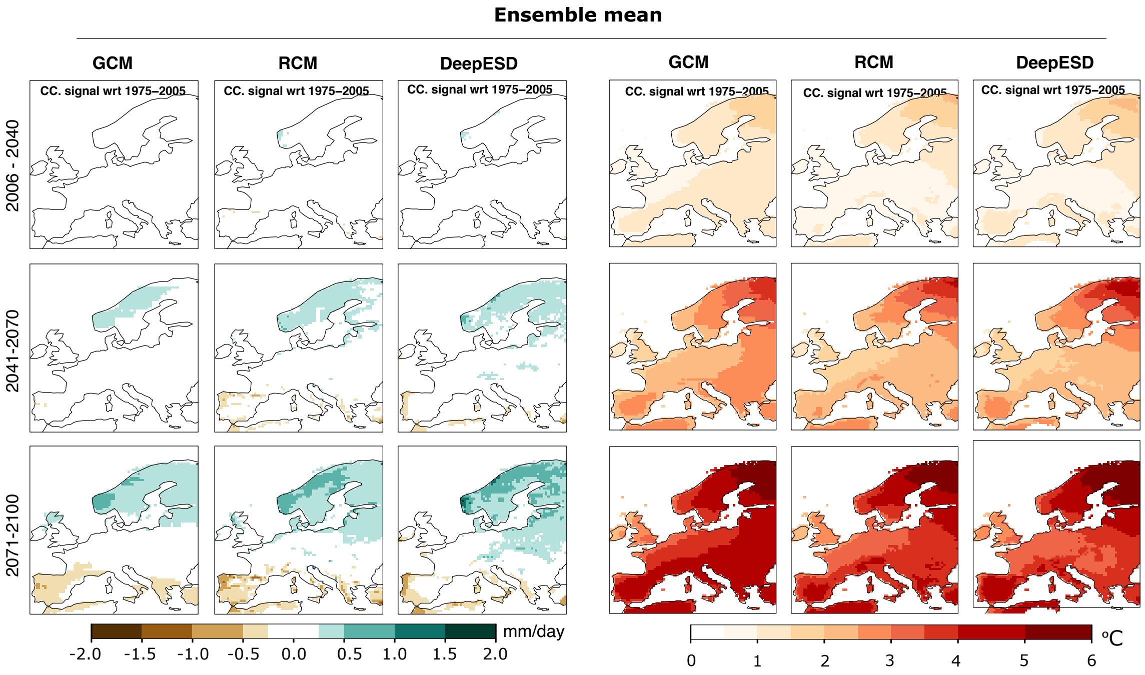


Figure 7.6: Climate change signals obtained for precipitation (temperature) from the multi-model ensemble mean of GCMs, RCMs and *DeepESD* —in columns— for the near (2006-2040), mid (2041-2070) and far (2071-2100) future —in rows— under the RCP8.5 scenario. In all cases, absolute differences (with respect to the historical scenario: 1975-2005) are shown.

Europe (see, e.g., Boé et al. (2020) and references therein). In particular, Gutiérrez et al. (2020) reports a reduction in surface solar radiation —which is directly related to surface temperature— over Central and Eastern Europe, as projected by RCMs which were not driven with time-varying anthropogenic aerosols. Despite further studies are needed to robustly explain the differences exhibited by both GCM and RCM ensembles, nowadays the literature reinforces the plausibility of the warming signals exhibited by GCMs for these particular regions and justifies the use of both RCMs but also GCMs as "pseudo-reality" in this experiment. With regards to DeepESD, further analyses mostly focused on the stationarity assumption have to be conducted, to characterize the differences in their climate signals (especially for temperature) with respect to those simulated by their driving GCMs.

Besides the spatial results shown in Figures 7.5 and 7.6, Figures 7.8 and 7.9 show the yearly time-series for the precipitation and temperature indicators (see Table 5.6), averaged over the eight PRUDENCE regions (Figure 7.7), which are broadly representative of the different European climate regimes. For every indicator, the ensemble of GCMs (red), RCMs (blue) and DeepESD (yellow) for the total period 1975-2100, plus the obser-

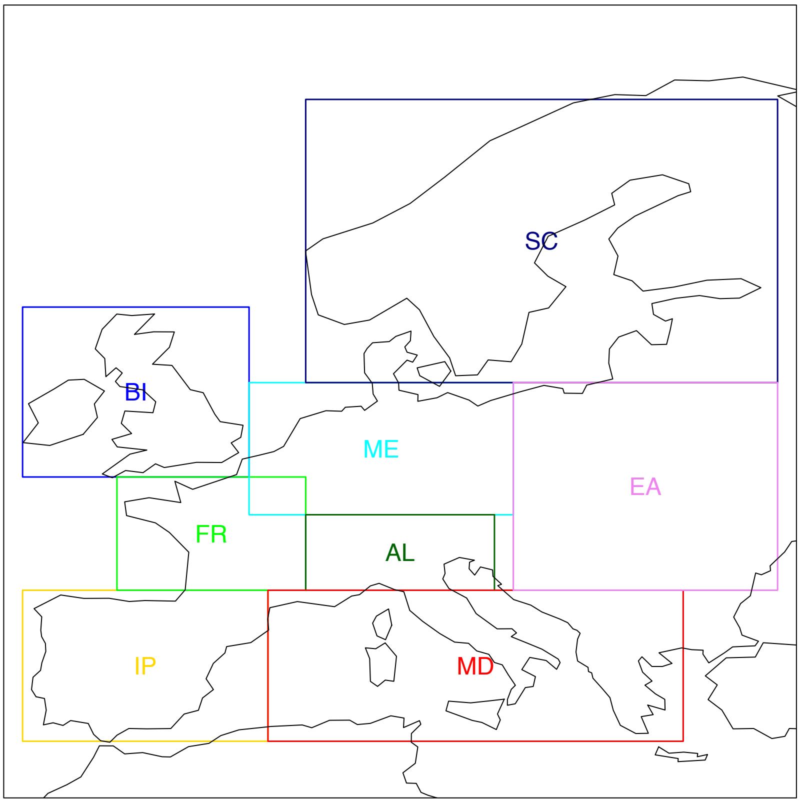


Figure 7.7: The eight PRUDENCE regions defined in Christensen and Christensen (2007): British Islands (BI), France (FR), Iberian Peninsula (IP), Alps (AL), Mid-Europe (ME), Mediterranean (MD), Scandinavia (SC), and Eastern Europe (EA).

vational reference, E-OBS (black), for the period 1979-2008, are shown. In all cases, the solid lines represent the multi-model ensemble mean whilst the shadows encompass all the models contributing to the ensemble. For a more detailed analysis, we also include a 5-fold cross-validated time-series (green) for 1979-2008 —obtained in "perfect conditions", that is, using ERA-Interim and E-OBS as predictor and predictand datasets, respectively,— which is used as reference to assess the performance of *DeepESD* to reproduce these indicators over that period. Moreover, these figures allow to qualitatively assess the plausibility of the projections provided by *DeepESD* by comparing them with those returned by the GCMs and the RCMs, which are considered as "pseudo-reality."

In particular, Figure 7.8 shows the results obtained for the R01, SDII and P98Wet indicators of precipitation (in different panels from top to bottom). Lower R01 values are found for the Southern European regions (IP, MD) than for the Northern (BI, ME) and the mountainous ones (AL, SC). *DeepESD* provides unbiased estimates of this indicator both in "perfect conditions" and for the historical scenario across the eight PRUDENCE

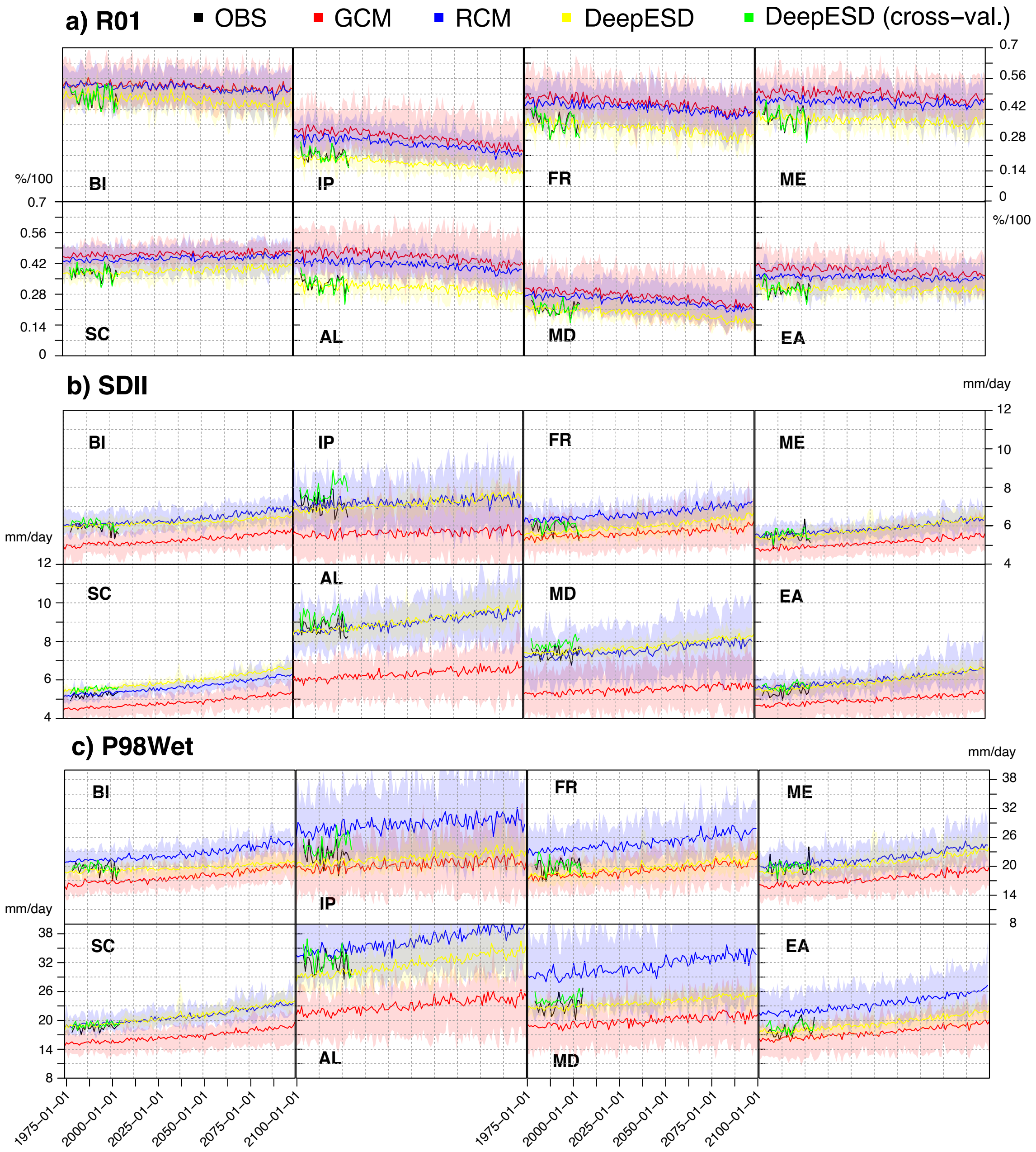


Figure 7.8: Yearly time-series for R01, SDII and P98Wet, averaged over the eight PRUDENCE regions. For every indicator, the ensemble of GCMs (red), RCMs (blue) and DeepESD (yellow) for the total period 1975-2100, plus the observational reference, E-OBS (black), for the period 1979-2008, are shown. In all cases, the solid lines represent the multi-model ensemble mean whilst the shadows encompass all the models contributing to the ensemble. For a more detailed analysis, we also include a 5-fold cross-validated (obtained in "perfect" conditions) time-series (green) for 1979-2008.

regions. However, both GCM and RCM ensembles (especially the latter) overestimate the frequency of wet days, which is consistent with the drizzle effect. In terms of changes, the three ensembles project similar signals, with an increase (decrease) in the Northern (Southern) regions, and no noticeable variations in Mid- and Eastern-Europe. Regarding the reproduction of rainfall amount, the GCMs underestimate both SDII and P98Wet, especially over the Alps —where biases of about -3 and -10 mm/day, respectively, are found for these two indicators,— due to the miss-representation of orographic convection. As compared to GCMs, the added value of RCMs to reproduce mean precipitation is reflected by their ability to provide better estimates of the SDII in all regions. Nevertheless, extreme precipitation is still quite challenging for RCMs, which clearly overestimate the P98Wet in some regions, especially over the Southern part of the continent (IP and MD zones). In contrast to GCMs and RCMs, *DeepESD* provides in general more robust estimates for both SDII and P98Wet under the historical scenario. In particular, nearly unbiased results are found for these two indicators in all regions with the exception of the Iberian Peninsula and the Alps, where P98Wet is underestimated. Nevertheless, this effect is not seen in the cross-validated time-series, which suggests that some inconsistency between ERA-Interim and the GCMs may exist for some of the predictor variables considered over these regions —recall from section 6.2.2 that CNNs are able to automatically take advantage of site-dependent windows of information from the predictor field.— Beyond these differences, the three ensembles project an increase in both the SDII and P98Wet indicator across all regions.

Figure 7.9 is the equivalent to 7.8 but for temperature. In this case, the indicators assessed are P02, the mean and P98. As reflected by the E-OBS curves, the regions located at the highest latitudes (BI, SC) or in mountainous areas (AL) present the lowest observed records for the three indices (in between -10 and 22°C for P02, 0-10°C for the mean and 15-20°C for the P98). On the contrary, the Southern regions (IP, MD) exhibit the highest ones (1-5°C for P02, 12-15°C for the mean and 23-28°C for P98). This latitudinal gradient in temperatures is induced by the solar radiation received by these regions across the year, but it is also related to the orography and local processes such as land-sea contrasts, among others. In general, the three ensembles perform similarly for the three indicators and across all regions. In particular, whilst the GCM ensemble alternates positive (MD, EA) with negative (AL, SC) biases in the historical scenario, the RCM ensemble underestimates the three indicators considered across all regions. This is consistent with the maps shown in Figure 7.5 and proves that the results from that figure are not due to a particular climate model, since all of them behave similarly. In contrast to GCMs and RCMs, *DeepESD* exhibit unbiased results for the three indicators and across all regions under the historical scenario. Indeed, note that the cross-validated time-series follow precisely the

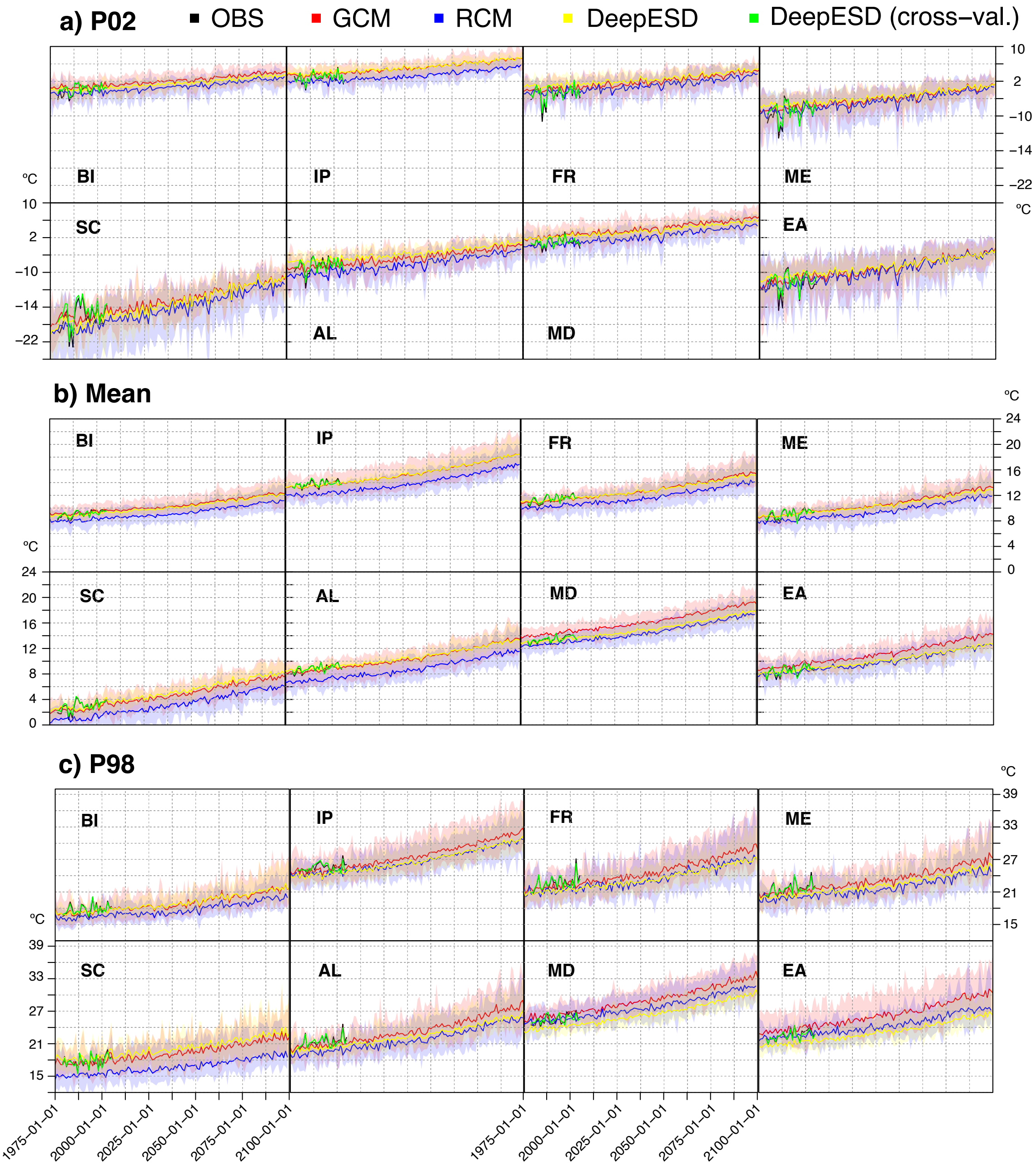


Figure 7.9: As Figure 7.8 but for temperature. In this case, the indicators analyzed are P02, the mean and P98.

E-OBS curves, and this behaviour does not worsen when passing from the reanalysis to the GCMs world. As per the projected signals of change, the three ensembles point out to a (quasi) linear increase for the three indicators studied along the century and across all regions, with warming values of about 4-6°C for the far-future in most of cases.

Overall, our results show that *DeepESD* reliably reproduce the observed precipitation and temperature fields over Europe in the historical period which gives certain confidence on the plausibility of the future projections developed within this new dataset, as significant deviations —on average for the eight PRUDENCE zones— from the "pseudo-reality" provided by GCMs and RCMs are not encountered. Nevertheless, DeepESD projects lower warming signals for the far-future than the ensemble of the driving GCMs. This aspect has to be analyzed in future studies to assess whether these differences are consequence of a better reproducibility of the local scale or to violations in the stationarity assumption. The main contribution of DeepESD is the dataset itself, which represents the first of its kind at a continental-scale, and is expected to fasten the analysis of SD-based climate projections with views to a possible integration of these products in climate impact studies.

To close this chapter, we assess the contribution of various factors to the uncertainty of the future projections (Hertig and Jacobeit, 2008; San-Martín et al., 2017; Manzanas et al., 2020a). To do this, we first measure the overall spread of the different ensembles that were represented with shadows in Figures 7.8 and 7.9. This is done by calculating the average value of the year-to-year spreads —understood as the standard deviation across all contributing climate models— for the period 2071-2100. The bottom row in Table 7.1 (7.2) shows the results obtained for the different precipitation (temperature) indicators analyzed: R01, SDII and P98Wet (P02, mean and P98). These values provide an estimation for the uncertainty that is due to the use of different climate models (either GCMs or RCMs). Note that the comparison between the *DeepESD* and RCM ensembles is not totally rigorous since, whilst the former includes eight GCMs, the latter is formed by eleven RCMs. Differently, *DeepESD* and GCM ensembles are based on exactly the same climate models and therefore the comparison is fairer in this case.

For completeness, three additional sources of uncertainty were also analyzed for the case of *DeepESD*. These are related to different details of the particular CNN setup considered for downscaling which can affect the spread of the future projections, namely: 1) the use of different predictor configurations, 2) the use of different versions of the same network topology, differing only in the values of their parameters, and 3) the choice of different realization from the stochastic sampling. In particular, 1) is addressed by selecting three different predictor sets from those available in Table 5.2, 2) is addressed by training ten versions of the CNN1 and CNN10 models —for precipitation and temperature, respectively,— and 3) is addressed by sampling ten times from the Bernoulli-Gamma(Gaussian) conditional distributions estimated with the CNN1(CNN10) topologies. For this analysis, we focus on one single GCM, the EC-Earth (which was already analyzed in detail in section 7.1) and form a different ensemble for each of the uncertainty sources analyzed —referred previously to as 1), 2) and 3).— The spread of each of these ensembles is then calculated as the average value of its year-to-year standard deviations for the period 2071-2100.

Precip.	**GCM**			**DeepESD**			**RCM**		
Source of uncertainty	**R01**	**SDII**	**P98Wet**	**R01**	**SDII**	**P98Wet**	**R01**	**SDII**	**P98Wet**
Predictor conf.	-	-	-	0.00	0.12	0.57	-	-	-
Model training	-	-	-	0.00	0.15	0.52	-	-	-
Conditional sampling	-	-	-	0.00	0.01	0.10	-	-	-
Ensemble	0.05	0.73	2.78	0.02	0.24	1.18	0.04	0.50	3.44

Table 7.1: For the three ensembles analyzed (GCMs, RCMs and *DeepESD*), the bottom row shows the average value of the year-to-year spreads —understood as the standard deviation across all contributing climate models— for the period 2071-2100 for R01, SDII and P98Wet. These values provide an estimation for the uncertainty that is due to the use of different climate models (either GCMs or RCMs). For *DeepESD*, three additional sources of uncertainty, which are related to different details of the particular CNN setup considered for downscaling, are also analyzed (see the text for details). In this case, a single climate model is considered, the EC-Earth. The R01 values are expressed in %/100 whilst SDII and P98Wet in mm/day.

Temperature	**GCM**			**DeepESD**			**RCM**		
Source of uncertainty	**P02**	**Mean**	**P98**	**P02**	**Mean**	**P98**	**P02**	**Mean**	**P98**
Predictor conf.	-	-	-	0.10	0.26	0.59	-	-	-
Model training	-	-	-	0.11	0.07	0.17	-	-	-
Conditional sampling	-	-	-	0.01	0.00	0.00	-	-	-
Ensemble	1.52	1.36	2.26	1.10	0.85	2.03	1.60	1.20	1.71

Table 7.2: As Table 7.1 but for temperature. In this case, the indicators analyzed are P02, the mean and P98. All values are expressed in °C.

For precipitation, the ensemble spread is substantially lower for *DeepESD* than for the GCMs and the RCMs —less than the half— for the three indicators analyzed. With regards to the additional sources of uncertainty analyzed for *DeepESD*, the choice of predictor configuration and network re-training are similarly important. However, the uncertainty due to the choice of realization from the stochastic sampling is almost irrelevant.

For temperature, the uncertainty due to the choice of climate model is lower for *DeepESD* than for the GCM and RCM ensembles for the three metrics. For the particular case of P98, the RCM ensemble exhibits the lowest spread, followed by *DeepESD* and the GCMs. More in detail for *DeepESD*, the uncertainty related to the parameters variability is lower than that due to the choice of predictor configuration (particularly for the mean and P98). As per precipitation, the uncertainty due to the conditional sampling is almost negligible.

Part IV

Concluding Remarks

CHAPTER 8

Conclusions, Achievements and Future Work

8.1 Main Conclusions

Building on the experimental frameworks defined in the COST action VALUE (Maraun et al., 2015) and EURO-CORDEX ESD (Jacob et al., 2020) we have designed in this Thesis a series of analysis that allow to comprehensively assess the potential benefits and limitations of CNNs for climate downscaling tasks. The main conclusions obtained are next summarized in relation to the objectives posed in section 2.2 (in italics).

1. *To test the applicability and performance of CNNs for climate downscaling in "perfect" conditions —i.e. based on reanalysis predictors.— In this regard, one of the key features to examine will be their ability to deal with high-dimensional input spaces.*
   As compared to classical GLMs, we have proved that CNNs provide better validation results in "perfect" conditions. In particular, among the different topologies intercompared, CNN1 (based on three convolutional layers of fifty, twenty-five, and one feature maps, respectively) and CNN10 (same CNN1, but with ten feature maps in the last hidden layer instead of one) were shown to be the best-performing configurations for the downscaling of temperature and precipitation, respectively. The superiority of CNNs over GLMs is due to 1) their ability to learn complex and non-linear patterns from data — which has been found to be particularly relevant for the case of precipitation— and 2) their ability to efficiently handle high-dimensional input spaces. The latter constitutes a clear advantage over traditional SD methods since tedious and human-guided dimensionality reduction techniques —which may entail a loss of relevant information for the downscaling— are no longer needed.

2. *To evaluate the benefits and disadvantages of CNN multi-site topologies, as compared to the equivalent single-site versions. We will analyze for this aim the implicit regularization that occurs in multi-site architectures.*

   We have seen that, whereas single-site CNNs are prone to overfitting in certain locations, the equivalent multi-site topologies perform an implicitly regularization which allows the network to treat simultaneously the high-dimensional predictor space, avoiding overfitting and leading to improved forecast accuracy.

3. *To gain understanding about the internal functioning of CNNs, which are typically seen as "black-box" models. To do this, we will focus on the study of the predictor-predictand link (i.e., influence of every input feature in the downscaling model outputs).*

   By studying the connection between the last hidden layer and the output space, we have demonstrated that the largest weights are found over an area of approximately 5x5 gridboxes surrounding the location of interest, with (quasi) zeroed-values elsewhere. This indicates that the network automatically neglects the regions which are not of interest for downscaling, taking advantage thus of site-dependent windows of information. Moreover, based on saliency maps, we have proved that CNNs are able to extract useful knowledge from the most informative predictors at each site, neglecting those whose influence for downscaling is scarce.

4. *To study the suitability of CNNs to downscale future climate change scenarios. To do so, we will first evaluate the ability of CNNs to reproduce the observed climate based on the historical scenario of a GCM. Then, we will explore their potential for moderate and coherent extrapolation under one emission scenario, based on various GCMs.*

   For a single GCM, the EC-Earth, we have proved that our CNNs produce unbiased high-resolution fields of daily precipitation and temperature over Europe for the period 1979-2008 (under the historical scenario). Moreover, the climate change signals obtained with our CNNs when downscaling the RCP8.5 scenario (for the far future 2071-2100) are notoriously more compatible with EC-Earth's raw outputs — considered as "pseudo-reality"— than those provided by classical GLMs. Based on this promising result, we used our CNNs to downscale a subset of eight GCMs from CMIP5, producing *DeepESD*, the first ensemble of high-resolution projections (up to 2100) of daily precipitation and temperature over Europe based on DL. Based on this new dataset, we have demonstrated that CNNs allow for moderate extrapolation, providing projections which are broadly compatible with those given by the driving CMIP5 GCMs and a subset of RCMs from EURO-CORDEX.

8.2 Key Achievements

8.2.1 Publications

The main results of this Thesis (Part III) have led to a series of publications in international journals and conference proceedings of relevance in the fields of atmospheric sciences and artificial intelligence. In particular,

- Section 6.1 in Chapter 6 is based on **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "Configuration and intercomparison of deep learning neural models for statistical downscaling", *Geoscientific Model Development*, vol. 13, pp. 2109–2124, 2020, DOI: 10.5194/gmd-2019-278 (1^{st} decile in JCR[1])

- Section 6.2.1 in Chapter 6 is based on **J. Baño-Medina** and J. M. Gutiérrez, "The importance of inductive bias in convolutional models for statistical downscaling", *Proceedings of the 9th International Workshop on Climate Informatics: CI 2019*, 2019, DOI:10.5065/y82j-f154

- Section 6.2.2 in Chapter 6 is based on **J. Baño-Medina** and J. M. Gutiérrez, "Deep convolutional networks for feature selection in statistical downscaling", *Proceedings of the 8th International Workshop on Climate Informatics: CI 2018*, 2018, DOI: 10.5065/D6BZ64XQ

- Section 6.2.2 in Chapter 6 is based on **J. Baño-Medina**, "Understanding deep learning decisions in statistical downscaling models", *Association for Computing Machinery, New York, NY, USA, p 79–85*, 2020, DOI: 10.1145/3429309.3429321

- Section 7.1 in Chapter 7 is based on **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "On the suitability of deep convolutional neural networks for downscaling climate change projections", *Climate Dynamics*, 2021, DOI: 10.1007/s00382-021-05847-0 (1^{st} quartile in JCR).

- Section 7.2 in Chapter 7 is based on **J. Baño-Medina**, R. Manzanas, and J. M. Gutiérrez, "DeepESD: An Ensemble of Regional Climate Change Projections over Europe based on Deep Learning Downscaling", *Submitted to Nature Scientific Data.* (1^{st} quartile in JCR)

Additionally, as a result of the activities carried out in parallel to the development of this Thesis at the Santander Meteorology Group (SMG), two more publications related to software development have been released:

[1] Journal Citation Reports (JCR) is a tool that permits to measure the relative importance of a journal within its corresponding thematic based on the number of citations its papers receive.

- Section 8.2.2 in Chapter 8 is based on M. Iturbide, J. Bedia, S. Herrera, **J. Baño-Medina**, J. Fernández, M.D. Frías, R. Manzanas, D. San-Martín, E. Cimadevilla, A.S. Cofiño and J.M. Gutiérrez, "The R-based climate4R open framework for reproducible climate data access and post-processing", *Environmental Modelling Software, vol. 111, pp. 42-54*, 2019, DOI: 10.1016/j.envsoft.2018.09.009 (1^{st} quartile in JCR).

- Section 8.2.2 in Chapter 8 is based on J. Bedia, **J. Baño-Medina**, M.N. Legasa, M. Iturbide, R. Manzanas, S. Herrera, D. San-Martín, A.S. Cofiño and J.M. Gutiérrez, "Statistical downscaling with the downscaleR package (v3.1.0): Contribution to the VALUE intercomparison project", *Geoscientific Model Development*, 2019, DOI: 10.5194/gmd-2019-224 (1^{st} decile in JCR)

Furthermore, the following contributions have been presented in national and international conferences:

- "downscaleR: An R-based package for statistical downscaling and bias correction within the climate4R framework", 2^{nd} Workshop on Bias Correction in Climate Studies, Santander, Spain (poster)

- "Deep convolutional networks for feature selection in statistical downscaling", 8^{th} International workshop on Climate Informatics (CI), Colorado, EE.UU (poster)

- "Climate research reproducibility with the climate4R R-based framework", 8^{th} International workshop on Climate Informatics (CI), Colorado, EE.UU (poster)

- "Deep neural networks for statistical downscaling of climate change projections", XVIII Conference of the Spanish association for Artificial Intelligence (CAEPIA), Granada, Spain (oral)

- "The influence of inductive bias in convolutional models for statistical downscaling", 9^{th} International workshop on Climate Informatics (CI), Paris, France (oral & poster)

- "Statistical downscaling with deep learning: A contribution to CORDEX-CORE", International Conference for Regional Climate ICRC-CORDEX, Beijing, China (oral)

- "On the suitability of convolutional neural networks for climate downscaling", 1^{st} Artificial Intelligence for Copernicus Workshop, Reading, England (oral)

- "Understanding deep learning decisions in statistical downscaling models", 10^{th} International Conference on Climate Informatics (CI), Oxford, UK (oral)

Note also that this Thesis has won the 3^{rd} prize award at the doctoral consortium of the *"Asociación Española de Inteligencia Artificial (AEPIA)"*, that took place at the XVIII Conference on Artificial Intelligence in Granada, Spain (2018).

Finally, building on the methodological knowledge gained during the realization of this Thesis, we have produced *DeepESD* (see section 7.2), the first dataset based on DL that provides downscaled daily projections (up to 2100) of precipitation and temperature for an ensemble of eight GCMs, covering the entire Europe at a resolution of 0.5°. *DeepESD* is publicly available through the Earth System Grid Federation (ESGF), at the University of Cantabria's node (`https://data.meteo.unican.es/thredds/catalog/esgcet/collections/CORDEX-DeepESD-EE/catalog.html`[2]).

8.2.2 Software

This Thesis builds on (and contributes to) climate4R (C4R), a bundle of *R* packages developed at the Santander Meteorology Group (SMG) which allow to meet the particularities and requirements of (almost) every climate data application —Fig.8.1 shows schematically the role of every C4R library within a typical workflow in a climate experiment.— In particular, in this Thesis we make an extensive use of `loadeR` (data access), `transformeR` (data manipulation; e.g., re-gridding), `downscaleR` (statistical downscaling), `climate4R.value` (validation) and `visualizeR` (visualization of results). We refer the reader to the reference manuscript (Iturbide et al., 2019) and/or to the GitHub repository (`https://github.com/SantanderMetGroup/climate4R`) for more details about C4R.

Despite the collaboration in different core libraries —especially in `downscaleR` (Bedia et al., 2020), which provides functions to carry out every step of the downscaling workflow from model setup to model training and prediction (see Fig.8.2),— the main contribution of this Thesis to C4R is `downscaleR.keras`, which provides an interface to Keras (Chollet et al., 2015), an extremely popular high-level API for building and training deep learning models. Keras supports arbitrary network architectures and is seamlessly integrated with TensorFlow (Abadi et al., 2016). `downscaleR.keras` has allowed to incorporate sophisticated convolutional or recurrent neural networks (among others), to the set of classical SD methods included in `downscaleR` —for instance, the GLMs used in this Thesis.— More information about this package can be found in `https://github.com/SantanderMetGroup/downscaleR.keras`.

[2] This URL is temporal, since the paper describing *DeepESD* is currently under review. Once the paper is published, a final version of the dataset will be released.

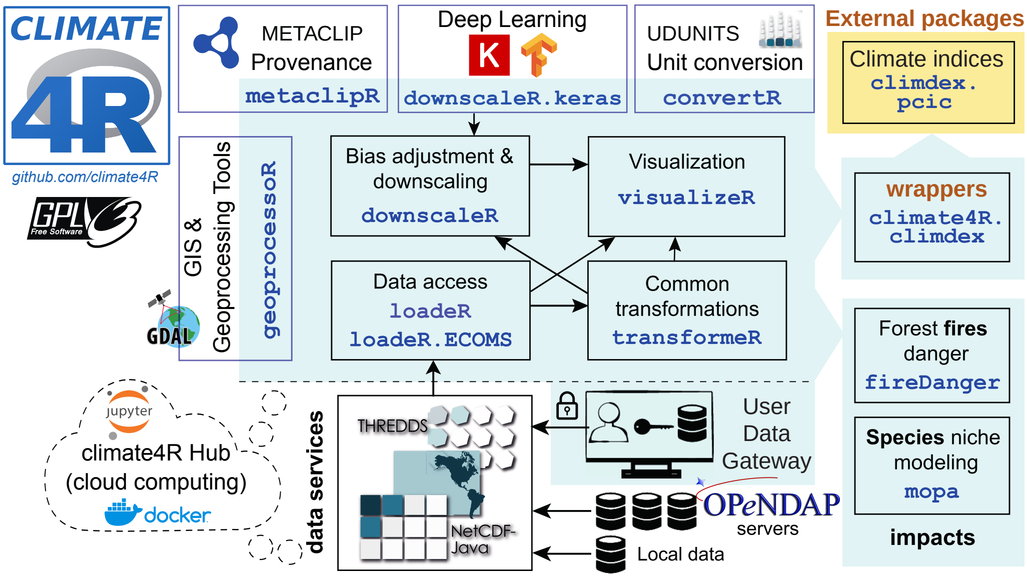


Figure 8.1: Diagram illustrating the climate4R framework (figure taken from Iturbide et al. (2019)).

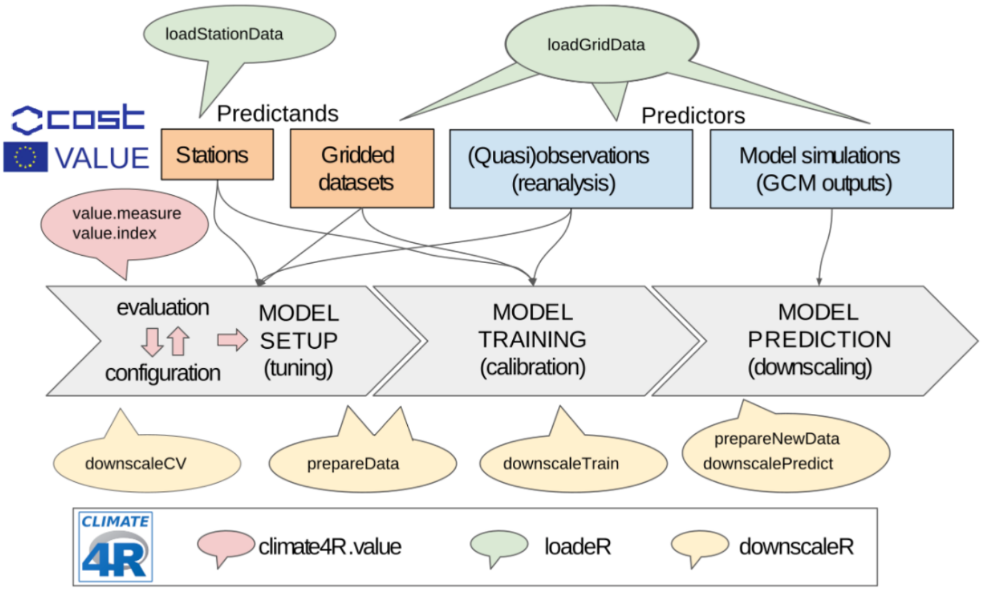


Figure 8.2: Schematic representation of the key functions included in `downscaleR`, which are used at different steps of the downscaling process (figure taken from Bedia et al. (2020))
.

8.2.3 Reproducibility

Transparency and reproducibility are key ingredients to develop high-quality science. Recently, the community has gathered to define a set of FAIR (Findability, Accesibility, Interoperability and Reuse) principles that provide guidelines for users to promote re-usability of their data and code (Wilkinson et al., 2016). In alignment with these principles, we created a GitHub repository (`https://github.com/SantanderMetGroup/DeepDownscaling`) which hosts a series of (R) Jupyter notebooks (Pérez and Granger, 2007) which allow the user not only to fully reproduce the results presented in this Thesis, but also to adapt our original code to his/her particular needs. Every notebook (see Table 8.1) is associated with a published paper. To ease compatibility, we have "freezed" the specific versions of the C4R libraries used within each notebook, which can be installed simply via *conda* (see `https://github.com/SantanderMetGroup/climate4R` for details on the installation of C4R).

Section		Manuscript	Notebook	C4R version
Chap. 6	Sec. 6.1	*Configuration and Intercomparison of Deep Learning Neural Models for Statistical Downscaling*	2018_Bano_CI.ipynb	v1.3.0
Chap. 6	Sec. 6.2.1	*The Importance of Inductive Bias in Convolutional Models for Statistical Downscaling*	2019_Bano_CI.ipynb	v1.5.0
Chap. 6	Sec. 6.2.2	*Deep Convolutional Networks for Feature Selection in Statistical Downscaling*	2020_Bano_GMD.ipynb	v1.5.0
Chap. 6	Sec. 6.2.2	*Understanding Deep Learning Decisions in Statistical Downscaling Models*	2020_Bano_CI.ipynb	v1.5.0
Chap. 7	Sec. 7.1	*On the Suitability of Deep Convolutional Neural Networks for Downscaling climate change projections*	2020_Bano_CD.ipynb	v1.3.0
Chap. 7	Sec. 7.2	*DeepESD: An Ensemble of Regional Climate Change Projections over Europe based on Deep Learning Downscaling*	2021_Bano_NSD.ipynb	v1.5.0

Table 8.1: Information regarding the reproducibility of the results presented in this Thesis, which are based on different published manuscripts, each with a corresponding (R) Jupyter notebook. These notebooks can be found in the SMG GitHub repository (`https://github.com/SantanderMetGroup/DeepDownscaling`, DOI: 10.5281/zenodo.3461087), and build on specific C4R versions of the libraries (conda installation available).

All the results presented along this Thesis have been produced with a virtual machine with the following technical specifications:

- Operating system: Ubuntu 18.04.3 LTS (64 bits)
- Memory: 60 GiB
- Processor: 2x Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz (16 cores, 32 threads)

Nonetheless, note that some of the notebooks listed in Table 8.1 can be perfectly run in machines with lower memory capacities.

8.3 Research Stays

During the realization of this Thesis two research stays have been done at international centres of reference in the fields of climate research and deep learning:

- January 2019, Institute of Data Science of the German Aerospace Agency (DLR) in the Climate Informatics Group, in Jena (Germany). The focus of this stay was to gain experience in the development of deep architectures designed for climate-related applications.

- March-April 2021, Centre National de Recherches Météorologiques (CNRM), in Toulouse (France), Under the umbrella of the CORDEX-FPS convection project (see `https://www.hymex.org/cordexfps-convection/`), we started a collaboration with CNRM (which is still on-going) aimed at building statistical emulators[3] based on DL models.

8.4 Future Work

Part of the results from this Thesis have opened the door for the development of new works which constitute the natural continuation of some of the analysis presented here.

For instance, we plan to assess the suitability of CNNs for the rest of CORDEX domains (beyond Europe) with the idea of providing a world-wide dataset of high-resolution climate change scenarios based on DL. Likewise, we also plan to move to finer than 0.5° spatial resolutions. The adaptation of CNNs to other domains and resolutions will require to introduce variations with respect to the configurations presented herein (e.g., fully-convolutional networks, batch normalization layers). This study will allow us to gain more knowledge on the potential benefits and limitations of different CNN topologies for climate downscaling tasks.

Also, we have seen in this Thesis that the lack of informativeness power in the predictors results in a limited representation of extremes in the downscaled fields. To date, this is solved by sampling out from the conditional distributions learnt at a gridbox level, which leads to a loss of temporal and spatial structure. However, DL may offer alternatives to

[3] Statistical emulators based on DL have recently emerged as a potential alternative to mimic the work done by a RCM, avoiding thus the associated long simulation times and high computational requirements. This can be done either by 1) using the GCM-RCM fields as input-output pairs to construct the DL model, or by 2) upscaling the circulation RCM variables to a coarser spatial resolution (predictors) and use the original high-resolution RCM fields as “pseudo-reality” (predictands).

cope with this issue such as Variational Autoencoders (VAE, Kingma and Welling (2013)) or Generative Adversarial Modeling (GAN, Goodfellow et al. (2014)). Another interesting approach to tackle this problem may be related to the quantification of the uncertainty in the estimation of the model parameters. In this regard, Bayesian Deep Learning (BDL, Gal (2016)) may help to build more robust models for downscaling at all time-scales, not only climate change projections but also weather and seasonal forecasts.

Furthermore, despite this has been partially addressed in this Thesis (see section 6.2), advancing towards a better understanding of the internal functioning of DL is crucial to make these models more appealing to the research community. Efforts in this direction are yet to be done, not only for building better DL models, but also for their adoption in a wider number of applications.

Finally, in the framework of an international collaboration between the SMG and CNRM, started during one of the research stays above described, we contemplate to open a new research line focused on the use of DL models as statistical emulators. This line will try to answer some of the key questions which have been posed by the CORDEX Flagship Pilot Study on convection[4]: 1) is a DL model capable to learn the non-linear system of differential equations that characterize a particular RCM? 2) if this relationship is learned for a particular GCM-RCM pair, is it applicable to downscale other GCMs? and 3) is this relationship able to extrapolate to other emission scenarios different to those used for learning? In this regard, the first results obtained during the stay in CNRM[5] indicate that the CNNs proposed in this Thesis show promising capabilities to emulate RCMs.

[4] See https://www.hymex.org/cordexfps-convection/wiki/doku.php?id=home for details about the Flagship Pilot Study (FPS) on convective phenomena at high resolution over Europe and the Mediterranean, one of the 13 FPS endorsed by CORDEX (the full list can be visited at https://cordex.org/experiment-guidelines/flagship-pilot-studies/endorsed-cordex-flagship-pilote-studies).

[5] Note that, since these results are very preliminary, we have decided to not include them in this document.

Bibliography

Abadi, M., et al., 2016: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467.*

Abaurrea, J. and J. Asín, 2005: Forecasting local daily precipitation patterns in a climate change scenario. *Climate Research*, **28 (3)**, 183–197.

Ailliot, P., D. Allard, V. Monbet, and P. Naveau, 2015: Stochastic weather generators: an overview of weather type models. *Journal de la Société Française de Statistique*, **156 (1)**, 101–113.

Akhtar, N., J. Brauch, and B. Ahrens, 2018: Climate modeling over the mediterranean sea: impact of resolution and ocean coupling. *Climate Dynamics*, **51 (3)**, 933–948.

Amin, J., M. Sharif, M. Yasmin, and S. L. Fernandes, 2018: Big data analysis for brain tumor detection: Deep convolutional neural networks. *Future Generation Computer Systems*, **87**, 290–297.

Baño-Medina, J., 2020: Understanding deep learning decisions in statistical downscaling models. *Proceedings of the 10th International Conference on Climate Informatics*, Association for Computing Machinery, New York, NY, USA, 79–85, CI2020.

Baño-Medina, J. and J. M. Gutiérrez, 2018: Deep convolutional networks for feature selection in statistical downscaling. *Proceedings of the 8th International Workshop on Climate Informatics: CI 2018.*

Baño-Medina, J. and J. M. Gutiérrez, 2019: The importance of inductive bias in convolutional models for statistical downscaling. *Proceedings of the 9th International Workshop on Climate Informatics: CI 2019.*

Babaousmail, H., R. Hou, G. T. Gnitou, and B. Ayugi, 2021: Novel statistical downscaling emulator for precipitation projections using deep convolutional autoencoder over northern africa. *Journal of Atmospheric and Solar-Terrestrial Physics*, **218**, 105 614.

Baño-Medina, J., R. Manzanas, E. Cimadevilla, J. Fernández, A. Cofiño, and J. M. Gutiérrez, 2021a: Deepesd: An ensemble of regional climate change projections over europe based on deep learning downscaling. *Submitted to Scientific Data.*

Baño-Medina, J., R. Manzanas, and J. M. Gutiérrez, 2021b: On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, 1–11.

Bardossy, A. and E. J. Plate, 1992: Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, **28 (5)**, 1247–1259.

Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an ai lens. *Geophysical Research Letters*, **46 (22)**, 13 389–13 398.

Baxter, J., 1995: Learning internal representations. *Proceedings of the eighth annual conference on Computational learning theory*, 311–320.

Baño-Medina, J., R. Manzanas, and J. M. Gutiérrez, 2020: Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, **13 (4)**, 2109–2124.

Bedia, J., et al., 2020: Statistical downscaling with the downscaler package (v3.1.0): contribution to the VALUE intercomparison experiment. *Geoscientific Model Development*, **13 (3)**, 1711–1735.

Beecham, S., M. Rashid, and R. K. Chowdhury, 2014: Statistical downscaling of multisite daily rainfall in a south australian catchment using a generalized linear model. *International Journal of Climatology*, **34 (14)**, 3654–3670.

Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle, 2007: Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 153–160.

Bengio, Y., P. Simard, and P. Frasconi, 1994: Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5 (2)**, 157–166.

Bentsen, M., et al., 2013: The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, **6 (3)**, 687–720.

Bergstra, J., et al., 2010: Theano: A cpu and gpu math compiler in python. *Proc. 9th python in science conf*, Vol. 1, 3–10.

Beucler, T., S. Rasp, M. Pritchard, and P. Gentine, 2019: Achieving conservation of energy in neural network emulators for climate modeling. *arXiv preprint arXiv:1906.06622*.

Bishop, C. M., 1995: Regularization and complexity control in feed-forward networks.

Boé, J., S. Somot, L. Corre, and P. Nabat, 2020: Large discrepancies in summer climate change over europe as projected by global and regional climate models: causes and consequences. *Climate Dynamics*, **54 (5)**, 2981–3002.

Bottou, L., F. E. Curtis, and J. Nocedal, 2018: Optimization methods for large-scale machine learning. *Siam Review*, **60 (2)**, 223–311.

Brands, S., S. Herrera, J. Fernández, and J. M. Gutiérrez, 2013: How well do cmip5 earth system models simulate present climate conditions in europe and africa? *Climate dynamics*, **41 (3)**, 803–817.

Brands, S., S. Herrera, D. San-Martín, and J. M. Gutiérrez, 2011a: Validation of the ensembles global climate models over southwestern europe using probability density functions, from a downscaling perspective. *Climate Research*, **48 (2-3)**, 145–161.

Brands, S., J. Taboada, A. Cofino, T. Sauter, and C. Schneider, 2011b: Statistical downscaling of daily temperatures in the nw iberian peninsula from global climate models: validation and future scenarios. *Climate Research*, **48 (2-3)**, 163–176.

Brandsma, T. and T. A. Buishand, 1997: Statistical linkage of daily precipitation in switzerland to atmospheric circulation and temperature. *Journal of hydrology*, **198 (1-4)**, 98–123.

Breiman, L., 1996: Bagging predictors. *Machine learning*, **24 (2)**, 123–140.

Bürger, G., 1996: Expanded downscaling for generating local weather scenarios. *Climate Research*, **7**, 111–128.

Cannon, A. J., 2008: Probabilistic Multisite Precipitation Downscaling by an Expanded Bernoulli–Gamma Density Network. *Journal of Hydrometeorology*, **9 (6)**, 1284–1300.

Caruana, R., 1998: Multitask Learning. *Learning to Learn*, S. Thrun and L. Pratt, Eds., Springer US, Boston, MA, 95–133.

Casanueva, A., S. Herrera, J. Fernández, M. Frías, and J. M. Gutiérrez, 2013: Evaluation and projection of daily temperature percentiles from statistical and dynamical downscaling methods. *Natural Hazards and Earth System Sciences*, **13 (8)**, 2089–2099.

Casanueva, A., S. Herrera, M. Iturbide, S. Lange, M. Jury, A. Dosio, D. Maraun, and J. M. Gutiérrez, 2020: Testing bias adjustment methods for regional climate change applications under observational uncertainty and resolution mismatch. *Atmospheric Science Letters*, **21 (7)**, e978.

Casanueva, A., S. Kotlarski, S. Herrera, A. M. Fischer, T. Kjellstrom, and C. Schwierz, 2019: Climate projections of a multivariate heat stress index: the role of downscaling and bias correction. *Geoscientific Model Development*, **12 (8)**, 3419–3438.

Chandler, R. E. and H. S. Wheater, 2002: Analysis of rainfall variability using generalized linear models: A case study from the west of ireland. *Water Resources Research*, **38 (10)**, 10–1.

Chang, Y.-S., H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin, 2020: An lstm-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, **11 (8)**, 1451–1463.

Chapman, W. E., A. C. Subramanian, L. D. Monache, S. P. Xie, and F. M. Ralph, 2019: Improving Atmospheric River Forecasts With Machine Learning. *Geophysical Research Letters*, **46 (17-18)**, 10 627–10 635.

Chaudhuri, C. and C. Robertson, 2020: Cligan: A structurally sensitive convolutional neural network model for statistical downscaling of precipitation from multi-model ensembles. *Water*, **12 (12)**, 3353.

Chen, S.-T., P.-S. Yu, and Y.-H. Tang, 2010: Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology*, **385 (1)**, 13–22.

Cheng, C., G. Li, Q. Li, and H. Auld, 2008: Statistical downscaling of hourly and daily climate scenarios for various meteorological variables in south-central canada. *Theoretical and Applied Climatology*, **91 (1)**, 129–147.

Chevallier, F., F. Chéruy, N. Scott, and A. Chédin, 1998: A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of applied meteorology*, **37 (11)**, 1385–1397.

Chollet, F. et al., 2015: Keras. https://keras.io.

Christensen, J. H., F. Boberg, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, **35 (20)**.

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

Christian, J., et al., 2010: The global carbon cycle in the canadian earth system model (canesm1): Preindustrial control simulation. *Journal of Geophysical Research: Biogeosciences*, **115 (G3)**.

Collins, M., et al., 2013: Long-term climate change: projections, commitments and irreversibility. *Climate Change 2013-The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 1029–1136.

Collobert, R., K. Kavukcuoglu, and C. Farabet, 2011: Torch7: A Matlab-like Environment for Machine Learning.

Cornes, R. C., G. v. d. Schrier, E. J. M. v. d. Besselaar, and P. D. Jones, 2018: An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, **123 (17)**, 9391–9409.

Cybenko, G., 1989: Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2 (4)**, 303–314.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *Journal of climate*, **19 (18)**, 4605–4630.

Daniely, A., R. Frostig, and Y. Singer, 2016: Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances In Neural Information Processing Systems*, **29**, 2253–2261.

Dee, D. P., et al., 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597.

Demory, M.-E., et al., 2020: Can high-resolution gcms reach the level of information provided by 12–50km cordex rcms in terms of daily precipitation distribution?, geosci. model dev. discuss. *Geosci Model Dev Discuss*, **2020**, 1–33.

Doblas Reyes, F., et al., 2018: Using ec-earth for climate prediction research. *ECMWF Newsletter*, **(154)**, 35–40.

Dosio, A., 2016: Projections of climate change indices of temperature and precipitation from an ensemble of bias-adjusted high-resolution euro-cordex regional climate models. *Journal of Geophysical Research: Atmospheres*, **121 (10)**, 5488–5511.

Doury, A., S. Somot, S. Gadat, A. Ribes, and L. Corre, 2021: Regional climate model emulator based on deep learning: Concept and first evaluation of a novel hybrid downscaling approach. *Climate Dynamics*.

Dueben, P. D. and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, **11 (10)**, 3999–4009.

Dufresne, J.-L., et al., 2013: Climate change projections using the ipsl-cm5 earth system model: from cmip3 to cmip5. *Climate dynamics*, **40 (9)**, 2123–2165.

Dunn, P., 2004: Occurrence and quantity of precipitation can be modeled simultaneously. *International Journal of Climatology*, **24**, 1231–1239.

Dunne, J. P., et al., 2013: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part II: Carbon System Formulation and Baseline Simulation Characteristics. *Journal of Climate*, **26 (7)**, 2247–2267.

Durman, C., J. M. Gregory, D. C. Hassell, R. Jones, and J. Murphy, 2001: A comparison of extreme european daily precipitation simulated by a global and a regional climate model for present and future climates. *Quarterly Journal of the Royal Meteorological Society*, **127 (573)**, 1005–1015.

Enke, W. and A. Spegat, 1997: Downscaling climate model outputs into local and regional weather elements by classification and regression. *Climate Research*, **8 (3)**, 195–207.

Fauzi, F., H. Kuswanto, and R. Atok, 2020: Bias correction and statistical downscaling of earth system models using quantile delta mapping (qdm) and bias correction constructed analogues with quantile mapping reordering (bccaq). *Journal of Physics: Conference Series*, IOP Publishing, Vol. 1538, 012050.

Fealy, R. and J. Sweeney, 2007: Statistical downscaling of precipitation for a selection of sites in ireland employing a generalised linear modelling approach. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **27 (15)**, 2083–2094.

Fowler, H., C. Kilsby, and P. O'Connell, 2000: A stochastic rainfall model for the assessment of regional water resource systems under changed climatic condition. *Hydrology and Earth System Sciences*, **4**.

François, B., S. Thao, and M. Vrac, 2021: Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, 1–31.

Gaertner, M. Á., et al., 2018: Simulation of medicanes over the mediterranean sea in a regional climate model ensemble: impact of ocean–atmosphere coupling and increased resolution. *Climate dynamics*, **51 (3)**, 1041–1057.

Gal, Y., 2016: Uncertainty in deep learning. Ph.D. thesis, University of Cambridge.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, 2018: Could Machine Learning Break the Convection Parameterization Deadlock? *Geophysical Research Letters*, **45 (11)**, 5742–5751.

Giffard-Roisin, S., M. Yang, G. Charpiat, C. Kumler Bonfanti, B. Kégl, and C. Monteleoni, 2020: Tropical Cyclone Track Forecasting Using Fused Deep Learning From Aligned Reanalysis Data. *Frontiers in Big Data*, **3**.

Gilbert, R. C., M. B. Richman, T. B. Trafalis, and L. M. Leslie, 2010: Machine learning methods for data assimilation. *Computational Intelligence in Architecturing Complex Engineering Systems*, 105–112.

Giorgi, F. and W. J. Gutowski Jr, 2015: Regional dynamical downscaling and the cordex initiative. *Annual Review of Environment and Resources*, **40**, 467–490.

Giorgi, F. and P. Lionello, 2008: Climate change projections for the mediterranean region. *Global and planetary change*, **63 (2-3)**, 90–104.

Giorgi, F., C. Torma, E. Coppola, N. Ban, C. Schär, and S. Somot, 2016: Enhanced summer convective rainfall at alpine high elevations in response to climate warming. *Nature Geoscience*, **9 (8)**, 584–589.

Glorot, X. and Y. Bengio, 2010: Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research - Proceedings Track*, **9**, 249–256.

Gómez-Gonzalez, C. A., L. Palma Garcia, L. Lledó, R. Marcos, N. Gonzalez-Reviriego, G. Carella, and A. Soret Miravet, 2021: Deep learning-based downscaling of seasonal forecasts over the iberian peninsula. *EGU General Assembly Conference Abstracts*, EGU21–12 253.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning*. MIT Press.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014: Generative adversarial nets. *Advances in neural information processing systems*, **27**.

Gunn, S. R. et al., 1998: Support vector machines for classification and regression. *ISIS technical report*, **14 (1)**, 5–16.

Gutiérrez, C., S. Somot, P. Nabat, M. Mallet, L. Corre, E. van Meijgaard, O. Perpiñán, and M. Á. Gaertner, 2020: Future evolution of surface solar radiation and photovoltaic potential in europe: investigating the role of aerosols. *Environmental Research Letters*, **15 (3)**, 034 035.

Gutiérrez, J. M., A. S. Cofiño, R. Cano, and M. A. Rodríguez, 2004: Clustering methods for statistical downscaling in short-range weather forecasts. *Monthly Weather Review*, **132 (9)**, 2169–2183.

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

Gutiérrez, J. M., et al., 2019: An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, **39 (9)**, 3750–3785.

Hanssen-Bauer, I., C. Achberger, R. Benestad, D. Chen, and E. Førland, 2005: Statistical downscaling of climate scenarios over scandinavia. *Climate Research*, **29 (3)**, 255–268.

Hauser, M. and A. Ray, 2017: Principles of Riemannian Geometry in Neural Networks. *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2807–2816.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess, 2006: Downscaling heavy precipitation over the united kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **26 (10)**, 1397–1415.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hertig, E. and J. Jacobeit, 2008: Assessments of mediterranean precipitation changes for the 21st century using statistical downscaling techniques. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **28 (8)**, 1025–1045.

Hertig, E., S. Seubert, A. Paxian, G. Vogt, H. Paeth, and J. Jacobeit, 2013: Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: Statistical assessments for the mediterranean area. *Theoretical and Applied Climatology*, **111 (1)**, 1–20.

Hertig, E., et al., 2019: Comparison of statistical downscaling methods with respect to extreme events over Europe: Validation results from the perfect predictor experiment of the COST Action VALUE. *International Journal of Climatology*, **39 (9)**, 3846–3867.

Hewage, P., M. Trovati, E. Pereira, and A. Behera, 2021: Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, **24 (1)**, 343–366.

Hewitson, B. and R. Crane, 2002: Self-Organizing Maps: Applications to synoptic climatology. *Climate Research*, **22**, 13–26.

Hewitson, B. C. and R. G. Crane, 1996: Climate downscaling: techniques and application. *Climate Research*, **7 (2)**, 85–95.

Hinton, G. E., J. L. McClelland, and D. E. Rumelhart, 1990: Distributed Representations. *The Philosophy of Artificial Intelligence.*

Hinton, G. E., S. Osindero, and Y.-W. Teh, 2006: A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18 (7)**, 1527–1554.

Hochreiter, S., 1998: The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **6**, 107–116.

Hochreiter, S. and J. Schmidhuber, 1997: Long short-term memory. *Neural computation*, **9 (8)**, 1735–1780.

Hope, P. K., 2006: Projected future changes in synoptic systems influencing southwest western australia. *Climate Dynamics*, **26 (7-8)**, 765–780.

Hornik, K., 1991: Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4 (2)**, 251–257.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew, 2006: Extreme learning machine: Theory and applications. *Neurocomputing*, **70 (1)**, 489–501.

Hutengs, C. and M. Vohland, 2016: Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, **178**, 127–141.

Huth, R., 1999: Statistical downscaling in central Europe: evaluation of methods and potential predictors. *Climate Research*, **13**, 91–101.

Huth, R., 2002: Statistical downscaling of daily temperature in central europe. *Journal of Climate*, **15 (13)**, 1731–1742.

Huth, R., 2004: Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. *Journal of Climate*, **17 (3)**, 640–652.

Huth, R., 2005: Downscaling of humidity variables: a search for suitable predictors and predictands. *International Journal of Climatology*, **25 (2)**, 243–250.

Huth, R., S. Kliegrova, and L. Metelka, 2008: Non-linearity in statistical downscaling: does it bring an improvement for daily temperature in europe? *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **28 (4)**, 465–477.

Iizumi, T., M. Nishimori, K. Dairaku, S. A. Adachi, and M. Yokozawa, 2011: Evaluation and intercomparison of downscaled daily precipitation indices over japan in present-day climate: Strengths and weaknesses of dynamical and bias correction-type statistical downscaling methods. *Journal of Geophysical Research: Atmospheres*, **116 (D1)**.

Ioffe, S. and C. Szegedy, 2015: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*.

Irrgang, C., N. Boers, M. Sonnewald, E. A. Barnes, C. Kadow, J. Staneva, and J. Saynisch-Wagner, 2021: Will artificial intelligence supersede earth system and climate models? *arXiv preprint arXiv:2101.09126*.

Iturbide, M., et al., 2019: The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, **111**, 42–54.

Jacob, D., et al., 2020: Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Regional Environmental Change*, **20 (2)**, 51.

Jaitly, N. and G. E. Hinton, 2013: Vocal tract length perturbation (vtlp) improves speech recognition. *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, Vol. 117, 21.

Jones, P. W., 1999: First-and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, **127 (9)**, 2204–2210.

Keller, D. E., A. M. Fischer, M. A. Liniger, C. Appenzeller, and R. Knutti, 2017: Testing a weather generator for downscaling climate change projections over switzerland. *International Journal of Climatology*, **37 (2)**, 928–942.

Kharin, V. and F. Zwiers, 2003: On the roc score of probability forecasts. *Journal of Climate*, **16**, 4145–4150.

Kilsby, C., et al., 2007: A Daily Weather Generator for Use in Climate Change Studies. *Environmental Modelling and Software*, **22**, 1705–1719.

Kingma, D. P. and M. Welling, 2013: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klein Tank, A., E. Manzini, P. Braconnot, F. Doblas-Reyes, T. Buishand, and A. Morse, 2009: Evaluation of the ensembles prediction system. *ENSEMBLES: Climate Change and its Impacts-Summary of research and results from the ENSEMBLES project*, Met Office Hadley Centre, 95–106.

Koenker, R. and K. F. Hallock, 2001: Quantile regression. *Journal of economic perspectives*, **15 (4)**, 143–156.

Koller, D. and N. Friedman, 2009: *Probabilistic graphical models: principles and techniques*. MIT press.

Kotlarski, S., et al., 2014: Regional climate modeling on European scales : A joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geoscientific Model Development*, **7**, 1297–1333.

Krasnopolsky, V. M. and M. S. Fox-Rabinovitz, 2006: A new synergetic paradigm in environmental numerical modeling: Hybrid models combining deterministic and machine learning components. *Ecological Modelling*, **191 (1)**, 5–18.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012: ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 1097–1105.

Kumar, B., R. Chattopadhyay, M. Singh, N. Chaudhari, K. Kodari, and A. Barve, 2021: Deep learning–based downscaling of summer monsoon rainfall data over indian region. *Theoretical and Applied Climatology*, **143 (3)**, 1145–1156.

Kühnlein, M., T. Appelhans, B. Thies, and T. Nauss, 2014: Improving the accuracy of rainfall rates from optical satellite sensors with machine learning — A random forests-based approach applied to MSG SEVIRI. *Remote Sensing of Environment*, **141**, 129–143.

Lafon, T., S. Dadson, G. Buys, and C. Prudhomme, 2013: Bias correction of daily precipitation simulated by a regional climate model: a comparison of methods. *International Journal of Climatology*, **33 (6)**, 1367–1381.

Landschützer, P., N. Gruber, D. C. E. Bakker, U. Schuster, S. Nakaoka, M. R. Payne, T. P. Sasse, and J. Zeng, 2013: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences*, **10 (11)**, 7793–7815.

Larraondo, P. R., L. J. Renzullo, I. Inza, and J. A. Lozano, 2019: A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv preprint arXiv:1903.10274*.

LeCun, Y., Y. Bengio, et al., 1995: Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, **3361 (10)**, 1995.

Lee, R. W., 2015: Storm track biases and changes in a warming climate from an extratropical cyclone perspective using cmip5. Ph.D. thesis, University of Reading.

Lguensat, R., J. L. Sommer, S. Metref, E. Cosme, and R. Fablet, 2019: Learning generalized quasi-geostrophic models using deep neural numerical models. *arXiv preprint arXiv:1911.08856*.

Liu, Y., et al., 2016: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of atmospheric sciences*, **20 (2)**, 130–141.

MacKay, D. J., 1997: Gaussian processes-a replacement for supervised neural networks?

Manzanas, R., L. Fiwa, C. Vanya, H. Kanamaru, and J. M. Gutiérrez, 2020a: Statistical downscaling or bias adjustment? a case study involving implausible climate change projections of precipitation in malawi. *Climatic Change*, **162 (3)**, 1437–1453.

Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, 2014: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *Journal of Geophysical Research: Atmospheres*, **119 (4)**, 1708–1719.

Manzanas, R., J. Gutiérrez, J. Fernández, E. Van Meijgaard, S. Calmanti, M. Magariño, A. Cofiño, and S. Herrera, 2018: Dynamical and statistical downscaling of seasonal temperature forecasts in europe: Added value for user applications. *Climate Services*, **9**, 44–56.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, E. Penabad, and A. Brookshaw, 2020b: Statistical adjustment, calibration and downscaling of seasonal forecasts: a case-study for southeast asia. *Climate Dynamics*, **54 (5)**, 2869–2882.

Maraun, D. and M. Widmann, 2018: *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press.

Maraun, D., M. Widmann, and J. M. Gutiérrez, 2019: Statistical downscaling skill under present climate conditions: A synthesis of the VALUE perfect predictor experiment. *International Journal of Climatology*, **39 (9)**, 3692–3703.

Maraun, D., et al., 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48 (3)**.

Maraun, D., et al., 2015: VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, **3 (1)**, 1–14.

Maraun, D., et al., 2017a: Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, **7 (11)**, 764–773.

Maraun, D., et al., 2017b: The VALUE perfect predictor experiment: Evaluation of temporal variability. *International Journal of Climatology*, **39 (9)**, 3786–3818.

Markatou, M., H. Tian, S. Biswas, and G. Hripcsak, 2005: Analysis of variance of cross-validation estimators of the generalization error. *Journal of Machine Learning Research*, **6 (Jul)**, 1127–1168.

McCulloch, W. S. and W. Pitts, 1943: A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5 (4)**, 115–133.

McFarlane, N., 2011: Parameterizations: representing key processes in climate models without resolving them. *WIREs Climate Change*, **2 (4)**, 482–497.

Meyer, D., R. J. Hogan, P. D. Dueben, and S. L. Mason, 2021: Machine learning emulation of 3d cloud radiative effects. *arXiv preprint arXiv:2103.11919*.

Miao, Q., B. Pan, H. Wang, K. Hsu, and S. Sorooshian, 2019: Improving Monsoon Precipitation Prediction Using Combined Convolutional and Long Short Term Memory Neural Network. *Water*, **11 (5)**, 977.

Mirza, M. and S. Osindero, 2014: Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Montavon, G., W. Samek, and K.-R. Müller, 2018: Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, **73**, 1–15.

Monteleoni, C., G. A. Schmidt, and S. McQuade, 2013: Climate Informatics: Accelerating Discovering in Climate Science with Machine Learning. *Computing in Science & Engineering*, **15 (5)**, 32–40.

Mu, B., B. Qin, S. Yuan, and X. Qin, 2020: A climate downscaling deep learning model considering the multiscale spatial correlations and chaos of meteorological events. *Mathematical Problems in Engineering*, **2020**.

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

Müller, W., et al., 2018: A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM 1.2-HR). *Journal of Advances in Modeling Earth Systems*, **10**.

Nair, V. and G. E. Hinton, 2010: Rectified Linear Units Improve Restricted Boltzmann Machines. *ICML*.

Nelder, J. A. and R. W. M. Wedderburn, 1972: Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135 (3)**, 370–384.

Nikulin, G., et al., 2018: Dynamical and statistical downscaling of a global seasonal hindcast in eastern africa. *Climate Services*, **9**, 72–85.

Nowack, P., J. Runge, V. Eyring, and J. D. Haigh, 2020: Causal networks for climate model evaluation and constrained projections. *Nature communications*, **11 (1)**, 1–11.

Olmo, M. E. and M. L. Bettolli, 2021: Statistical downscaling of daily precipitation over southeastern south america: assessing the performance in extreme events. *International Journal of Climatology*.

Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate data challenges in the 21st century. *science*, **331 (6018)**, 700–702.

Pan, B., K. Hsu, A. AghaKouchak, and S. Sorooshian, 2019: Improving precipitation estimation using convolutional neural network. *Water Resources Research*, **55 (3)**, 2301–2321.

Panofsky, H. A., G. W. Brier, and W. H. Best, 1958: Some application of statistics to meteorology.

Papale, D. and R. Valentini, 2003: A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, **9 (4)**, 525–535.

Pérez, F. and B. E. Granger, 2007: IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, **9 (3)**, 21–29.

Piani, C., J. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over europe. *Theoretical and Applied Climatology*, **99 (1)**, 187–192.

Pinto, I., C. Jack, and B. Hewitson, 2018: Process-based model evaluation and projections over southern africa from coordinated regional climate downscaling experiment and coupled model intercomparison project phase 5 models. *International Journal of Climatology*, **38 (11)**, 4251–4261.

Pradhan, R., R. S. Aygun, M. Maskey, R. Ramachandran, and D. J. Cecil, 2017: Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing*, **27 (2)**, 692–702.

Preisendorfer, R. W. and C. D. Mobley, 1988: Principal component analysis in meteorology and oceanography. *Developments in atmospheric science*, **17**.

Quesada-Chacón, D., K. Barfus, and C. Bernhofer, 2021: Climate change projections and extremes for costa rica using tailored predictors from cordex model output through statistical downscaling with artificial neural networks. *International Journal of Climatology*, **41 (1)**, 211–232.

Ranzato, M., C. Poultney, S. Chopra, Y. LeCun, et al., 2007: Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, **19**, 1137.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, **12 (11)**.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, **115 (39)**, 9684–9689.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al., 2019: Deep learning and process understanding for data-driven earth system science. *Nature*, **566 (7743)**, 195–204.

Reimers, C., J. Runge, and J. Denzler, 2019: Using causal inference to globally understand black box predictors beyond saliency maps. *Proceedings of the 9th International Workshop on Climate Informatics (2019)*, Vol. 6.

Rodrigues, E. R., I. Oliveira, R. Cunha, and M. Netto, 2018: Deepdownscale: a deep learning strategy for high-resolution weather forecast. *2018 IEEE 14th International Conference on e-Science (e-Science)*, IEEE, 415–422.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Rosenblatt, F., 1958: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, **65 (6)**, 386.

Ruder, S., 2017: An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986: Learning representations by back-propagating errors. *nature*, **323 (6088)**, 533–536.

Rummukainen, M., 2010: State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (1)**, 82–96.

Sachindra, D., K. Ahmed, M. M. Rashid, S. Shahid, and B. Perera, 2018: Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, **212**, 240–258.

Salathé Jr, E. P., 2005: Downscaling simulations of future global climate with application to hydrologic modelling. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **25 (4)**, 419–436.

San-Martín, D., R. Manzanas, S. Brands, S. Herrera, and J. M. Gutiérrez, 2017: Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *Journal of Climate*, **30 (1)**, 203–223.

Scher, S., 2018: Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, **45 (22)**, 12–616.

Scher, S. and G. Messori, 2018: Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, **144 (717)**, 2830–2841.

Scher, S. and G. Messori, 2019: Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, **12 (7)**, 2797–2809.

Schmidli, J., C. Goodess, C. Frei, M. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith, 2007: Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the european alps. *Journal of Geophysical Research: Atmospheres*, **112 (D4)**.

Schneider, T., S. Lan, A. Stuart, and J. Teixeira, 2017: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, **44 (24)**, 12–396.

Schoof, J. T. and S. C. Pryor, 2001: Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **21 (7)**, 773–790.

Seifert, A. and S. Rasp, 2020: Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, **12 (12)**.

Serifi, A., T. Günther, and N. Ban, 2021: Spatio-temporal downscaling of climate data using convolutional and error-predicting neural networks. *Frontiers in Climate*, **3**, 26.

Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020a: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part i: Daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology*, **59 (12)**, 2057–2073.

Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020b: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part ii: Daily precipitation. *Journal of Applied Meteorology and Climatology*, **59 (12)**, 2075–2092.

Sietsma, J. and R. J. Dow, 1991: Creating artificial neural networks that generalize. *Neural networks*, **4 (1)**, 67–79.

Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualising image classification models and saliency maps. *In Workshop at International Conference on Learning Representations*, Citeseer.

Soares, P. M., et al., 2019: Process-based evaluation of the value perfect predictor experiment of statistical downscaling methods. *International Journal of Climatology*, **39 (9)**, 3868–3893.

Sørland, S. L., C. Schär, D. Lüthi, and E. Kjellström, 2018: Bias patterns and climate change signals in gcm-rcm model chains. *Environmental Research Letters*, **13 (7)**, 074 017.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, 2014: Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15 (1)**, 1929–1958.

Steininger, M., D. Abel, K. Ziegler, A. Krause, H. Paeth, and A. Hotho, 2020: Deep learning for climate model output statistics. *arXiv preprint arXiv:2012.10394*.

Stengel, K., A. Glaws, D. Hettinger, and R. N. King, 2020: Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, **117 (29)**, 16 805–16 815.

Sun, L. and Y. Lan, 2021: Statistical downscaling of daily temperature and precipitation over china using deep learning neural models: Localization and comparison with other methods. *International Journal of Climatology*, **41 (2)**, 1128–1147.

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

Terray, L. and J. Boé, 2013: Quantifying 21st-century france climate change and related uncertainties. *Comptes Rendus Geoscience*, **345 (3)**, 136–149.

Teutschbein, C. and J. Seibert, 2012: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, **456-457**, 12–29.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58 (1)**, 267–288.

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

Timbal, B. and B. McAvaney, 2001: An analogue-based method to downscale surface air temperature: application for australia. *Climate Dynamics*, **17 (12)**, 947–963.

Toms, B. A., E. A. Barnes, and J. W. Hurrell, 2021: Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophysical Research Letters*.

Touretzky, D. S. and G. E. Hinton, 1985: Symbols among the neurons: Details of a connectionist inference architecture. *IJCAI*, Vol. 85, 238–243.

Tran Anh, D., S. P. Van, T. D. Dang, and L. P. Hoang, 2019: Downscaling rainfall using deep learning long short-term memory and feedforward neural network. *International Journal of Climatology*, **39 (10)**, 4170–4188.

Tripathi, S., S. Venkata, and R. Nanjundiah, 2006: Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach. *Journal of Hydrology*, **330**, 621–640.

Turco, M., M. C. Llasat, S. Herrera, and J. M. Gutiérrez, 2017: Bias correction and downscaling of future rcm precipitation projections using a mos-analog technique. *Journal of Geophysical Research: Atmospheres*, **122 (5)**, 2631–2648.

Uvo, C., J. Olsson, O. Morita, K. Jinno, A. Kawamura, K. Nishiyama, N. Koreeda, and T. Nakashima, 2001: Statistical atmospheric downscaling for rainfall estimation in Kyushu Island, Japan. *Hydrology and Earth System Sciences*, **5**.

Vaittinada Ayar, P., M. Vrac, S. Bastin, J. Carreau, M. Déqué, and C. Gallardo, 2016: Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: present climate evaluations. *Climate Dynamics*, **46 (3)**, 1301–1329.

Van Vuuren, D. P., et al., 2011: The representative concentration pathways: an overview. *Climatic change*, **109 (1)**, 5–31.

Vandal, T., E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly, 2018a: Quantifying uncertainty in discrete-continuous and skewed data with bayesian deep learning. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2377–2386.

Vandal, T., E. Kodra, and A. R. Ganguly, 2017: Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *arXiv preprint arXiv:1702.04018*.

Vandal, T., E. Kodra, and A. R. Ganguly, 2019: Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, **137 (1)**, 557–570.

Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, 2018b: Generating high resolution climate change projections through single image super-resolution: An abridged version. *International Joint Conferences on Artificial Intelligence Organization*.

Vaughan, A., W. Tebbutt, J. S. Hosking, and R. E. Turner, 2021: Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development Discussions*, 1–25.

Vesely, F. M., L. Paleari, E. Movedi, G. Bellocchi, and R. Confalonieri, 2019: Quantifying uncertainty due to stochastic weather generators in climate change impact studies. *Scientific reports*, **9 (1)**, 1–8.

Voldoire, A., et al., 2013: The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics*, **40 (9-10)**, 2091–2121.

Vrac, M. and P. Ayar, 2016: Influence of Bias Correcting Predictors on Statistical Downscaling Models. *Journal of Applied Meteorology and Climatology*, **56**.

Vrac, M., P. Drobinski, A. Merlo, M. Herrmann, C. Lavaysse, L. Li, and S. Somot, 2012: Dynamical and statistical downscaling of the french mediterranean climate: uncertainty assessment. *Natural Hazards and Earth System Sciences*, **12 (9)**, 2769–2784.

Vrac, M., P. Marbaix, D. Paillard, and P. Naveau, 2007a: Non-linear statistical downscaling of present and lgm precipitation and temperatures over europe. *Climate of the Past*, **3 (4)**, 669–682.

Vrac, M., M. Stein, K. Hayhoe, and X.-Z. Liang, 2007b: A general method for validating statistical downscaling methods under future climate change. *Geophysical Research Letters*, **34 (18)**.

Walton, D., N. Berg, D. Pierce, E. Maurer, A. Hall, Y.-H. Lin, S. Rahimi, and D. Cayan, 2020: Understanding differences in california climate projections produced by dynamical and statistical downscaling. *Journal of Geophysical Research: Atmospheres*, **125 (19)**, e2020JD032 812.

Wang, F., D. Tian, L. Lowe, L. Kalin, and J. Lehrter, 2021: Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, **57 (4)**, e2020WR029 308.

Weyn, J. A., D. R. Durran, and R. Caruana, 2019: Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, **11 (8)**, 2680–2693.

Widmann, M., et al., 2019: Validation of spatial variability in downscaling results from the value perfect predictor experiment. *International Journal of Climatology*, **39 (9)**, 3819–3845.

Wilby, R. and T. Wigley, 2000: Precipitation Predictors for Downscaling: Observed and General Circulation Model Relationships. *International Journal of Climatology*, **20**, 641–661.

Wilby, R. L., T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, **34 (11)**, 2995–3008.

Wilkinson, M. D., et al., 2016: The fair guiding principles for scientific data management and stewardship. *Scientific data*, **3 (1)**, 1–9.

Wilks, D. S., 2010: Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, **1 (6)**, 898–907.

Wilks, D. S. and R. L. Wilby, 1999: The weather generation game: a review of stochastic weather models. *Progress in physical geography*, **23 (3)**, 329–357.

Williams, P. M., 1998: Modelling seasonality and trends in daily rainfall data. *Advances in neural information processing systems*, 985–991.

Xingjian, S., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, 2015: Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 802–810.

Yang, C., N. Wang, S. Wang, and L. Zhou, 2018: Performance comparison of three predictor selection methods for statistical downscaling of daily precipitation. *Theoretical and Applied Climatology*, **131 (1)**, 43–54.

Yosinski, J., J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, 2015: Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Yuval and W. Hsieh, 2006: The impact of time-averaging on the detectability of nonlinear empirical relations. *Quarterly Journal of the Royal Meteorological Society*, **128**, 1609–1622.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals, 2021: Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, **64 (3)**, 107–115.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, 2016: Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zintgraf, L. M., T. S. Cohen, T. Adel, and M. Welling, 2017: Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.

Zorita, E. and H. Von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of climate*, **12 (8)**, 2474–2489.

IFCA

Instituto de Física de Cantabria

Santander Meteorology Group
http://www.meteo.unican.es
Santander (Spain)
2021