UNIVERSIDAD NACIONAL DE TUCUMÁN
FACULTAD DE CIENCIAS EXACTAS Y TECNOLOGÍA (FACET)

Aplicaciones de métodos de aprendizaje automático al estudio de variables meteorológicas en la región del NOA

Tesis presentada para obtener el título de Licenciada en Física

Autora:
María Ailén Vega Caro

Director:
Dr. Bruno S. Zossi

Codirector:
Dr. Franco D. Medina

Tucumán – Argentina
Octubre 2025

Resumen

Este trabajo se centra en la evaluación y comparación de distintos métodos de interpolación espacial aplicados a la precipitación acumulada anual en la región del Noroeste Argentino (NOA), una zona caracterizada por una topografía compleja y una red de estaciones meteorológicas escasa y distribuida de manera irregular. El objetivo principal fue analizar la capacidad de modelos de aprendizaje automático y métodos tradicionales para generar estimaciones precisas de esta variable climática. Se trabajó con una base de datos integrada por observaciones provenientes de estaciones meteorológicas en el periodo temporal 1993-2019 para el entrenamiento de los modelos y su testeo. Los métodos implementados incluyeron Kriging, Procesos Gaussianos (GPR) y Redes Neuronales (NN). En el caso de Kriging, se desarrollaron y compararon cuatro variantes: modelos con y sin deriva de altitud, combinados con un modelado isotrópico y anisotrópico. Esto se realizó con el fin de evaluar el impacto de considerar la altitud como variable explicativa y de modelar la direccionalidad en la variabilidad espacial. Además, los resultados se compararon con los obtenidos a partir de la base de reanálisis ERA5-LAND, un modelo basado en la física del sistema climático ampliamente utilizado por la facilidad de acceso a sus datos.

Los resultados muestran que el mejor desempeño general se obtuvo con el modelo NN, que presentó los menores errores absolutos medios y mayor capacidad para reproducir patrones espaciales y temporales. Por otro lado, los modelos de Kriging mostraron resultados mixtos que variaron en función de la ubicación geográfica de las estaciones de testeo. En particular, la inclusión de la altitud como deriva en general mejoró los resultados en zonas con grandes variaciones altitudinales, mientras que el modelado de la anisotropía no siempre implicó una mejora significativa, posiblemente debido a la distribución irregular de datos. Por su parte, la base de reanálisis ERA5-LAND presentó desempeños consistentemente inferiores a los de los otros modelos, lo que resalta el valor de los modelos simples desarrollados localmente.

Este trabajo aporta evidencia sobre el valor de integrar técnicas físicas y computacionales para el estudio del clima regional. En particular, destaca el potencial de las redes neuronales como herramienta complementaria para la generación de mapas climáticos en áreas con limitaciones observacionales, y sugiere líneas futuras de investigación como el entrenamiento por regiones o la incorporación de más variables físicas (como la temperatura, humedad, etc.) al entrenamiento de los modelos.

Índice

Capítulo 1: Introducción

La variabilidad de la precipitación es un tema ampliamente estudiado debido a sus importantes implicaciones socioeconómicas. Particularmente, la precipitación está asociada al riesgo de seguridad hídrica (O'Neill et al., 2022), lo cual incluye tanto eventos de sequía como inundaciones. En Sudamérica, en el periodo 1970-2021, se reportaron 115,2 billones de dólares en pérdidas económicas por eventos relacionados al clima, siendo la mayoría de estos relacionados a cuestiones hídricas (Organización Meteorológica Mundial, 2021).

La precipitación tiene una fuerte variabilidad espacio-temporal en respuesta a factores locales, regionales y globales en un gran rango de escalas temporales. Estos factores pueden ser naturales (topografía, radiación solar, etc.) o antropogénicos (aumento de gases de efecto invernadero, deforestación, etc.). Por ejemplo, a nivel global y en la escala de tiempo centenaria, la precipitación está aumentando en una tasa consistente con lo esperable como respuesta al calentamiento global generado por la actividad antropogénica (~7%/°C) (Fowler et al., 2021). Sin embargo, en escalas multidecadales, variaciones del tipo oscilatorio, como la Oscilación Decadal del Pacífico (Mantua y Hare, 2002), pueden superponerse a esta señal antropogénica y amplificar o contrarrestar los cambios de la precipitación. Por otro lado, en la escala de tiempo interanual, una confluencia de modos de variabilidad tales como El Niño Oscilación del Sur y el Modo Anular del Sur pueden generar fuertes variaciones en la lluvia entre un año y otro (Cai et al., 2020; Fogt y Marshall, 2020). La complejidad del análisis es alta debido a que existe una interacción entre los forzantes interanuales, interdecadales y de mayor plazo (por ej., Medina et al., 2025). Sumado a esta complejidad, la falta de datos de precipitación genera una importante incertidumbre observacional que dificulta la evaluación y obtención de conclusiones en algunas regiones del mundo (Skansi et al., 2013).

En vista de lo expuesto, es crucial disponer de datos de precipitación para entender su variabilidad, mejorar su predicción y disminuir los riesgos asociados. En este sentido, la disponibilidad de conjuntos de datos de alta resolución espacial es un prerrequisito para la reducción y gestión del riesgo de desastres (Shrestha et al., 2021). Estas bases de datos pueden ser mediciones interpoladas o estimaciones que provienen de combinar mediciones con modelos físicos (reanálisis, bases relacionadas a mediciones satelitales, entre otras). Sin embargo, existen muchas regiones del mundo donde estas bases de datos tienen un desempeño deficiente debido a la falta de datos medidos para la generación de las mismas, a la poca resolución espacial con la que se construyen, a fallas en los modelos de asimilación de datos,

problemas en la física de los modelos, entre otros (Funk et al., 2015; Gupta et al., 2019; Medina et al. 2023; Sun et al., 2018).

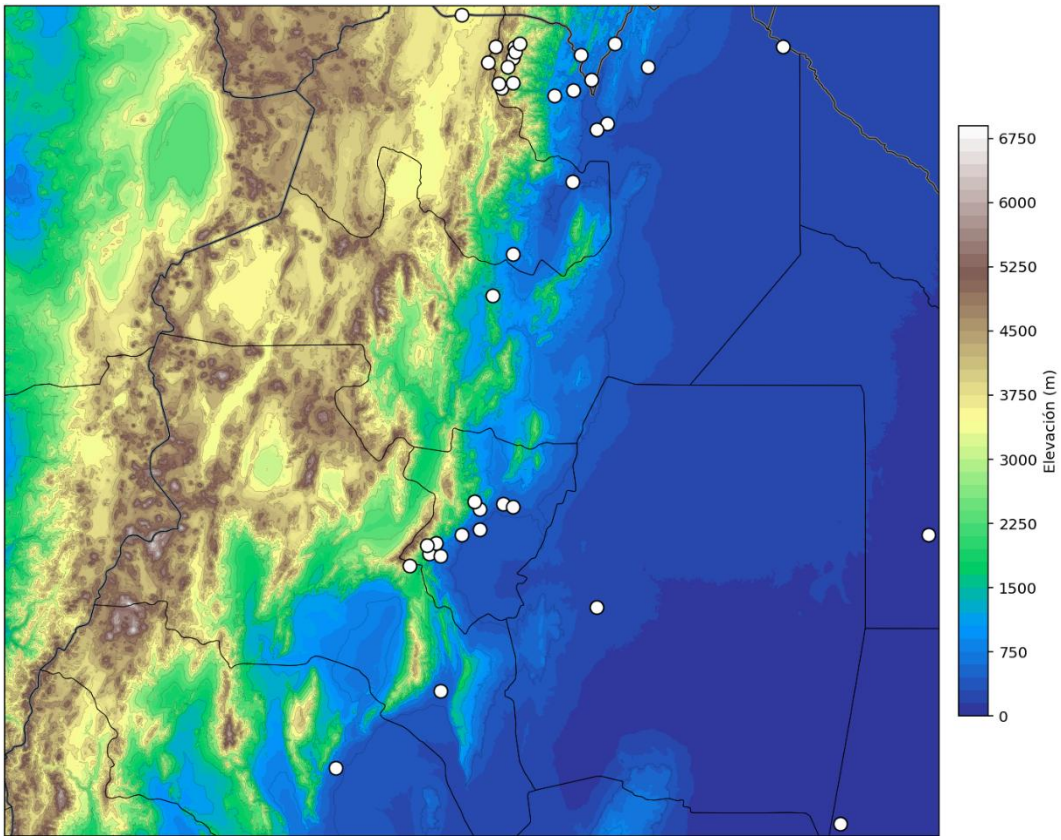

Fig.1.1. Mapa del Noroeste Argentino (NOA), se indican la altura en metros sobre el nivel del mar (color) y las estaciones meteorológicas utilizadas en este trabajo (puntos blancos).

El Noroeste Argentino (NOA), ubicado entre 20-30°S y 69-62°O (Figura 1.1), es un claro ejemplo de las fuertes variaciones espacio-temporales en la precipitación. En esta región, la Cordillera de los Andes determina un patrón espacial de precipitación con fuertes contrastes en distancias relativamente cortas. Existe un importante gradiente de alturas en la dirección este-oeste (Figura 1.1), con elevaciones que van desde los 300 m (zona este) a más de 3000 m (zona oeste). Esto a su vez genera un fuerte gradiente de precipitación (Ferrero y Villalba, 2019; Medina et al., 2023). Particularmente, los valores máximos de lluvia anual en el NOA se observan sobre la pendiente este de Los Andes, en la transición entre la llanura semiárida y el desierto andino de la Puna. Allí el aire cálido y húmedo proveniente del noreste es forzado a ascender, lo cual promueve su enfriamiento, condensación del vapor y consecuente precipitación (Minetti et al., 2009). Sobre el oeste del NOA, en altas elevaciones, las

precipitaciones anuales se encuentran por debajo de los 500 mm, mientras que en elevaciones intermedias del NOA se observan valores de hasta 2500 mm (Figura 1.2). Dada la fuerte variabilidad de la precipitación en el NOA, es necesario contar con datos de alta resolución espacial y con una gran extensión temporal que permitan evaluar estas variaciones. Si bien existen conjuntos de datos de alta resolución generados por reconocidos centros climáticos, los mismos fallan en representar la variabilidad de la lluvia en el NOA. En este sentido, Medina et al. (2023) muestran que una de las causas de esta deficiencia es la falta de observaciones que se usan para construir estas bases de datos de alta resolución. Particularmente en Argentina, la administración de la red de observaciones de precipitación está fragmentada, siendo responsabilidad de diversos organismos mantenerla y poner a disposición dichos datos (Servicio Meteorológico Nacional-SMN, Subsecretaría de Recursos Hídricos-SRH, Instituto Nacional de Tecnología Agropecuaria-INTA, entre otros). Esto provoca que el acceso a los datos sea a través de diversos mecanismos, tales como la solicitud vía e-mail (SMN) o el acceso vía diferentes páginas web (SRH e INTA), lo cual puede dificultar a los investigadores la recolección de todos los datos disponibles. Por esta razón, es de crucial importancia realizar una exhaustiva recolección de los datos de precipitación disponibles en el NOA, que luego permita la generación de bases de datos confiables y la investigación de la variabilidad de la precipitación.

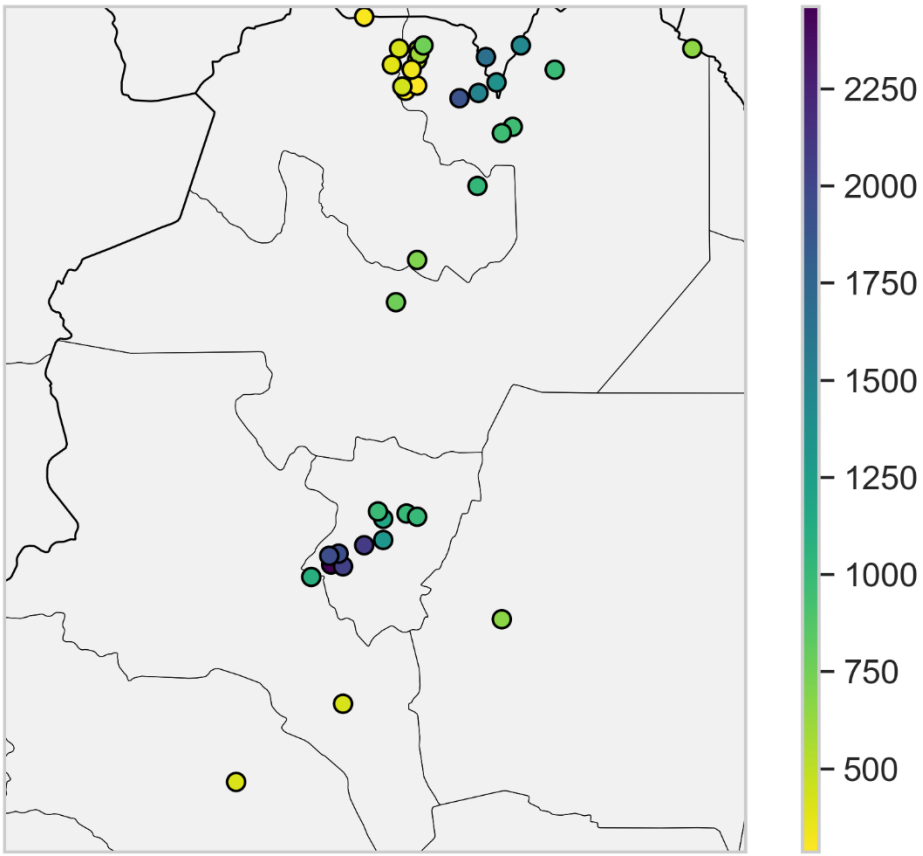

Fig. 1.2. Mapa promedio de las precipitaciones anuales acumuladas (mm) en las estaciones meteorológicas para el período (1993-2019).

Particularmente, esta Tesis de Licenciatura busca contribuir en parte a la generación de una base de datos de precipitación en el NOA. Para esto, el objetivo de la Tesis es explorar la habilidad de diversas técnicas de interpolación de datos para estimar temporal y espacialmente la precipitación anual en el NOA. También se propone comparar las estimaciones provenientes de las técnicas estadísticas propuestas con una base de datos global de alta resolución. Las técnicas seleccionadas incluyen técnicas estadísticas clásicas como Kriging Ordinario y Universal con deriva externa, y técnicas basadas en aprendizaje automático o "machine learning", como Redes Neuronales (NN) y Proceso de Regresión Gaussiana (GPR). Se espera que los resultados de este trabajo exploratorio permitan determinar las técnicas más adecuadas para estimación de patrones espaciales y temporales de precipitación en la región. Esto permitirá en un futuro la generación de una base de datos que sea mejor a las disponibles actualmente.

Capítulo 2: Marco Teórico

2.1. Redes Neuronales Artificiales

2.1.1. Introducción

Las Redes Neuronales (Neural Network, NN) son modelos no lineales inspirados en el funcionamiento de las neuronas biológicas, aplicados al aprendizaje automático, donde la unidad básica de la red es la neurona (Colamonaco, 2024; Bottaro, 2021).

El objetivo del aprendizaje automático es desarrollar algoritmos capaces de aprender automáticamente a partir de los datos, sin necesidad de programar explícitamente la relación entre las variables (Bottaro, 2021). Las Redes Neuronales se enmarcan en el aprendizaje supervisado, en el cual el objetivo es extraer la relación fundamental entre las variables de entrada y la variable objetivo. Para ello, se utilizan datos conocidos de entrada y salida durante el entrenamiento, para luego realizar predicciones con nuevos datos de entrada, donde los valores de salida son desconocidos. Es importante destacar que el modelo debe aprender, no memorizar; es decir, debe evitar el sobreajuste y el subajuste (Bottaro, 2021).

Una neurona artificial recibe información de entrada, las procesa mediante pesos, aplica una transformación no lineal mediante una función de activación y devuelve una información de salida que puede ser pasada a otras neuronas (Bottaro, 2021; Colamonaco, 2024). Una red neuronal consiste en muchas de estas neuronas apiladas y puede aproximar funciones complejas a partir de los datos de entrenamiento. La primera capa se denomina capa de entrada, las capas intermedias capas ocultas, y la última capa es la de salida (Bottaro, 2021).

Para construir una red neuronal artificial (ANN), se deben organizar las neuronas en capas y definir el tipo de conexión entre ellas. En una estructura feed-forward, la capa de entrada procesa los datos y genera valores de salida según los pesos asignados. Esta salida se utiliza como entrada para la primera capa oculta, repitiéndose el proceso hasta la capa de salida (Bottaro, 2021).

El uso de capas ocultas incrementa el poder de representación de la red, permitiendo modelar relaciones complejas entre entradas y salidas. Teóricamente, una sola capa oculta puede aproximar cualquier función continua de múltiples entradas y salidas (Teorema de Aproximación Universal) (Bottaro, 2021).

Las Redes Neuronales se pueden entender como una extensión de los modelos paramétricos de regresión lineal y logística, que resuelven problemas de regresión y clasificación (Lindholm

et al., 2022). Es por ello, que para entender cómo funcionan las redes neuronales, se deben analizar estos modelos.

2.1.2. Modelos paramétricos

Para explicar el comportamiento de los modelos paramétricos, Lindholm et al. (2022) utiliza a un modelo de regresión lineal:

$$y = f_\theta(x) + \varepsilon \tag{2.1}$$

donde θ denota explícitamente que la función depende de un conjunto de parámetros, para enfatizar que esta ecuación debe verse como un modelo de la verdadera relación entre los datos de entrada y de salida. Esta formulación permite generalizar a funciones no lineales arbitrarias, siempre que sean adaptables, es decir, dependan de un parámetro θ que controle la forma de la función. Diferentes valores de θ generan diferentes funciones. El aprendizaje consiste en encontrar el vector θ de manera que f_θ describa de manera adecuada la relación entre las entradas y la salida. En términos matemáticos se dice que es una familia paramétrica de funciones:

$$\{f_\theta(\cdot) : \theta \in \Theta\} \tag{2.2}$$

donde Θ es el espacio que contiene a todos los posibles vectores de parámetros. Una vez especificada esta familia, se debe optimizar θ para que el modelo describa correctamente la relación entre variables de entrada y salida. Esto se logra minimizando la función de costo, que se define como el promedio de alguna función de perdida L evaluada para los datos de entrenamiento.

$$\hat{\theta} = \arg\min \frac{1}{n} \sum L\big(y_i, f_\theta(x_i)\big) \tag{2.3}$$

donde L es la función de pérdida evaluada en los datos de entrenamiento. En la mayoría de los casos, esta optimización requiere métodos numéricos e iterativos, que se repiten hasta aproximar una solución.

El propósito final no es simplemente minimizar el error en los datos de entrenamiento y adaptar el modelo lo más posible a los datos de entradas, sino que el propósito es lograr un modelo que generalice a nuevos datos. Entonces, el problema que se quiere resolver es minimizar el error esperado para los nuevos datos de entrada. Como este error esperado no es conocido, la optimización sobre el conjunto de entrenamiento actúa como un intermediario.

La elección de la función de pérdida es una decisión de diseño crítica, ya que diferentes funciones producen soluciones distintas para $\hat{\theta}$. Un aspecto muy importante a la hora de elegir

la función de perdida es su robustez. Una función de pérdida es robusta si los valores atípicos tienen un bajo impacto; de lo contrario, no es robusta. Algunas funciones de pérdida comunes en problemas regresión son:

- Error cuadrático: $L(y, \hat{y}) = (\hat{y} - y)^2$
- Error absoluto: $L = |\hat{y} - y|$ (2.4)

La primera es la opción por defecto en las regresiones lineales, pues simplifica el entrenamiento. Sin embargo, no es tan utilizada en otros modelos de regresión como las redes neuronales, pues es menos robusta frente a valores atípicos. La segunda es más robusta y se utiliza frecuentemente en redes neuronales.

La función de pérdida calcula la discrepancia entre las predicciones y los valores verdaderos, siendo un elemento central para el entrenamiento y ajuste de los parámetros (Colamonaco, 2024).

Lindholm et al. (2022) destacan además algunos conceptos clave acerca de la optimización:

- Precisión de optimización vs. precisión estadística: La función de costo (J(θ)) se calcula a partir de datos de entrenamiento, que contienen ruido. Optimizar más allá del error estadístico no aporta mejoras.
- Función de pérdida vs. función de error: La función de pérdida L usada para entrenar puede diferir de la función error o métrica de evaluación E, buscando mejorar la generalización o reducir costos computacionales.
- Early stopping: Durante la optimización iterativa, puede ocurrir que al pasar de una iteración a la siguiente, se pase de una solución útil (con buenas propiedades de generalización) a una solución peor (por ejemplo, debido al sobreajuste). Para evitar esto, se puede detener el entrenamiento antes de alcanzar el mínimo de la función de costo. Esto evita sobreajuste, se denomina regularización implícita y se puede aplicar a cualquier método que se entrene con optimización iterativa. Se puede implementar usando un conjunto de validación; cuando el error en validación comienza a aumentar, se detiene el entrenamiento.
- Regularización explícita: Su propósito es favorecer los valores de parámetros más pequeños, para disminuir la flexibilidad del modelo, lo que puede mejorar el desempeño de este. Consiste en modificar la función de costo agregando un término independiente que penalice parámetros grandes, favoreciendo la generalización. La regularización busca mantener los parámetros θ lo más pequeños posible, salvo que los datos indiquen lo contrario. Es decir, si un modelo con θ pequeños ajusta a los datos igual de bien que un modelo con θ mayores, entonces se preferirá el modelo con valores más pequeños. Entre las técnicas de regularización,

la L^2 penaliza directamente la magnitud de los parámetros, agregando el término $\|\theta\|^2$ a la función de costo.

2.1.3. Redes Neuronales como extensión de modelos paramétricos

Las Redes Neuronales se pueden entender como una extensión de los modelos de regresión lineal y logística (Lindholm et al., 2022). Esto se logra apilando múltiples capas de regresiones lineales y funciones de activación no lineales, formando una estructura jerárquica capaz de modelar relaciones complejas.

En un modelo de regresión lineal se considera la siguiente relación entre la variable de salida $\hat{y}$ y las variables de entrada $x_1, \ldots, x_p$:

$$\hat{y} = W_1 x_1 + W_2 x_2 + \cdots + W_p x_p + b \qquad (2.5)$$

Donde $W_1, \ldots, W_p$ son los pesos y b es el término de offset. En lo siguiente se utilizará el valor 1 para denotar a la variable de entrada correspondiente al offset.

Para describir la relación no lineal entre los parámetros de entrada $x = [1, x_1, x_2, \ldots, x_p]^T$ y la variable de salida $\hat{y}$ se introduce una función escalar no lineal llamada función de activación. De esta manera, la regresión lineal mostrada en la ec. 2.5 es transformada en un modelo generalizado de la regresión lineal en donde la combinación lineal de las variables de entrada es transformada por la función de activación, de modo que:

$$\hat{y} = h\big(W_1 x_1 + W_2 x_2 + \cdots + W_p x_p + b\big) \qquad (2.6)$$

Algunas funciones de activación son la función logística y la función de unidad lineal rectificada (ReLU por por sus siglas en inglés), donde la función logística está dada por:

$$h(z) = \frac{1}{1 + e^{-z}} \qquad (2.7)$$

y la función ReLU está dada por:

$$h(z) = \max(0, z) \qquad (2.8)$$

En la actualidad, la función de activación ReLU es la opción más utilizada para Redes Neuronales, probablemente debido a su simplicidad.

2.1.4. Redes Neuronales Profundas

El modelo de regresión lineal generalizado que se observa en la ec. 2.6 es demasiado simple para explicar relaciones complejas entre variables de entrada y salida, por lo que se lo debe generalizar. Para describir relaciones complejas, se utilizan múltiples modelos de regresión

lineal paralelos para formar una capa, y luego se apilan capas secuencialmente para construir una red neuronal profunda.

Las redes neuronales densas, también llamadas Multi-Layer Perceptron (MLP), tienen tres o más capas, donde cada neurona se conecta con todas las neuronas de la capa anterior. El ancho (número de neuronas por capa) y la profundidad (número de capas) determinan la complejidad y la no linealidad que la red puede representar, incrementando el número total de parámetros que deben ajustarse (Colamonaco, 2024).

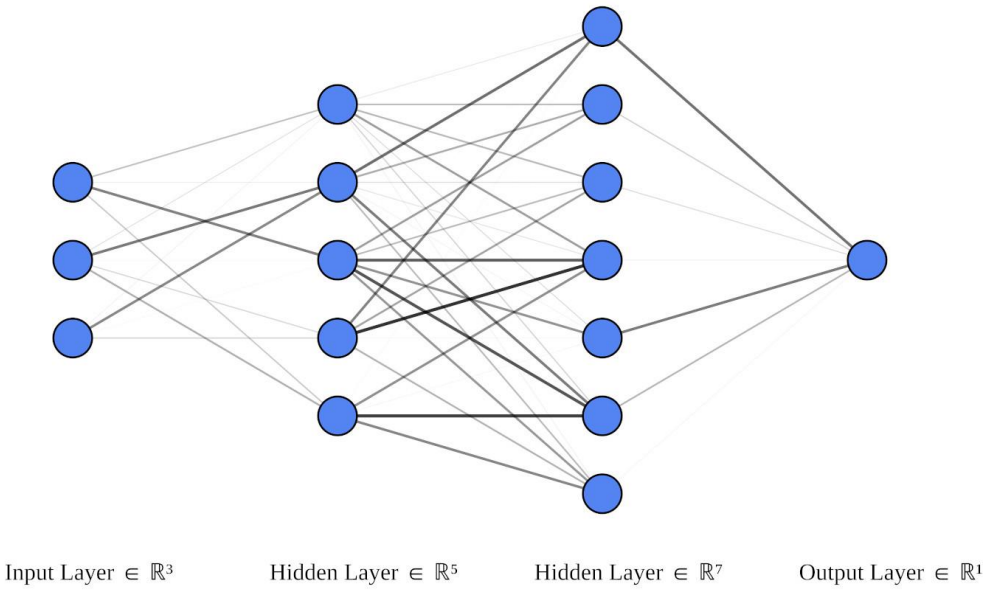

Fig 2.1. Esquema explicativo de la arquitectura de una red neuronal.

2.2. Proceso Gaussiano de Regresión (GPR)

2.2.1. Introducción

Los procesos gaussianos (GP) son métodos de aprendizaje automático supervisado de tipo no paramétrico, lo que significa que no suponen una forma funcional fija entre las variables de entrada y salida. En cambio, la relación se aprende directamente a partir de los datos de entrada y salida (Rasmussen & Williams, 2006).

Dado un conjunto de entrenamiento, formado por datos de entradas y salida, el objetivo es hacer predicciones para nuevos puntos de entrada que no están en este conjunto. Para lograr esto, se busca una función que describa cómo se relacionan los datos y que genere predicciones para todos los valores de entrada posibles. Como hay infinitas funciones posibles que podrían ajustarse a los datos, es necesario introducir algún tipo de supuesto sobre cuáles de ellas son más probables.

Una manera de encarar este problema del proceso de aprendizaje supervisado es otorgar a cada posible función una determinada probabilidad. Esta se asigna en función de cual se cree que sea más probable que represente correctamente la relación entre los datos. Es para esto que se utilizan los procesos gaussianos, que son una generalización de la distribución de probabilidad gaussiana que asignan una distribución de probabilidad sobre funciones. Mientras que una distribución gaussiana clásica describe la probabilidad de valores numéricos o vectores, un proceso estocástico, como un Proceso Gaussiano, determina las propiedades de las funciones y la probabilidad de *funciones completas*.

Dentro del aprendizaje supervisado, los procesos pueden aplicarse tanto a problemas de regresión (donde la salida es continua) como de clasificación (donde la salida es discreta). En este trabajo se utiliza el caso de regresión, denominado Proceso Gaussiano de Regresión (GPR).

En este enfoque, se parte de una distribución a *priori* sobre las posibles funciones. Luego, al observar los datos, se actualiza esta distribución para obtener una distribución a *posteriori*, que representa las funciones más probables dado lo observado. Entonces, se puede pensar que el proceso "descarta" aquellas funciones que no se ajustan a los datos. Luego, la distribución a posteriori se ajusta de modo que la función media pase por los puntos observados y la incertidumbre se reduzca cerca de las observaciones. De esta manera, esta distribución se ajusta considerando únicamente las funciones que se adecúan a las observaciones.

Este método es no paramétrico, lo que significa que el modelo no tiene que preocuparse por si puede ajustarse a los datos, a diferencia de un modelo paramétrico, como una regresión lineal, que podría fallar con datos no lineales.

Una parte esencial del proceso es la distribución a *priori*, que refleja los supuestos sobre la forma que se espera que tengan las funciones: si son suaves, si su variación es rápida y abrupta, etc. Esta información se relaciona con la función de covarianza (o kernel), que define cómo se relacionan los valores de la función en distintos puntos del espacio de entrada. Ajustar correctamente esta función es fundamental en el proceso gaussiano.

Al realizar un modelado, es más realista considerar que los datos tienen ruido, ya que no conocemos los valores de las funciones, si no que conocemos sus versiones ruidosas. Es importante destacar que, si se asume un ruido de observación con distribución gaussiana, entonces las predicciones del modelo pueden obtenerse de forma analítica. La distribución predictiva para los objetivos con ruido se calcula añadiendo la varianza del ruido (σn^2) a la varianza predictiva de la función subyacente.

El objetivo del proceso gaussiano es reconstruir la función subyacente eliminando el ruido en las observaciones, promediando las observaciones ruidosas de manera ponderada Como las predicciones son combinaciones lineales de los valores observados, la regresión gaussiana puede entenderse como un suavizador lineal. El comportamiento de este suavizado puede entenderse a través del kernel equivalente, que proporciona una ponderación a las observaciones localmente alrededor de un punto de prueba, es decir, determina qué tan fuerte se relacionan los puntos cercanos entre sí.

2.2.2. Funciones de covarianza

La función de covarianza (o kernel) es fundamental para el proceso gaussiano, ya que encierra los supuestos que se tienen de la función que se desea encontrar y define cómo se relacionan los valores de salida con los distintos puntos del espacio de entrada. En otras palabras, representa la noción de similitud entre los datos: si dos puntos de entrada son cercanos, se espera que las salidas tengan valores similares y el kernel asignará una alta covarianza. Por lo tanto, los puntos de entrenamiento que se encuentren cerca de un punto de testeo serán cruciales para la predicción en ese punto.

En este trabajo sólo se considerarán funciones de covarianza estacionarias, es decir, aquellas que son invariantes respecto de las traslaciones en el espacio de entradas (solo dependen de la distancia entre los puntos y no de su posición absoluta). Si además la función solo depende de la magnitud de la distancia entre los puntos, se llama isotrópica. Una función de covarianza isotrópica es invariante respecto de todas las rotaciones rígidas y se conocen como funciones de base radial (RBFs).

Las propiedades del kernel determinan la suavidad del proceso: un kernel más suave genera funciones más continuas y regulares, mientras que uno menos suave produce funciones más abruptas o irregulares. Matemáticamente, esto se refleja en la continuidad y diferenciabilidad en media cuadrática del proceso. Para definir ambos conceptos se considera:

- una secuencia de puntos [x1, x2, …] pertenecientes al dominio de entrada y
- un punto fijo x* perteneciente al dominio de entrada, de modo que los puntos de la secuencia se acercan a x*, es decir, $|x_k - x_*| \to 0$ cuando $k \to \infty$, siendo k el subindice que señala a cada punto de la secuencia.

Un proceso f(x) se considera continuo en media cuadrática en un punto x* si la esperanza $E[|f(x_k) - f(x*)|^2] \to 0$, a medida que $k \to \infty$. Es decir, a medida que x_k se aproxima a x*.

Esto ocurre, si y sólo si, su función de covarianza k(x,x') es continua en el punto $x = x' = x*$. Para funciones de covarianza estacionarias, este problema se reduce a verificar la continuidad en k(0).

Por otro lado, la derivada en media cuadrática de f(x) en una dirección específica *i* se define como el límite en media cuadrática del cociente diferencial:

$$\frac{\partial f(x)}{\partial x_i} = \text{l.i.m}_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{2.9}$$

Donde e_i es el vector unidad en la dirección analizada y l.i.m indica el límite en media cuadrática.

La covarianza de la derivada está dada por:

$$Cov\left(\frac{\partial f(x)}{\partial x_i}, \frac{\partial f(x')}{\partial x_i}\right) = \frac{\partial^2 k(x, x')}{\partial x_i \partial x_i'} \tag{2.10}$$

Por lo tanto, si la segunda derivada de $k(x, x')$ existe y es finita, el proceso $f(x)$ será diferenciable en media cuadrática.

Estas definiciones se pueden extender a derivadas de orden superior. Si un proceso es estacionario y las derivadas parciales de orden 2k de k(x) existen y son finitas en x=0, entonces las derivadas parciales de orden k de f(x) existen como límites en media cuadrática para todo x perteneciente al dominio de entrada.

Las propiedades de la función de covarianza (suavidad y diferenciabilidad) alrededor del cero son las que determinan la suavidad del proceso estacionario. Un k más suave alrededor de cero, generará un proceso más suave, mientras que, si el k es abrupto alrededor de cero, el proceso será más 'brusco'.

Las funciones de covarianza analizadas en este trabajo son: Exponencial Cuadrada, Matérn, Cuadrática Racional, a continuación, se desarrollará una breve explicación de cada una de ellas.

a) Función de Covarianza Exponencial Cuadrada (SE)

La función de covarianza exponencial cuadrada, o Radial Basis Function (RBF), tiene la forma:

$$k_{SE}(r) = exp\left(-\frac{r^2}{l^2}\right) \tag{2.11}$$

En donde $r = |x - x'|$ y l es la longitud de escala característica. Esta función de covarianza es infinitamente diferenciable, por lo que las funciones que modela son muy suaves. A pesar

de que esta función de covarianza es una de las más utilizadas en el aprendizaje automático, no siempre es adecuada para algunos procesos físicos, como la turbulencia.

b) Clase de Funciones de Covarianza de Matérn

Las funciones de covarianza de la clase Matérn están dadas por la siguiente ecuación:

$$k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}r}{l}\right) \qquad (2.12)$$

Con parámetros positivos l y ν y donde K_{ν} es una función de Bessel modificada. Se puede probar que cuando $\nu \to \infty$, se obtiene la función de covarianza SE. Para esta clase de funciones de covarianza, f(x) será k-veces diferenciable si y sólo si ν > k.

Este tipo de funciones se vuelven especialmente simples cuando ν es un semi-entero dado por ν=p+1/2, siendo p un entero no negativo. Cuando esto ocurre, las funciones se transforman en un producto de una exponencial y un polinomio de orden p. Entre estas funciones, se pueden destacar dos casos interesantes por su aplicación en el aprendizaje automático: ν=3/2 y ν=5/2.

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right) \qquad (2.13)$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) exp\left(-\frac{5r}{l}\right) \qquad (2.14)$$

Para valores de $\nu \geq 7/2$, es complicado distinguir entre diferentes valores de ν o incluso distinguirlos del caso $\nu \to \infty$, a partir de un conjunto finito y ruidoso de entrenamiento.

Por último, un caso especial de esta clase es el que se obtiene al considerar ν=1/2. Esta es la función de covarianza exponencial dada por:

$$k_{\nu=1/2}(r) = exp\left(-\frac{r}{l}\right) \qquad (2.15)$$

Esta función es continua en media cuadrada pero no diferenciable, lo cual implica que el modelo se vuelve más abrupto.

c). Función de Covarianza Cuadrática Racional (RQ)

La función cuadrática racional se puede considerar como una suma infinita de funciones de covarianza SE con diferentes longitudes de escala características. Está dada por:

$$\mathrm{k_{RQ}(r)} = \left(1 + \frac{\mathrm{r}^2}{2\alpha l^2}\right)^{-\alpha} \qquad (2.16)$$

Con α , $l > 0$. El límite de esta función cuando $\alpha \to \infty$ es la función de covarianza SE con longitud de escala característica l. A diferencia de la clase Matérn, el proceso correspondiente a la función RQ es infinitamente diferenciable en media cuadrática para cada α.

2.2.3. Selección del Modelo y de los Hiperparámetros

La elección de la forma funcional de la función de covarianza y de sus parámetros es fundamental para el entrenamiento de un modelo de GPR. Los parámetros ajustables de las funciones de covarianza son conocidos como hiperparámetros. Algunos de ellos son la longitud de escala (l), y la varianza del ruido (σn^2). La longitud de escala se puede interpretar como la distancia en el espacio de entrada que se debe recorrer para que el valor de la función cambie significativamente o para que las funciones dejen de estar correlacionadas. Una l corta implica funciones que varían más rápidamente, mientras que una l larga implica funciones más "suaves" o que varían lentamente. Por otro lado, la varianza del ruido (σn^2) explica la variabilidad en las observaciones que no puede ser atribuida a la función subyacente.

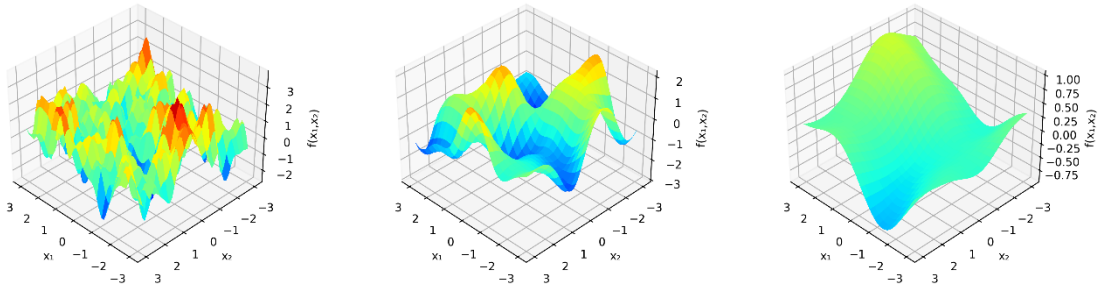

Fig 2.2. Funciones con dos variables de entrada generadas con una función de covarianza SE sin ruido. Las figuras muestran funciones de covarianza con diferentes hiperparámetros.

La elección adecuada de los hiperparámetros y de la función de covarianza es esencial para el buen rendimiento del modelo. La Fig. 2.2 muestra cómo los cambios en los hiperparámetros afectan al modelo desarrollado, resaltando la importancia de realizar una elección correcta de los mismos. Una forma de seleccionar modelos y ajustar los hiperparámetros es mediante el uso del Log-Marginal-Likelihood (LML), también conocido como 'evidencia' o 'verosimilitud' del modelo. Esta cantidad representa cuán probable es observar los datos Y, dados el conjunto de entrada X y los hiperparámetros del modelo. Se calcula integrando la verosimilitud de los datos multiplicada por la distribución a *priori* de los parámetros. Si se asume un ruido de observación Gaussiano, independiente e idénticamente distribuido, tomará la siguiente forma:

$$\log p(y|X,\theta) = -\frac{1}{2}y^T(K+\sigma_n^2 I)^{-1}y - \frac{1}{2}\log|K+\sigma_n^2 I| - \frac{n}{2}\log(2\pi) \qquad (2.17)$$

donde *y* es el vector de los valores objetivo en el conjunto de entrenamiento, *X* es la matriz de entradas del conjunto de entrenamiento, θ es el vector de hiperparámetros, *K* es la matriz de covarianza, calculada entre todos los puntos de entrada del entrenamiento, σ_n^2 es la varianza del ruido de observación, *I* es la matriz identidad y *n* es el número de puntos de entrenamiento u observaciones.

Una propiedad fundamental de la verosimilitud marginal es que incorpora automáticamente un equilibrio entre el ajuste del modelo a los datos y la complejidad del modelo. Este principio se conoce como la 'Navaja de Occam', que establece que los modelos deben ser lo suficientemente complejos para explicar los datos, pero no más de lo necesario, es decir, no deben ser ni muy simples ni demasiado complejos. Esto explica por qué la verosimilitud marginal no favorece simplemente a los modelos que se adecúan mejor a los datos de entrenamiento.

Un modelo demasiado simple no podrá describir adecuadamente los datos y tendrá una baja verosimilitud marginal. Por otro lado, un modelo demasiado complejo, o muy flexible, puede ajustarse a una gran variedad de posibles conjuntos de datos y al intentar adaptarse a cualquier dato de entrada, genera un sobreajuste. Esto hace que la probabilidad asignada a un conjunto de datos específicos sea menor, y por lo tanto que la verosimilitud marginal sea baja.

El primer término de la ecuación 2.17 es el que mide que tan bien el modelo explica las observaciones *y* (Data-Fit), mientras que el segundo miembro es el que penaliza la complejidad o flexibilidad del modelo. Por último, el término $\frac{n}{2}\log(2\pi)$ es una constante de normalización.

El LML proporciona una manera de comparar diferentes modelos GP, con distintas funciones de covarianza o hiperparámetros. Al maximizar el LML, se seleccionan los valores de los hiperparámetros que mejor "explican" los datos observados, equilibrando el ajuste y la complejidad y por lo tanto el modelo con el LML más alto se considera el más probable dadas las observaciones.

2.3. Kriging

2.3.1. Introducción

La mayor parte de las variables medioambientales, como las precipitaciones, se miden en puntos discretos entre los que existen grandes distancias. Sin embargo, en muchas aplicaciones es necesario contar con información continua en el espacio, es decir, obtener valores estimados

en ubicaciones donde no se dispone de mediciones. La geoestadística ofrece un conjunto de herramientas que permiten realizar este tipo de estimaciones de forma óptima, basadas en la correlación espacial entre las observaciones (Oliver y Webster, 2015).

Entre estas herramientas, Kriging es un método óptimo de predicción que proporciona estimaciones insesgadas y de varianza mínima. Se fundamenta en la Teoría de Variables Regionalizadas (TVR), según la cual las variables espaciales pueden modelarse como realizaciones de un proceso aleatorio que presenta autocorrelación espacial (Goovaerts, 2000; Sluiter, 2009).

Los procesos ambientales obedecen leyes físicas, por lo que en principio son deterministas (Oliver y Webster, 2015). Sin embargo, las múltiples interacciones entre procesos generan resultados tan complejos que, a escala de observación, se comportan como si fueran aleatorios. En este sentido, la geoestadística asume que la variación espacial de un atributo continuo, como las precipitaciones, puede considerarse el resultado de un proceso aleatorio espacialmente correlacionado.

La autocorrelación espacial describe la relación existente entre los valores de una variable en diferentes ubicaciones. Su cuantificación constituye el primer paso del análisis geoestadístico y se realiza mediante el variograma o la función de covarianza, que se pueden aproximar a partir de los datos muestrales (Oliver y Webster, 2015). Posteriormente, estos se ajustan con funciones matemáticas simples cuyas características (nugget, rango y sill) describen el comportamiento espacial de la variable. Los parámetros derivados del variograma son esenciales para el segundo paso: la predicción mediante Kriging.

2.3.2. Fundamentos del Kriging

El Kriging predice los valores de una variable en ubicaciones no muestreadas a partir de datos dispersos, basándose en un modelo estocástico de variación espacial continua.

Desde un punto de vista estadístico, el Kriging puede interpretarse como una técnica de regresión de mínimos cuadrados generalizados que incorpora la dependencia espacial entre observaciones vecinas mediante el variograma (Goovaerts, 2000).

En este marco, cada observación se considera una muestra particular de una variable aleatoria espacialmente dependiente. El método estima el valor en una ubicación no muestreada como la mejor combinación lineal e insesgada de las observaciones disponibles. El carácter "óptimo" del Kriging proviene de que, bajo los supuestos de estacionariedad y covarianza conocida, el estimador es lineal, insesgado y de varianza mínima.

No obstante, en muchos procesos meteorológicos o ambientales, la estacionariedad es una condición difícil de cumplir. En estos casos pueden adoptarse estrategias como limitar el tamaño o la forma del vecindario de búsqueda para mantener la validez local de los supuestos de Kriging (Sluiter, 2009).

En su forma más simple, el Kriging Ordinario (OK) no requiere información adicional más allá de las coordenadas espaciales y los valores medidos, siendo uno de los métodos más utilizados en ciencias ambientales por sus supuestos fácilmente cumplibles (Oliver y Webster, 2015).

El OK asume que la variable de interés es una variable aleatoria espacialmente dependiente, cuyo proceso subyacente es intrínsecamente estacionario. Esto implica que la media es constante y que la covarianza entre dos puntos depende únicamente de su separación (distancia y dirección), y no de su posición absoluta.

Matemáticamente, si la variable aleatoria Z se midió en N puntos $x_1, x_2, \dots, x_N$ obteniendo los valores $z(x_i)$, la predicción en un punto no muestreado $x_0$se calcula como una combinación lineal ponderada de las observaciones:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \qquad (2.18)$$

donde $\lambda_i$son los pesos asignados a cada observación. Para asegurar que la estimación sea insesgada, se impone la restricción:

$$\sum_{i=1}^{N} \lambda_i = 1 \qquad (2.19)$$

Los pesos λ_i se determinan a partir del modelo de variograma o covarianza, de forma que la varianza del error de estimación sea mínima. Cabe destacar que en la mayoría de las aplicaciones ambientales se trabaja con problemas bidimensionales.

2.3.3. Variograma

El variograma experimental (o semivariograma $\hat{\gamma}(h)$) es la representación gráfica de la semivarianza en función del vector de separación h entre los puntos. La semivarianza mide la disimilitud promedio entre los valores de pares de datos separados por cierta distancia. y se calcula como la mitad de la diferencia cuadrada promedio entre los valores observados de pares de datos separados por un vector h.

$$\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}\{z(x_i) - z(x_i + h)\}^2 \qquad (2.20)$$

Donde $z(x_i)$ y $z(x_i + h)$ son los valores observados de z en los puntos x_i y x_i+h. y n es el número de pares de puntos separados por h.

El variograma es la herramienta central de la geoestadística, crucial para evaluar si los datos están espacialmente correlacionados y en qué medida (Oliver y Webster, 2015). A partir del modelo de variograma ajustado a los datos experimentales, se pueden obtener predicciones óptimas mediante técnicas de *kriging*. Las tres características principales del variograma son:

1. Nugget: Se refiere a la discontinuidad o intersección positiva en la ordenada (h=0). Teóricamente, la semivarianza debería ser cero a una distancia cero. La varianza *nugget* surge del error de medición y de la variación espacial a distancias inferiores al intervalo de muestreo más corto (Oliver y Webster, 2015). En el contexto del kriging con deriva externa (KED), el efecto *nugget* se incluye para dar cuenta de las incertidumbres de medición y la variabilidad de corto alcance (Hiebl y Frei, 2017)

2. Sill: Es el límite superior que alcanza el variograma. Para los procesos estacionarios de segundo orden, el *sill* es el valor de la varianza *a priori* σ^2 del proceso.

3. Range: Es la distancia de retardo (lag) a la cual el variograma alcanza su *sill*. Esta distancia define el límite de la correlación espacial, de modo que los puntos separados por una distancia mayor al *range* se consideran espacialmente independientes.

Los modelos de variograma son fundamentales para la predicción. Al variograma experimental se le ajusta un modelo teórico. Si varios modelos ajustan bien entonces se elige el que presente un menor error cuadrado medio (MRE). Algunos de los más comunes son:

- Esférico: Caracterizado por un crecimiento lineal en el origen y alcanzando su sill a una distancia finita.
- Exponencial: Se aproxima asintóticamente al *sill*. Por lo que no tiene un rango finito. El rango efectivo suele definirse cuando se alcanza el 95% del *sill*.
- Gaussiano: Muestra un comportamiento parabólico cerca del origen, implicando mayor suavidad del fenómeno.

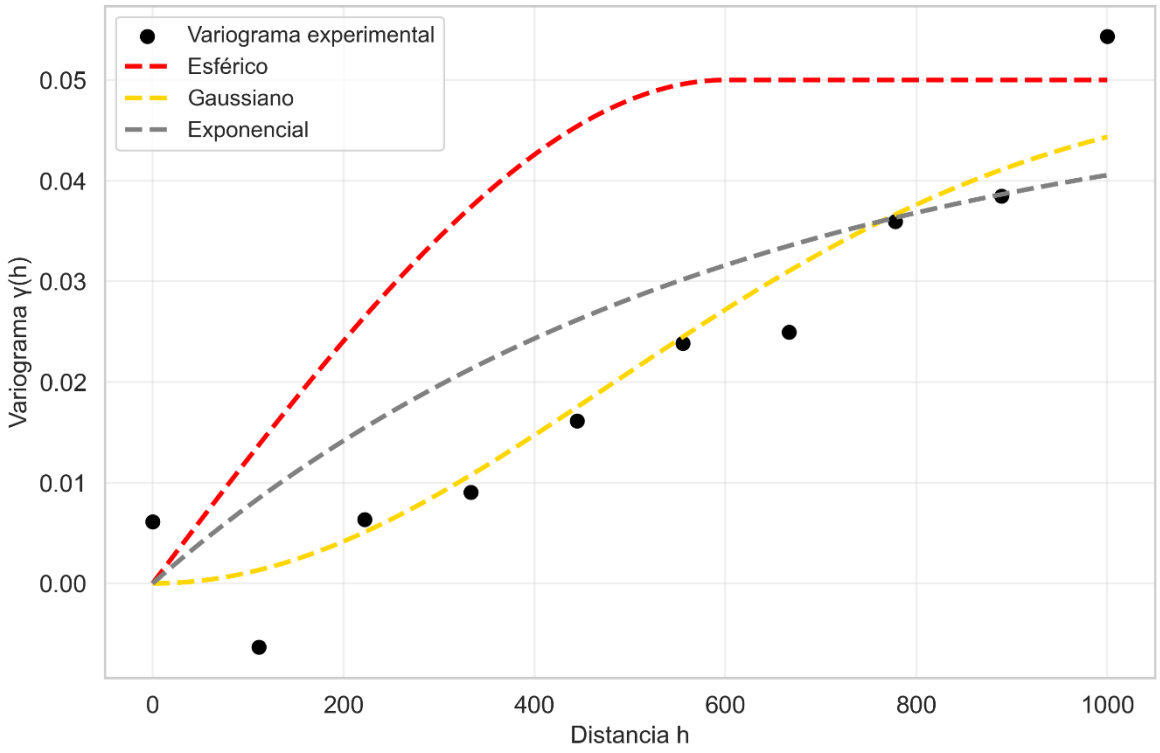

Fig 2.3. Ejemplo de variograma experimental y ajustes teóricos.

2.3.4. Estacionariedad, isotropía y anisotropía

El supuesto de estacionariedad intrínseca no siempre se cumple, por ejemplo, en presencia de una tendencia espacial pronunciada (denominada deriva*)*. En estos casos, la tendencia puede modelarse explícitamente mediante una componente determinística.

La estacionariedad de segundo orden se considera una relajación del criterio estricto de estacionariedad. Bajo este supuesto, cada punto del área de estudio tiene una distribución estadística potencial con la misma media y desviación estándar, y la covarianza entre puntos depende únicamente de la distancia y dirección que los separa (Sluiter, 2009).

La isotropía implica que la variación espacial es independiente de la dirección, mientras que la anisotropía se presenta cuando la correlación espacial varía según la dirección (Sluiter, 2009). En este último caso, la estructura del variograma (nugget y sill) puede mantenerse constante, pero el rango varía direccionalmente, indicando que la correlación se extiende más lejos en algunas direcciones que en otras. Esta situación, conocida como *anisotropía geométrica*, puede corregirse mediante una transformación lineal de las coordenadas (estiramiento del espacio) (Oliver y Webster, 2015).

También puede ocurrir una anisotropía zonal, cuando existen zonas preferenciales con medias diferentes que provocan variaciones en la varianza y en el sill según la dirección. Para

detectar anisotropías se recomienda calcular el variograma empírico al menos en cuatro direcciones distintas y analizar sus diferencias (Oliver y Webster, 2015; Goovaerts, 1997).

En términos generales, el variograma es una función que describe la variabilidad espacial en función de la distancia y, en casos anisotrópicos, también de la dirección. La comparación de semivariogramas experimentales en diferentes direcciones permite identificar la presencia y tipo de anisotropía, así como determinar la dirección de mayor continuidad espacial.

2.3.5. Kriging con deriva externa (KED)

Cuando existe una tendencia o deriva espacial explicable a partir de variables auxiliares es posible incorporar esta información en el modelo mediante el Kriging con deriva externa (KED).

El KED es un método geoestadístico multivariado que combina una componente determinística (deriva) y una componente estocástica (residual). La componente determinística describe la variable objetivo como una combinación lineal de campos espaciales conocidos (deriva externa), mientras que la estocástica modela los residuales como un campo aleatorio gaussiano con una estructura de covarianza espacial (Goovaerts, 2000; Hiebl y Frei, 2017).

Entonces, KED produce la interpolación como la mejor combinación lineal insesgada de las observaciones de muestra, asumiendo que el campo de interés en una muestra de un proceso gaussiano de segundo orden (Masson y Frei, 2014). Formalmente, el modelo puede expresarse como:

$$Y(s) = \mu(s) + Z(s), \text{ con } \mu(s) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k(s) \qquad (2.21)$$

donde $\mu(s)$ representa a la componente determinística (la tendencia o deriva externa) y está dada por la combinación lineal de K covariables $x_k(s)$, con β_k los coeficientes de tendencia. La componente $Z(s)$ describe a la parte estocástica del modelo, es decir, representa la componente aleatoria con media nula y covarianza estacionaria de segundo orden (Masson y Frei, 2014).

En términos prácticos, KED utiliza la información auxiliar para estimar la media local del atributo principal y luego aplica un Kriging simple sobre los residuales (Goovaerts, 2000).

Capítulo 3: Armado de la base de datos

3.1. Observaciones

Para el desarrollo de este trabajo fue fundamental la recolección de registros de precipitación total diaria del NOA, abarcando un período temporal lo suficientemente amplio. En esta sección se describen las fuentes utilizadas para la obtención de datos de precipitaciones, el período temporal abarcado, y los procesos de filtrado y organización de los datos previos al entrenamiento de los modelos.

Los datos utilizados provienen de estaciones meteorológicas pertenecientes al Sistema Nacional de Información Hídrica (SNIH, https://www.argentina.gob.ar/obras-publicas/hidricas/base-de-datos-hidrologica-integrada), al Instituto Nacional de Tecnología Agropecuaria (INTA, https://siga.inta.gob.ar/) y al Servicio Meteorológico Nacional (SMN, https://www.smn.gob.ar/). En particular, se emplearon series diarias de precipitaciones, que posteriormente fueron filtradas y procesadas. Además, se incorporaron datos geográficos de cada estación (latitud, longitud y altitud) con el fin de realizar la interpolación espacial.

Debido a que las coordenadas geográficas son esenciales para la interpolación espacial, se descartaron 16 estaciones del SNIH cuyas coordenadas geográficas y/o altitud no estaban accesibles públicamente al momento de realizar esta tesis. La obtención de esta información no es un proceso sencillo puesto que, aunque existen plataformas oficiales, como el portal del SNIH y el sitio *datos.gob.ar*, estas no ofrecen un listado descargable con las ubicaciones y alturas correspondientes a las estaciones. La falta de acceso a las coordenadas y las dificultades en el proceso de descarga representan una limitación significativa para el uso de estos datos por parte del público en general.

Por otra parte, la descarga de los datos también presenta dificultades, ya que debe realizarse de forma manual, accediendo a cada estación de manera individual. Aunque el portal del SNIH permite realizar un filtrado de estaciones según las variables medidas, esta opción no siempre funciona correctamente, puesto que muchas estaciones que pasaban el filtro no contenían registros meteorológicos disponibles en ningún período de tiempo.

Las estaciones seleccionadas para el análisis fueron aquellas cuyas coordenadas geográficas y altitud estaban disponibles, estas se listan en la Tabla 3.1.

Nombre	Latitud (°)	Longitud (°)	Altura(m)	Nombre	Latitud (°)	Longitud (°)	Altura (m)
Caimancito	-23.70	-64.53	356	**Salta Aero**	-24.80	-65.30	1227
Potrero de las Tablas	-26.85	-65.42	999	**Catamarca Aero**	-28.60	-65.80	470
Casa de Piedra	-27.28	-65.91	1200	**Santiago Del Estero Aero**	-27.80	-64.30	201
Piedra Grande	-27.3	-65.80	500	**Tartagal Aero**	-22.60	-63.80	450
Yampa II	-27.18	-65.84	1500	**Oran Aero**	-23.20	-64.30	360
El Nogalito	-26.78	-65.47	1100	**Jujuy Aero**	-24.40	-65.10	910
Las Mesadas	-27.20	-65.93	1850	**La Rioja**	-29.34	-66.81	330
Aguas Blancas	-22.72	-64.35	407	**Chamical**	-30.36	-66.31	470
Balapuca	-22.48	-64.45	581	**Los Sosa**	-27.10	-65.60	640
Cuatro Cedros	-22.82	-64.52	484	**Potrero del Clavillo**	-27.40	-66.10	1380
Astillero	-22.37	-64.12	500	**La Paz**	-22.40	-62.50	250
San José	-22.87	-64.70	877	**Roque Sáenz Peña**	-26.87	-60.45	90
Iruya	-22.80	-65.21	2741	**Las Breas**	-27.10	-61.10	104
Las Higueras	-22.75	-65.10	1972	**Tucumán Aero**	-26.83	-65.10	450
San Isidro	-22.76	-65.24	2956	**El Colmenar**	-26.80	-65.20	481
Nazareno	-22.51	-65.10	3101	**Famaillá**	-27.05	-65.42	370
Paltorco	-22.41	-65.09	3720	**Ceres**	-29.88	-61.95	89
Tuc-Tuca	-22.40	-65.27	3950	**Corrientes**	-27.45	-58.77	61
Poscaya	-22.45	-65.08	3238	**Formosa**	-26.20	-58.23	63
Trigo Huaico	-22.37	-65.04	3300	**Iguazú**	-25.73	-54.47	256

El Molino	-22.59	-65.15	2600	**Las Lomitas**	-24.70	-60.58	134
El Pabellón	-22.55	-65.34	4265	**Posadas**	-27.36	-55.96	81
Pozo Sarmiento	-23.14	-64.19	326	**Reconquista**	-29.18	-59.70	49
La Quiaca	-22.10	-65.60	3450	**Resistencia**	-27.45	-59.05	54

Tabla 3.1. Nombre, latitud, longitud y altura de las estaciones utilizadas en este trabajo. Notar que "Aero" se refiere a aeropuertos.

Se utilizaron 38 estaciones correspondientes a la región del NOA, y se incluyeron además 10 estaciones adicionales pertenecientes a la región del NEA. Esta inclusión tuvo como objetivo mejorar la interpolación espacial en el área cercana a Santiago del Estero, donde solo se cuenta de una única estación meteorológica en la base de datos armada. Se consideró que estas estaciones podrían mejorar la interpolación ya que el relieve y las altitudes de la región del NEA son similares a las de Santiago del Estero, como se observa en la Fig 3.1.

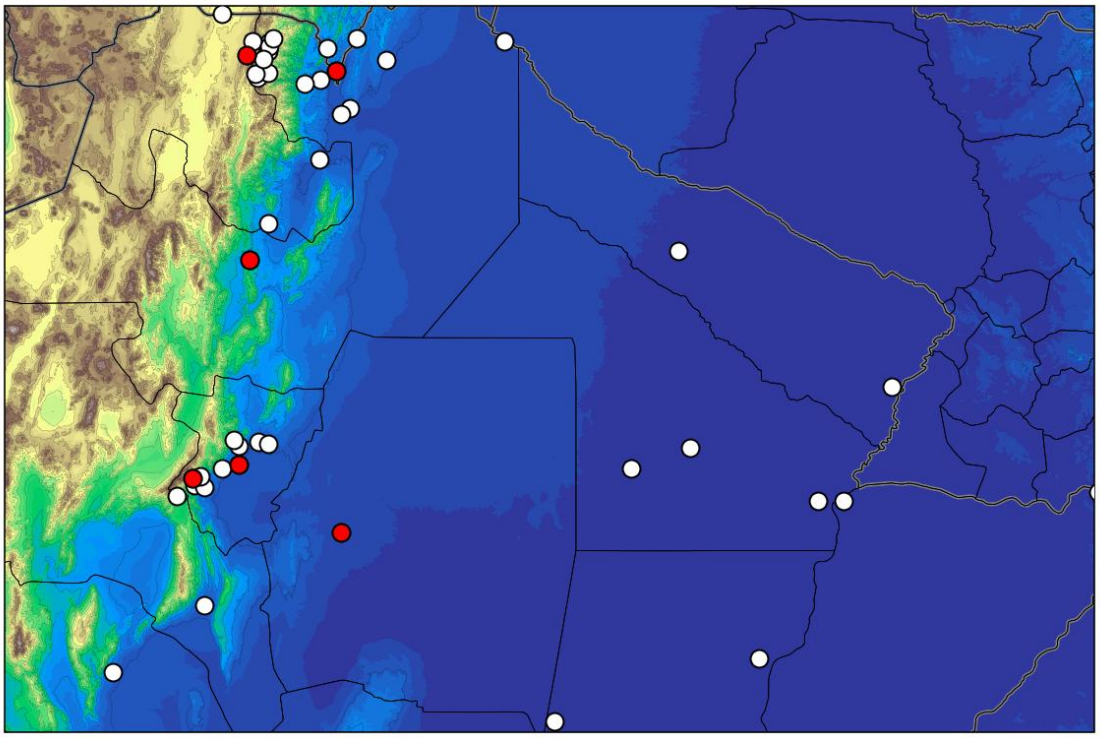

Fig. 3.1. Ubicación de las estaciones meteorológicas utilizadas para el entrenamiento de los modelos (puntos blancos) y para su testeo (puntos rojos), sobre mapa de relieve (color).

La Figura 3.1 muestra la distribución espacial de las estaciones utilizadas. Se observa que no se encuentran distribuidas de manera uniforme a lo largo de la región de análisis, sino que se concentran principalmente en dos núcleos: uno ubicado en el norte salteño, en el límite con Jujuy, y otro en la zona central de Tucumán.

3.2. ERA5-LAND

Si bien las estaciones meteorológicas constituyen una fuente de medición directa, no siempre se dispone de una cobertura espacial y temporal adecuada. En estos casos, es común recurrir a bases de datos de reanálisis, las cuales combinan observaciones con modelos físicos para generar estimaciones continuas en el tiempo y el espacio. En este trabajo, se utilizó la base ERA5-LAND (disponible en *Copermicus Data Store*: https://cds.climate.copernicus.eu/) (Muñoz Sabater, 2019), la cual se empleó con el objetivo de compararla con los resultados obtenidos. Esta comparación permite mostrar el aporte y valor de los modelos entrenados de manera local frente a una alternativa de fácil acceso y ampliamente utilizada como ERA5-LAND.

ERA5-LAND es un conjunto de datos de reanálisis desarrollado por el European Centre for Medium-Range Weather Forecasts (ECMWF) que proporciona una reconstrucción coherente y continua de las condiciones en superficie terrestre durante las últimas décadas (Muñoz-Sabater et al., 2021). Es una base de datos de última generación diseñada específicamente para aplicaciones terrestres. ERA5-LAND no asimila observaciones directamente, sino que este producto se genera mediante la ejecución del modelo de superficie terrestre de ERA5, utilizando como entrada variables atmosféricas del propio reanálisis (como temperatura, humedad, presión, radiación y viento) que actúan como forzantes y determinan los intercambios de energía y humedad entre la atmósfera y el suelo. A partir de estas condiciones, el modelo simula variables de superficie tales como la temperatura del suelo, la humedad, la evapotranspiración y la precipitación. ERA5-LAND presenta una resolución espacial más alta (~9 km) que ERA5, lo que mejora la representación de los procesos hidrológicos y la distribución espacial de la lluvia, especialmente en regiones con topografía compleja o elevada variabilidad climática.

Capítulo 4: Metodología

4.1. Procesado y Filtrado de Datos

En la construcción de los modelos fue necesario dividir las estaciones en dos subconjuntos: uno de entrenamiento y otro de testeo. Debido a la limitada cantidad de estaciones disponibles en la región del NOA, se seleccionaron seis estaciones para el testeo (puntos rojos en Figura 3.1). Esta elección se realizó con el objetivo de representar de manera adecuada las diferentes regiones que conforman al NOA y analizar la influencia del relieve y de la cercanía a estaciones meteorológicas en las predicciones realizadas con diferentes modelos.

De esta manera, se seleccionaron dos estaciones del núcleo central de Tucumán: la de menor altitud (Famaillá) y la de mayor altitud (Las Mesadas). Con el mismo criterio se eligieron dos estaciones del núcleo ubicado en el norte de Salta: la estación Aguas Blancas (menor altitud) y la estación El Pabellón (mayor altitud). Además, se incluyó la estación SALTA AERO, ubicada aproximadamente a mitad de camino entre ambos cúmulos (Tucumán y norte de Salta), y la estación SANTIAGO DEL ESTERO AERO, que constituye la única estación disponible en esa provincia. Esta selección permitió evaluar el desempeño de los modelos en distintas áreas y relieves, y se mantuvo constante para todos los modelos y pruebas comparativas.

Una vez armada la base de datos con los registros de precipitaciones correspondientes a las estaciones cuyas coordenadas geográficas estaban disponibles, se procedió al filtrado de los datos. Debido a que la mayoría de las estaciones tenían registros principalmente entre los años 1990 y 2024, se decidió limitar el análisis a ese período.

Como el objetivo principal del trabajo es realizar una interpolación espacial, se consideró necesario utilizar únicamente aquellos años en los que todas las estaciones disponibles tuvieran una cantidad significativa de datos. Para ello, se contaron los días en los que cualquier estación no registraba datos, y luego se evaluó la cantidad de datos anuales faltantes. Se decidió conservar únicamente los años en los que todas las estaciones presentaran al menos el 90 % de los datos anuales, es decir, se permitió un máximo de 35 días faltantes en cada año. Por lo tanto, cualquier año en el que al menos una estación tuviera más de 35 datos faltantes fue descartado. La Figura 4.1 muestra la cantidad de datos diarios faltantes por año en el período 1990–2024. A partir de este análisis, se excluyeron los años anteriores a 1993, posteriores a 2019 y los años 1995, 1997, 2000, 2004 y 2007, debido a que no cumplían con el umbral definido.

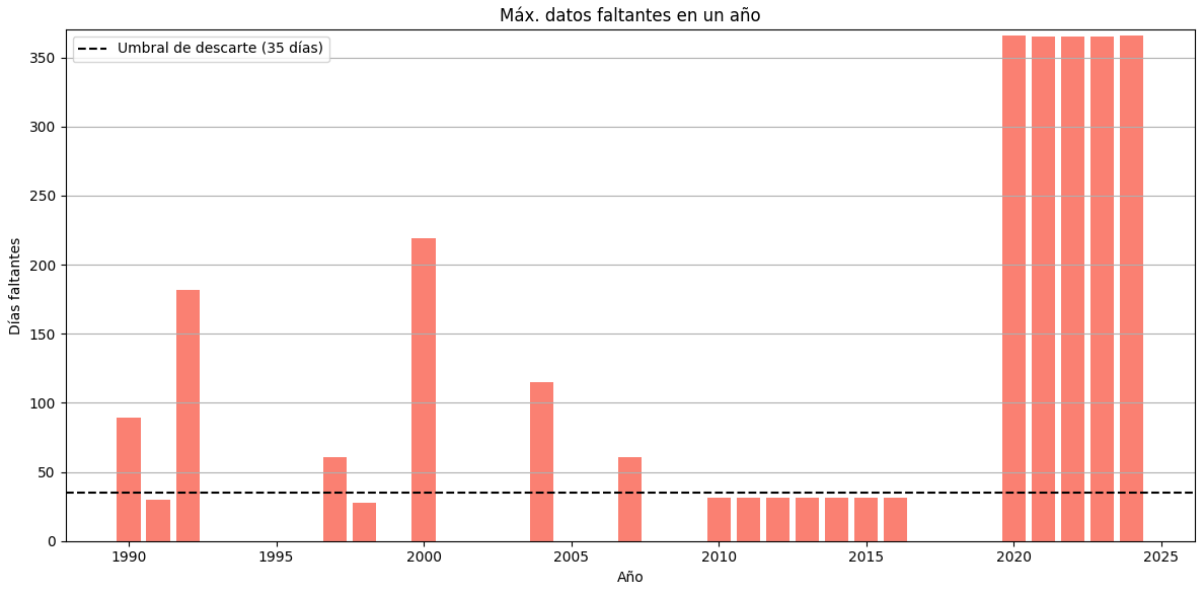

Fig. 4.1. Máximo de datos diarios faltantes por año entre todas las estaciones.

Una vez seleccionados los años, se realizó la suma de los datos diarios, para obtener el acumulado anual de las precipitaciones para cada estación. Además, antes de entrenar los modelos de GPR y NN, se aplicó un escalado estándar a los datos, eliminando la media y escalando a la varianza unitaria. Este es un paso necesario en la aplicación de varios métodos de aprendizaje automático, especialmente para los modelos de redes neuronales, utilizado en este trabajo. Cabe destacar que el escalado de los datos no es necesario para el modelo GPR. Sin embargo, para facilitar la comparación entre los modelos de aprendizaje automático, se decidió trabajar de la misma manera con los datos en ambos casos, y por lo tanto realizó el correspondiente escalado al desarrollar los modelos de ML.

4.2. Métodos de Interpolación

Los métodos evaluados en este trabajo son: redes neuronales (NN), Gaussian Process Regression (GPR), Kriging, y la base de reanálisis ERA5-LAND.

4.2.1. Kriging

Para los modelos de Kriging se consideraron dos variantes principales: Kriging Ordinario (OK) y Kriging con Deriva Externa (KED), ambos desarrollados con la librería gstools en Python. En el caso del KED, se utilizó la altitud como variable de deriva, considerando que las diferencias topográficas de la región influyen en gran medida en la distribución espacial de las precipitaciones (Ferrero y Villalba, 2019).

Para ambos modelos se analizaron variantes isotrópicas y anisotrópicas. En los modelos isotrópicos se asumió que la dependencia espacial de las precipitaciones es igual en todas las direcciones, mientras que en los anisotrópicos se modeló la dependencia direccional calculando la razón y el ángulo de anisotropía.

La incorporación de la altitud y el modelado de la anisotropía permitió determinar el impacto de estas variables, lo cual permite identificar en qué regiones del NOA estos factores influyen en la interpolación. Además, permite analizar si la irregularidad en la distribución espacial de las estaciones afecta la capacidad del modelo para representar adecuadamente la anisotropía presente en los patrones espaciales de precipitación. Una mala distribución puede generar que el número de pares de puntos en determinadas direcciones sea insuficiente, lo que puede dificultar la estimación de los patrones direccionales de las precipitaciones, afectando la confiabilidad del ajuste y la calidad de la interpolación.

Para modelar la anisotropía se estimó el variograma experimental en nueve direcciones distribuidas entre 0 y 180°. A cada variograma direccional se le ajustó un modelo gaussiano y se determinó la longitud de escala o rango característicos de cada dirección. A partir de los rangos máximos y mínimos se calculó el ángulo de anisotropía, correspondiente a la dirección de mayor correlación espacial, y la razón de anisotropía, definida como la relación entre los rangos mínimo y máximo. Estos parámetros fueron luego incorporados en los modelos de Kriging anisotrópico.

En los cuatro modelos de Kriging desarrollados, la variable a predecir fue la precipitación anual acumulada en cada estación y se utilizaron como variables de entrada las coordenadas y altitudes de las estaciones. Este tipo de herramienta no permite realizar interpolaciones temporales, por lo que la estimación del variograma se realizó anualmente, de manera independiente para cada año del período analizado. En cada año se ajustó un modelo gaussiano al variograma experimental, se entrenó el modelo con las estaciones de entrenamiento y se realizaron predicciones en las seis estaciones de testeo. Al ajustar el variograma teórico al experimental, la librería Gstools realiza una optimización minimizando la diferencia entre el variograma experimental y teórico, aunque se limitó el número máximo de evaluaciones para asegurar la convergencia del modelo.

4.2.2. Procesos Gaussianos de Regresión (GPR)

El método de GPR se aplicó utilizando la librería scikit-learn de Python. Se utilizaron como variables de entrada la latitud, longitud, altitud y año, en tanto, como variable de salida se utilizó la precipitación acumulada anual.

Durante el desarrollo del modelo, se probaron diferentes funciones de covarianza o kernels empleados usualmente al trabajar con GPR: Función de Base Radial o Exponencial Cuadrada (RBF), Matern, Cuadrática Racional (RQ), y combinaciones de ellos, todos acompañados de un modelado de ruido (White Kernel).

Se observó que en los modelos con kernel RQ, el parámetro α tendía sistemáticamente al límite superior definido, lo que indicaba que el modelo tenía un comportamiento similar al de un kernel RBF. Además, al combinar kernels RBF y Matern, el nivel de ruido tendía a valores mínimos, lo que sugería un sobreajuste del modelo (overfitting).

Por este motivo, se decidió analizar separadamente los kernels RBF y Matern, y seleccionar el modelo final en función del log-marginal-likelihood (LML). El kernel Matern ($\nu = 1.5$) presentó el menor LML y un ruido realista (no había indicios de sobreajuste por parte del modelo), por lo que se lo eligió configuración final.

4.2.3. Redes Neuronales (NN)

Finalmente, se entrenó un modelo de Redes Neuronales utilizando la librería Keras (TensorFlow). Al igual que en el modelo de GPR este modelo consideró como variable de salida la precipitación anual acumulada, y utilizó como variables de entrada la latitud, longitud, altitud y año. Sin embargo, también se incorporaron como variables predictoras el promedio regional de las lluvias anuales en las estaciones y la desviación estándar para cada año. Esto se realizó con el objetivo de proporcionar a la red neuronal información de la variabilidad de las lluvias, que no podrá encontrar únicamente con las coordenadas geográficas.

La red desarrollada constó de tres capas densas con 256, 32 y 4 neuronas respectivamente, todas con función de activación ReLU. Además, se aplicó regularización L2 en la primera capa seguida por una normalización por lotes (Batch Normalization) para evitar sobreajuste y estabilizar y acelerar el entrenamiento.

El modelo además utilizó el optimizador Adam (tasa de aprendizaje inicial 0.001) y se entrenó utilizando la función de pérdida “mean absolute error” (MAE). Además, para evitar el seguir entrenando cuando el modelo llegó al mejor rendimiento, se utilizaron las técnicas de EarlyStopping para detener el entrenamiento cuando no se observa una mejora en la pérdida, y ReduceLROnPlateau para disminuir la tasa de aprendizaje cuando la pérdida se estabiliza.

El entrenamiento se realizó durante un máximo de 250 épocas y con un tamaño de lote (batch) de 128 muestras. Los hiperparámetros (número de neuronas, tasa de aprendizaje, cantidad de capas) se determinaron mediante prueba y error, seleccionando la configuración que dio mejores resultados.

4.3. Medidas Estadísticas

La evaluación del rendimiento de los modelos se realizó utilizando las siguientes métricas estadísticas: el coeficiente de correlación de Pearson (R), el error absoluto medio (MAE) y el error relativo medio (MRE) calculados para cada estación de testeo teniendo en cuenta todos los años analizados.

El coeficiente de correlación de Pearson (R) se empleó para cuantificar la relación lineal entre los valores observados y los valores estimados por los modelos. Esta métrica permite evaluar en qué medida las predicciones reproducen la variabilidad de los datos reales, siendo valores cercanos a 1 indicativos de una alta concordancia entre la variabilidad de ambas series. El coeficiente de Pearson entre dos variables x, y está dado por:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (4.1)$$

donde $\bar{x}$ y $\bar{y}$ son los promedios de las variables.

El error absoluto medio (MAE) se utilizó para medir la magnitud promedio de los errores cometidos por los modelos sin considerar la dirección de las desviaciones. Esta métrica ofrece una estimación directa de la discrepancia promedio entre los valores predichos y los observados, facilitando la interpretación del desempeño general en unidades originales de la variable analizada.

$$MAE = \frac{\sum_{i=1}^{n}|y - \hat{y}|}{n} \qquad (4.2)$$

donde y es el valor real y $\hat{y}$ es el predicho por el modelo.

El error relativo medio (MRE) se empleó para evaluar el tamaño del error en relación con la magnitud de los valores observados. Además, al tener signo, permite analizar si el modelo tiende a sobreestimar o subestimar. Esta métrica permite comparar el rendimiento entre diferentes estaciones o conjuntos de datos al expresar los errores en términos relativos, lo que aporta una perspectiva proporcional de la precisión de las estimaciones.

$$MRE = \frac{\sum_{i=1}^{n}[(\hat{y} - y)/y]}{n} \qquad (4.3)$$

donde y es el valor real y $\hat{y}$ es el predicho por el modelo.

Capítulo 5: Resultados

En esta sección se presentan los resultados obtenidos a partir del entrenamiento y evaluación de los diferentes modelos utilizados para la interpolación espacial de precipitaciones anuales. El rendimiento de cada modelo se evaluó mediante las métricas propuestas, las cuales consisten en la comparación entre las precipitaciones medidas y las predichas. Además, se incluye una comparación con la base de reanálisis ERA5-LAND.

5.1. Análisis por distribución geográfica

Con el objetivo de evaluar cómo varía el rendimiento de los modelos en función de la ubicación geográfica se agruparon los resultados según 4 subregiones: Norte de Salta, Centro de Tucumán, Centro de Salta y Santiago del Estero. Esto permite visualizar cuáles son los modelos más adecuados para cada subregión.

5.1.1. Zona Norte de Salta

En esta región se seleccionaron dos estaciones meteorológicas: la estación Aguas Blancas, la cual presenta la menor altitud de la zona, y la estación El Pabellón, de mayor altitud. Las métricas obtenidas para estas estaciones con los siete métodos utilizados se resumen en las tablas 5.1 y 5.2, y se pueden observar gráficamente en las figuras 5.1 y 5.2, en las cuales están marcados en verde los métodos con mejores resultados y en rojo el método que obtiene el peor resultado para cada una de las métricas.

a) Estación El Pabellón (Zona Norte de Salta - Altura: 4265 m)

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.59	0.54	0.76	0.21	0.41	0.36	0.32
MAE	81.4	124.7	483.5	547.3	102.2	312.4	340.4
MRE	0.08	-0.17	1.34	1.57	0.28	0.93	-0.20

Tabla 5.1. Métricas para la estación El Pabellón.

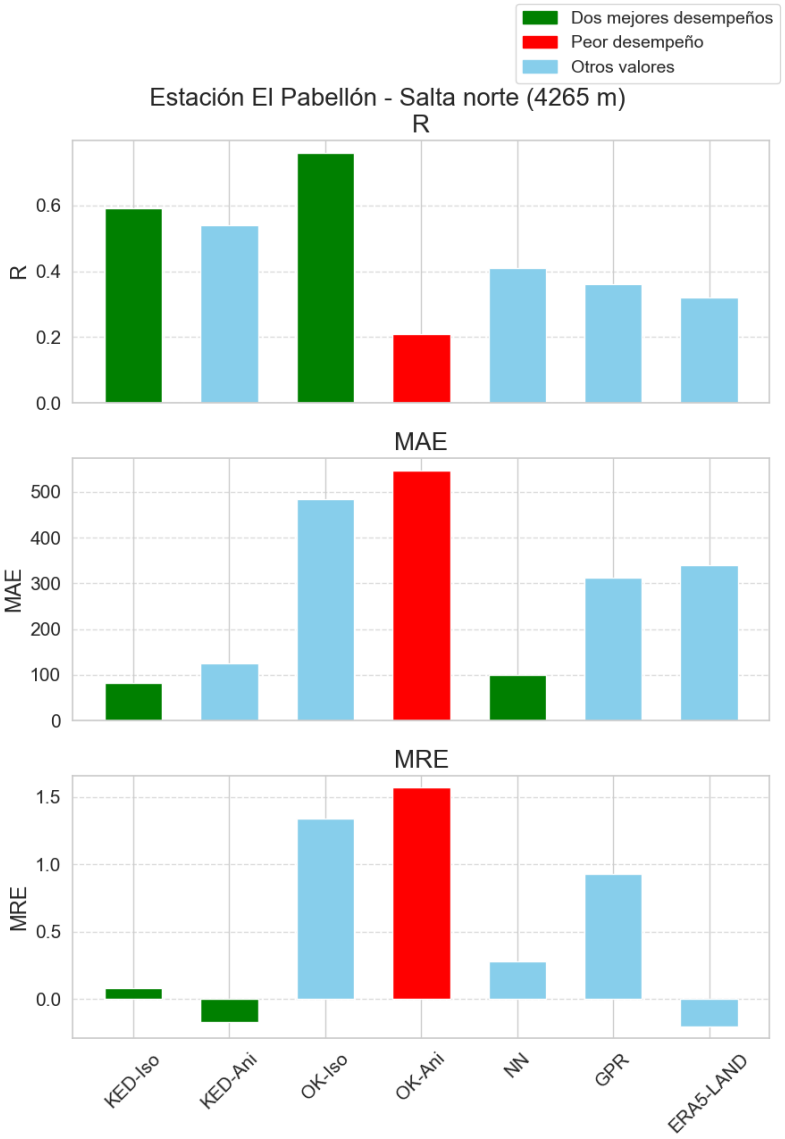

Fig 5.1. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación El Pabellón.

b) Estación Aguas Blancas (Zona norte de Salta - Altura: 407 m)

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.69	0.34	0.73	0.23	0.81	0.77	0.35
MAE	269.1	253.0	476.1	339.3	164.1	165.8	1198.2
MRE	-0.16	-0.09	-0.33	-0.21	-0.06	-0.01	3.36

Tabla 5.2. Métricas para la estación Aguas Blancas.

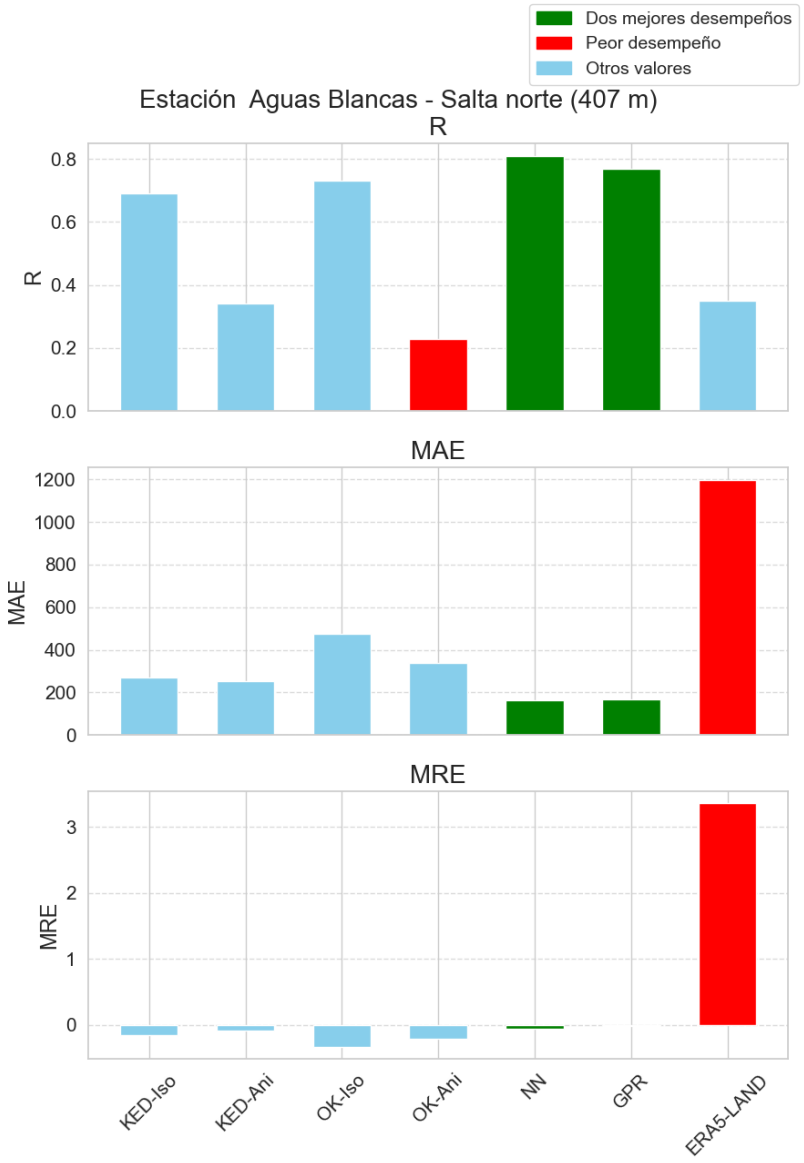

Fig 5.2. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación Aguas Blancas.

Se observa que, en esta región, los resultados de la interpolación mediante KED son muy buenos, en cuanto se obtiene un buen balance entre R alto y MAE/MRE bajos. En el caso de la estación El Pabellón (Tabla 5.1 y Fig. 5.1), el menor MAE (81 mm) se obtiene con el KED isotrópico, mientras que otros métodos muestran valores elevados (> 300 mm). Por otro lado, se debe destacar que también se logra un MAE bajo (102 mm) con el modelo de red neuronal (NN), cuyo valor de R (0,41) es el más alto entre los métodos de aprendizaje automático. Sin embargo, los mejores valores de R (> 0,5) se obtienen para los métodos basados en Kriging, indicando su superioridad para representar la variabilidad temporal de las precipitaciones en la estación El Pabellón.

En cambio, para la estación Aguas Blancas (Tabla 5.2 y Fig. 5.2) los mejores resultados se obtienen con NN, método con el cual se obtiene tanto el valor más alto de R (0.81) como el MAE más bajo (164 mm). Por otro lado, otro resultado a destacar es el que se obtiene con GPR (R~0.77, MAE~166 mm). Los resultados de NN y GPR son superiores a los demás, lo que indica que son capaces de capturar correctamente la variabilidad temporal en esta zona (alto R) y generar una buena predicción del valor observado (bajo MAE).

En relación con las distintas variantes de Kriging analizadas, se observa que la exclusión de la altitud como variable de deriva condujo a valores de MAE y MRE significativamente más altos, en particular en la estación ubicada a mayor altitud (Tabla 5.1). Este resultado destaca la relevancia de incorporar el relieve en el modelo, especialmente en regiones con marcadas variaciones altitudinales. En contraste, la consideración de la anisotropía no generó diferencias sustanciales en los valores de MAE en ninguna de las dos estaciones. En El Pabellón, el modelo anisotrópico presentó un MAE levemente superior, mientras que en Aguas Blancas fue ligeramente inferior.

Finalmente, se puede observar que los resultados obtenidos con los modelos estadísticos propuestos son mejores que los del ERA5-LAND. Las métricas obtenidas con el ERA5-LAND muestran que la base de reanálisis no es capaz de capturar correctamente la variabilidad temporal (bajo R) ni de generar buenas estimaciones del valor observado (altos MAE y MRE). Con esta base se observa un peor rendimiento en la estación Aguas Blancas, en la cual el MRE supera el 300%.

5.1.2. Zona Central de Tucumán

Para el análisis de esta región se eligieron dos estaciones: Famaillá y Las Mesadas, de menor y mayor altitud de la zona respectivamente. Las métricas obtenidas con los diferentes métodos para ambas estaciones se encuentran en las tablas 5.3 y 5.4 y se pueden observar gráficamente en las figuras 5.3 y 5.4.

a) Estación Famaillá (Zona Central de Tucumán - Altura: 370 m)

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.81	0.66	0.76	0.67	0.79	0.82	0.57
MAE	133.6	168.8	215.1	235.9	145.3	138.0	287.9
MRE	-0.009	-0.02	-0.11	-0.14	-0.03	0.06	0.11

Tabla 5.3. Métricas para la estación Famaillá.

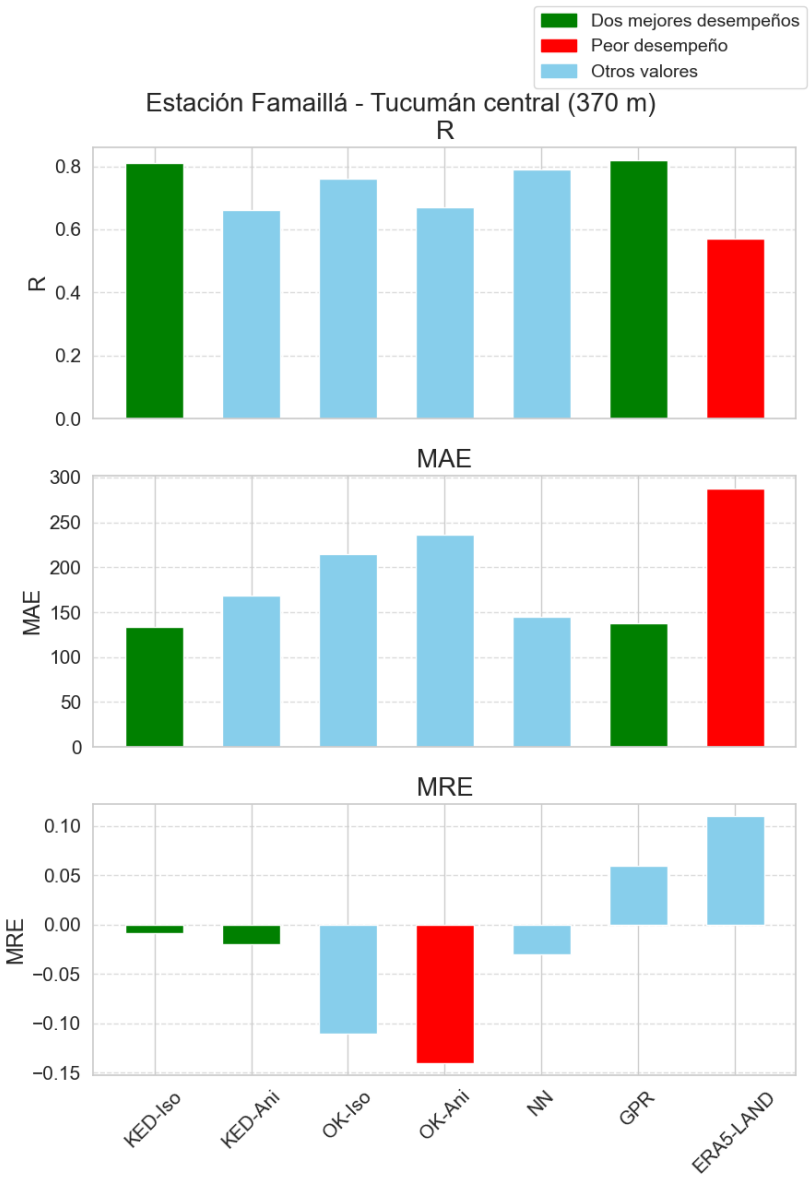

Fig 5.3. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación Famaillá.

b) Estación Las Mesadas (Zona Central de Tucumán - Altura: 1850 m)

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.66	-0.16	0.71	-0.25	0.72	0.42	0.02
MAE	952.0	938.1	797.5	890.2	276.0	432.0	1111.3
MRE	-0.48	-0.46	-0.40	-0.41	-0.10	-0.19	0.90

Tabla 5.4. Métricas para la estación Las Mesadas.

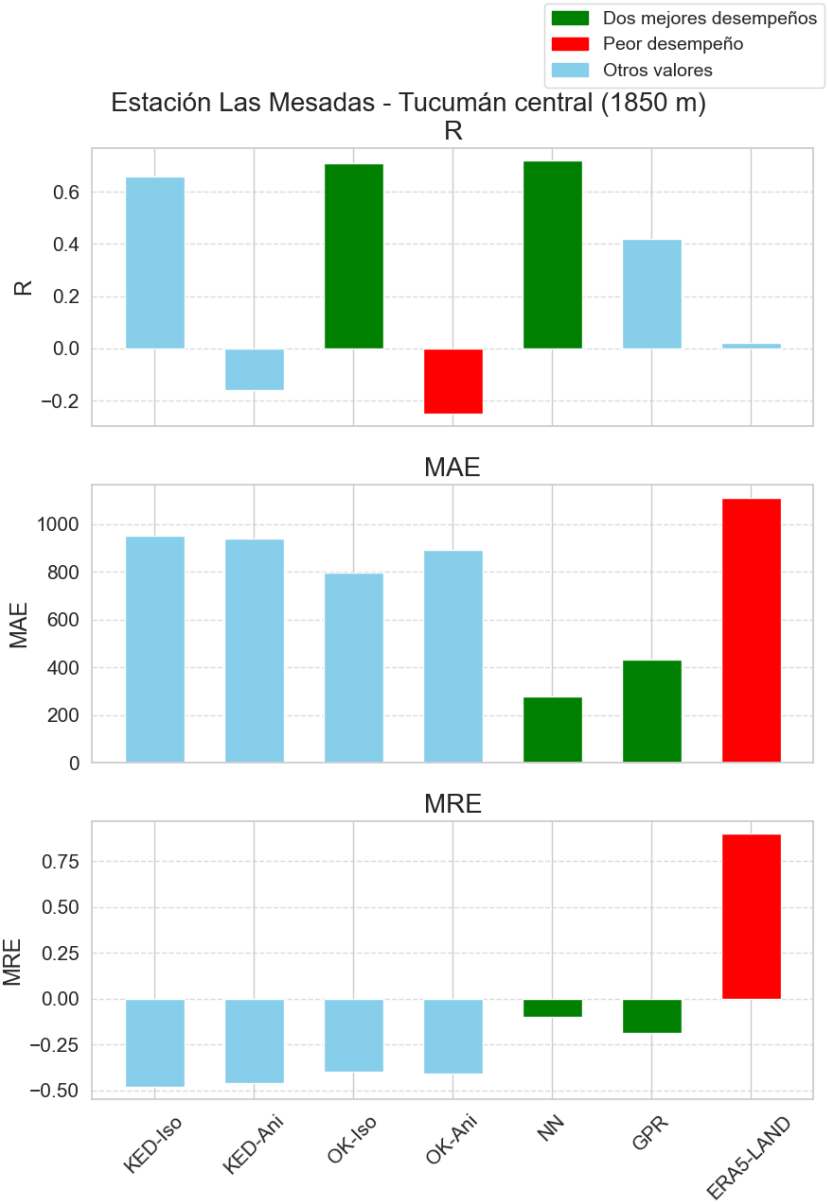

Fig 5.4. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación Las Mesadas.

Se observa que, en esta región, un muy buen desempeño del modelo NN para ambas estaciones. Se observa un mejor rendimiento del modelo en la estación Famaillá, en la cual R (0.79) es mayor y el MAE (145 mm) es menor.

En cuanto a la estación Famaillá (Tabla 5.3 y Fig. 5.3), los resultados obtenidos con KED isotrópico, GPR y NN son muy similares. Para estos tres métodos el valor de R es alto (R>0.79) y los MAE/MRE bajos (MAE<145mm, MRE<0.06), por lo que pueden detectar la variabilidad temporal en las precipitaciones y realizar estimaciones adecuadas del valor. El KED isotrópico presenta resultados ligeramente mejores, siendo el método con menor MRE y MAE.

En cuanto a la estación Las Mesadas (Tabla 5.4 y Fig. 5.4), los mejores resultados en general se obtienen para el modelo NN (R>0,7, MAE ~ 275 mm, MRE ~ -0,1). Sin embargo, el R para todos los métodos es menor que para el caso de Famaillá, lo que indica una mayor dificultad para predecir la variabilidad temporal en la zona elevada de Tucumán con los metodos propuestos. Particularmente, los modelos de Kriging presentan valores de MAE y MRE elevados (MAE>798 mm, |MRE|>0.40), con lo que no son confiables para realizar estimaciones precisas. Además, se observa que para todas las variantes de Kriging, los valores de los errores relativos son negativos, por lo que estos modelos tienden a subestimar las precipitaciones en la región.

Por otro lado, en ambas estaciones se observa que los modelos de Kriging anisotrópico presentan un desempeño inferior al de los modelos isotrópicos, lo que sugiere que no es óptimo modelar la anisotropía en la zona. En cuanto a la inclusión de la altura como variable de deriva en esta región, se observan resultados mixtos. En la estación Famaillá los modelos que incorporan a la altitud muestran un mejor rendimiento, mientras que en la estación Las Mesadas no se observan diferencias significativas entre ambos modelos. Sin embargo, es importante destacar que los resultados obtenidos con Kriging en la estación Las Mesadas muestran que un bajo rendimiento y, por lo tanto, no es recomendable utilizarlos para predicciones o interpolaciones en zonas elevadas de Tucumán.

Por último, en ambas estaciones, ERA5-LAND muestra un desempeño muy inferior al de los modelos entrenados, especialmente en zonas elevadas (R~0, MAE~1100 mm, MRE~0.9, ver Tabla 4). Esto indica que los modelos entrenados logran una mejor representación de los datos observados.

5.1.3. Zona Santiago del Estero

Para analizar esta región se utilizó como estación de testeo la estación SANTIAGO DEL ESTERO AERO, la única de la provincia en la base de datos armada. Las métricas obtenidas con todos los métodos para esta estación se encuentran resumidas en la tabla 5.5 y se puede observar gráficamente en la figura 5.5.

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.56	0.53	0.49	0.57	0.37	0.46	0.04
MAE	600.1	605.5	497.8	409.6	132.4	363.4	228.2
MRE	0.97	0.98	0.81	0.68	0.08	0.62	0.16

Tabla 5.5. Métricas para la estación SANTIAGO DEL ESTERO AERO (Zona Santiago del Estero - Altura: 201 m).

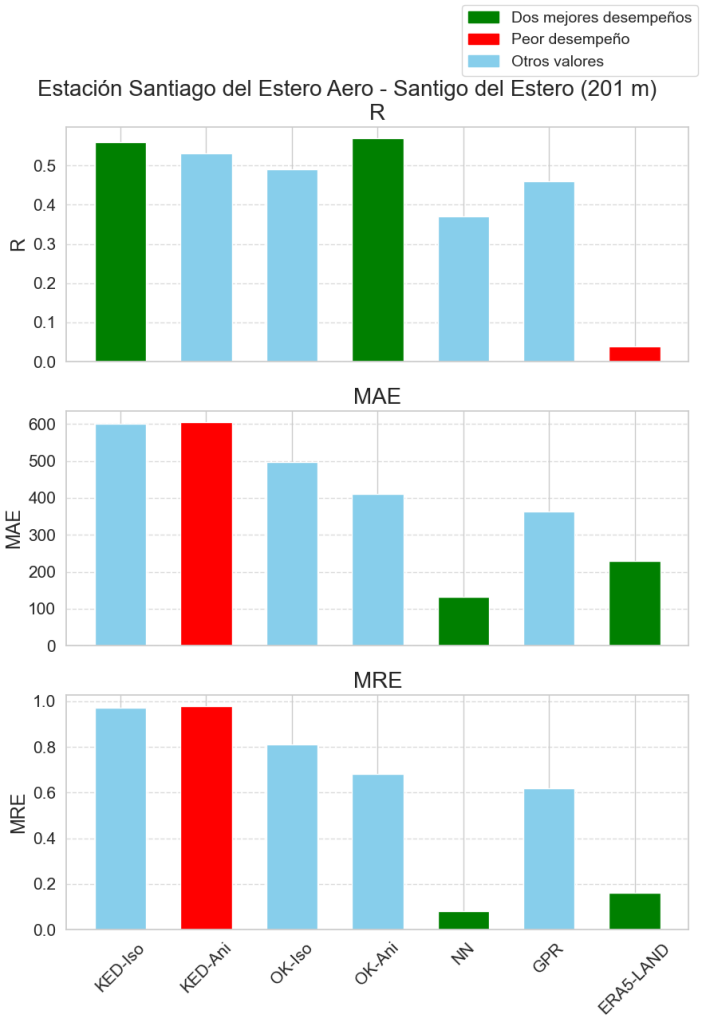

Fig 5.5. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación Santiago del Estero AERO.

Se observa que, en esta estación, el menor MAE se obtiene con NN (MAE ~ 132 mm) y, por lo tanto, este método proporciona las estimaciones de precipitaciones más precisas de la región. Sin embargo, con este método se obtuvo uno de los valores más bajos de R , por lo que no es capaz de captar la variabilidad temporal en las precipitaciones.

Un resultado a destacar es el obtenido con el modelo GPR que presenta un R más elevado que el modelo de red neuronal (R~0.46) y un valor de MAE (MAE ~363 mm) mucho menor que los obtenidos con las diferentes variantes de Kriging .

Por otro lado, los modelos de Kriging presentan algunos de los R más elevados (R>0.49). Sin embargo, también presentan los valores de MAE más elevados (MAE>410 mm), por lo que, a pesar de poder identificar la variabilidad temporal en las precipitaciones, no realizan estimaciones adecuadas del valor en sí. En las cuatro variantes de Kriging analizadas se observa que MRE es positivo y elevado, con lo que estos modelos tienden a realizar una sobreestimación significativa de las precipitaciones en la región.

Con respecto a la incorporación de la altitud en la interpolación, se observa que en esta estación el rendimiento de los modelos que la incluyen es inferior a los modelos que no la consideran. Esto sugiere que, en regiones llanas como Santiago del Estero, la inclusión de la altitud en la interpolación no solo no es necesaria, si no que puede empeorar el desempeño de los modelos. En cuanto al análisis con anisotropía, su incorporación mejoró los resultados cuando la altitud no fue utilizada como variable de deriva, mientras que no produjo cambios significativos al ser considerada dicha variable en el modelo.

Finalmente, en esta región se puede observar que las estimaciones realizadas por la base de reanálisis son razonablemente buenas. Este modelo presentó uno de los valores de MRE más bajos de todos los modelos evaluados (MRE~0.16). Sin embargo, su R es el más bajo (R~0.04), por lo que no es capaz de detectar la variabilidad temporal presente en las precipitaciones.

5.1.4. Zona Salta Central

Con el fin de analizar la región entre los núcleos de estaciones en el centro de Tucumán y en el norte de Salta, se eligió a la estación SALTA AERO para el testeo. Las métricas obtenidas con todos los métodos para esta estación se encuentran resumidas en la Tabla 5.6 y se puede observar gráficamente en la figura 5.6.

Métrica	KED Isotrópico	KED Anisotrópico	OK Isotrópico	OK Anisotrópico	NN	GPR	ERA5-LAND
R	0.45	0.20	0.40	0.17	0.56	0.02	0.14
MAE	306.5	274.2	281.9	312.8	257.2	130.2	978.1
MRE	0.45	0.40	0.41	0.46	-0.33	-0.01	1.59

Tabla 5.6. Métricas para la estación SALTA AERO (Zona Central de Salta - Altura: 1227 m)

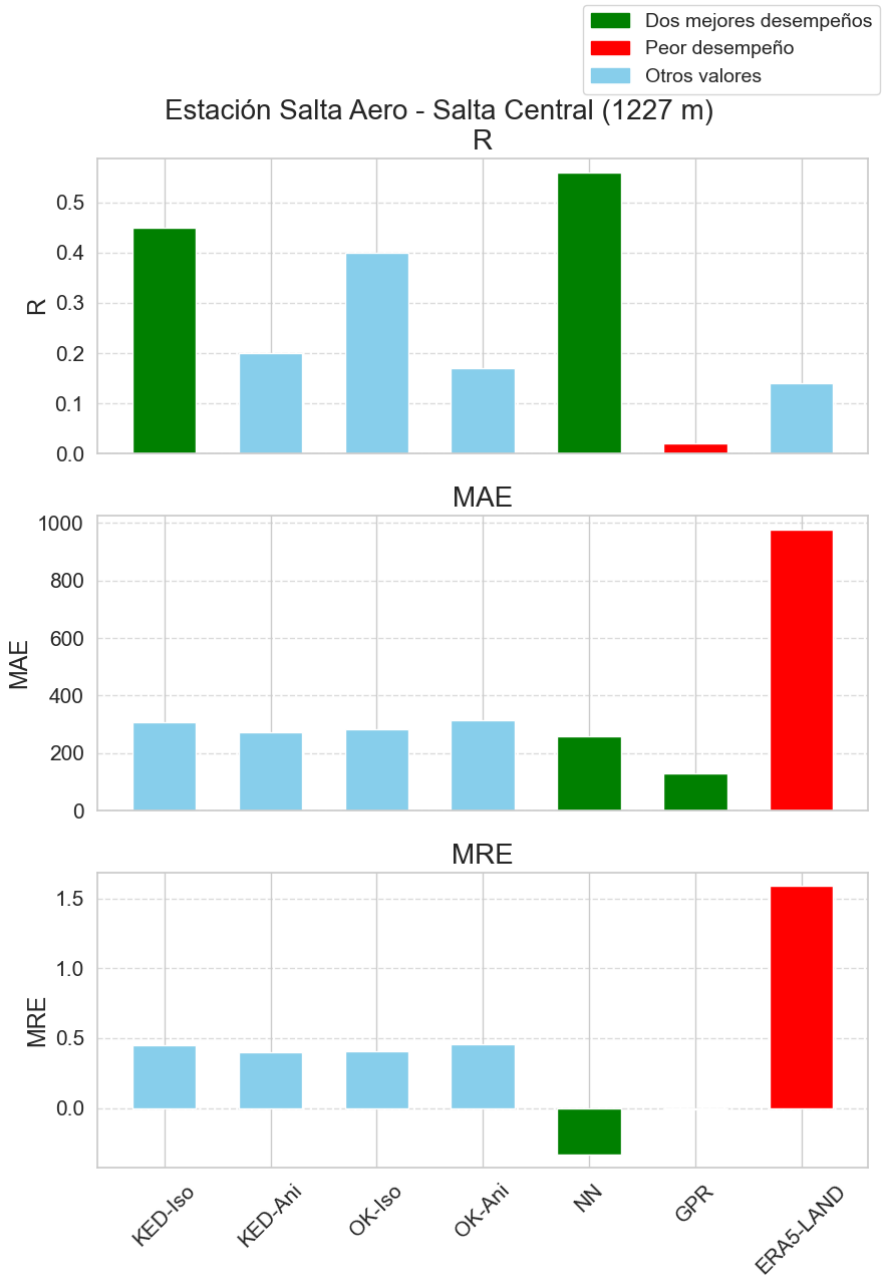

Fig 5.6. Métricas (eje vertical) obtenidas para cada método (eje horizontal) para la estación Salta AERO.

Se observa que en esta estación el menor valor de MAE se obtuvo con el modelo GPR (MAE ~ 130 mm), lo que indica que este método es capaz de estimar el valor medio de las precipitaciones de la subregión de manera adecuada. Sin embargo, presenta el valor de R más bajo (R~0.02), por lo que el modelo no logra detectar la variabilidad temporal presente en las precipitaciones. Por otro lado, los métodos con mayores R son los modelos de Kriging isotrópicos en sus dos variantes (R>0.40) y el modelo NN (R~0.56). Este último muestra un rendimiento superior, al presentar los menores valores de MAE y MRE entre los tres (MAE~257 mm, MRE~ -0.33) y el mayor valor de R.

En cuanto al modelado de la anisotropía en Kriging, se observa que empeora el rendimiento del modelo, generando una disminución de R en más de la mitad. Así, el modelo isotrópico es el más adecuado entre los de Kriging para representar la variabilidad temporal, aunque esto no supera el desempeño del modelo NN. Por su parte, la inclusión de la altura como variable de deriva generó un ligero aumento de R. Las variaciones del MAE con la inclusión de la deriva de altura y del modelado de anisotropía no presenta un patrón claro, aunque entre los cuatro modelos, el de peor rendimiento es el del modelo OK anisotrópico (MAE~313 mm). En las cuatro variantes de Kriging analizadas se observa que MRE es positivo y elevado (MRE>0.40), con lo que estos modelos tienden a sobreestimar significativamente las precipitaciones en la región.

Finalmente, se puede observar que en esta región las estimaciones realizadas por ERA5-LAND presentan uno de los peores desempeños. Este modelo obtuvo el mayor valor de MAE (MAE~978 mm) y un error relativo mayor al 150%. Si bien su R es mayor al obtenido con GPR, es bajo (R~0.14), lo que indica que no es capaz de captar la variabilidad temporal ni de realizar estimaciones adecuadas de los valores.

5.2. Análisis por modelo

Además del análisis comparativo entre modelos para determinar cuál presenta mejores resultados en cada zona, se realizó un análisis de cada modelo, con el objetivo de evaluar cómo varía el rendimiento de un mismo modelo según la ubicación geográfica. Con esto se busca identificar en qué zonas cada modelo logra mejores estimaciones.

Para llevar a cabo este análisis, se utilizaron gráficas que muestran, para cada estación, la relación entre las precipitaciones medidas y las precipitaciones predichas por el modelo a lo largo del período analizado (Figuras 5.7 a 5.20). Estas gráficas permiten observar las diferencias entre los valores observados y estimados y, además, determinar visualmente si el

método logró captar la variabilidad temporal. Por otro lado, se presentan las gráficas de error relativo por año, de manera de analizar en qué años se observa el mejor rendimiento.

Las estaciones fueron agrupadas por modelo, lo que permite observar cómo se comporta cada uno en las distintas regiones del área analizada. De esta forma, es posible detectar patrones en el rendimiento de los métodos, tales como un mejor ajuste en determinadas zonas o un sesgo sistemático en otras.

5.2.1. KED isotrópico

Las gráficas presentadas a continuación muestran las precipitaciones estimadas por el modelo y las observadas por año, para cada estación. Como este modelo incorpora la altitud como variable de deriva, es posible analizar si esta inclusión mejora el rendimiento del modelo en regiones con diferencias topográficas marcadas, en comparación con zonas en las que no hay diferencias marcadas de altitudes.

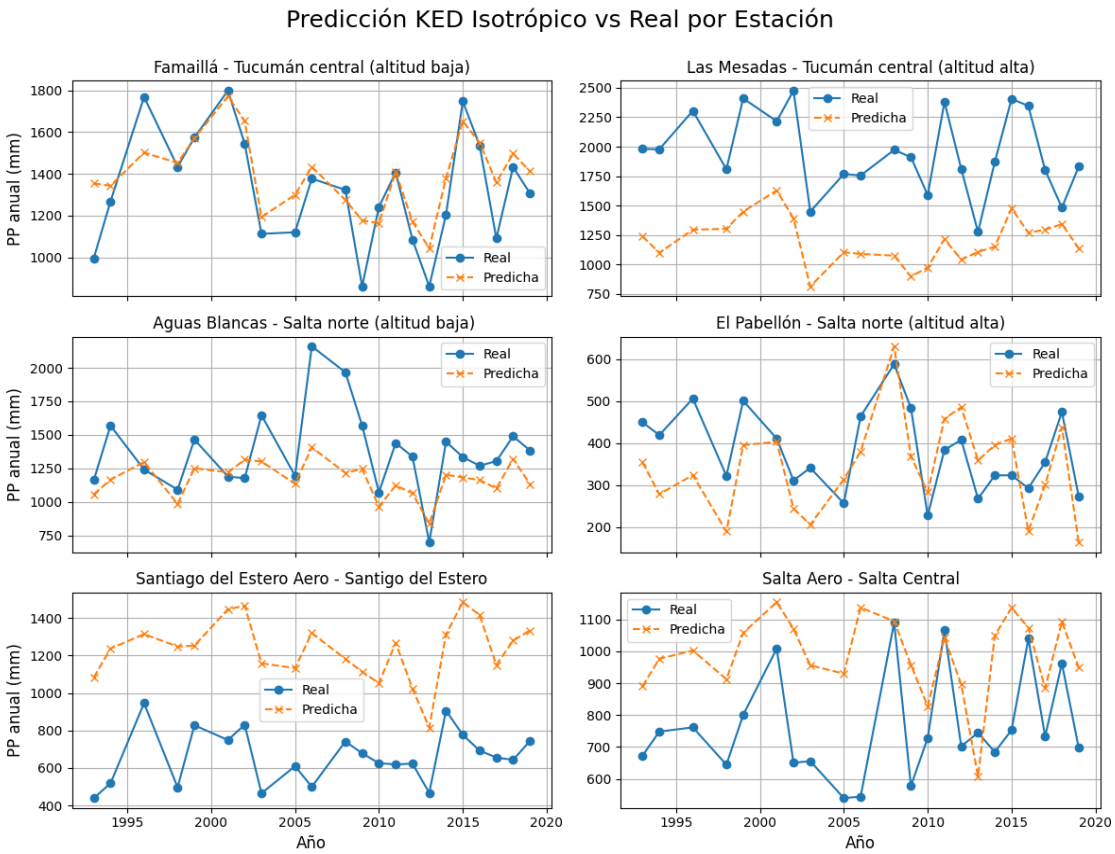

Fig 5.7. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de KED isotrópico para cada año.

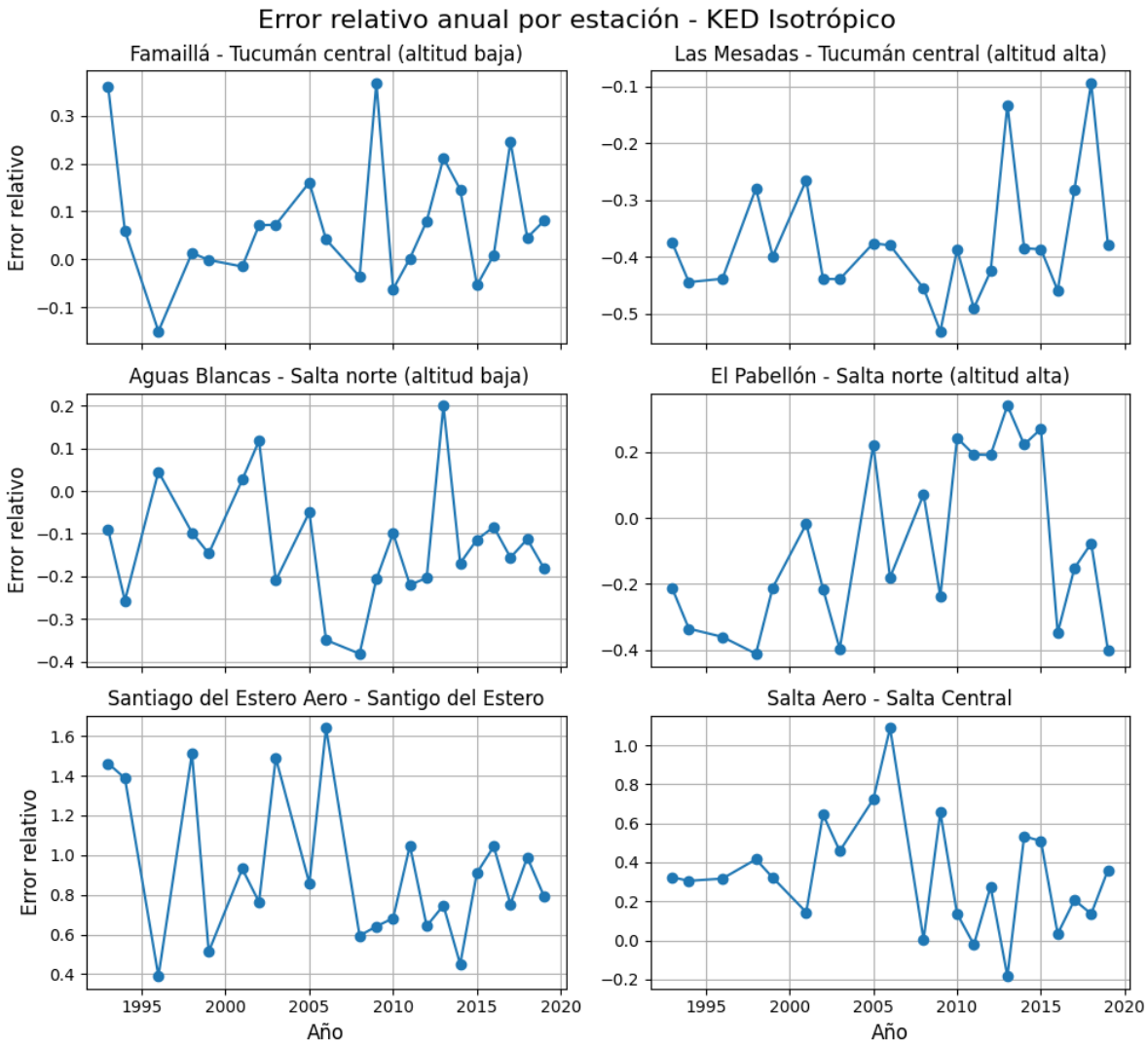

Fig 5.8. Errores relativos por año para el modelo de KED isotrópico.

Se observa que el modelo presenta su mejor desempeño en la estación Famaillá, ubicada en la región central de Tucumán y caracterizada por una altitud baja. En esta estación, el modelo logra predecir de manera adecuada tanto los valores de las precipitaciones como las tendencias temporales observadas (Figura 5.7). Dado que las estimaciones no se reducen a una media o a un valor promedio, es posible visualizar la variación interanual de las precipitaciones y distinguir los años más lluviosos o secos de la región. Además, se observa que para la mayoría de los años el error relativo es cercano al 10%, con picos cercanos al 20% y uno del 40% (Figura 5.8).

En la región de mayor altitud del centro tucumano, el modelo logra encontrar la variabilidad temporal (Figura 5.7, estación Las Mesadas), pero no logra captar la magnitud de esta

variación. Además, presenta un sesgo sistemático, observado en la subestimación de las precipitaciones para todos los años. Este sesgo genera errores relativos de entre el 30% y el 55% para todos los años (Figura 5.8, estación Las Mesadas).

En cuanto a la región norte de Salta, se observan buenos resultados para ambas estaciones. Para la estación Aguas Blancas (de menor altitud), el modelo logra captar parte de la variabilidad, pero tiene un mal desempeño en los años con precipitaciones anormalmente elevadas. Sin embargo, en general las predicciones se pueden considerar aceptables en comparación con lo observado en otras regiones del NOA, con errores relativos en la mayoría de los casos entre el 10% y 20%.

Con respecto a la estación El Pabellón, logra captar la variabilidad temporal, aunque con un rendimiento ligeramente inferior al observado en la estación Aguas Blancas. No se observa un sesgo sistemático en la predicción, pues genera tanto sobreestimaciones como subestimaciones. En cuanto a las predicciones numéricas, se observan errores relativos generalmente altos, llegando a superar el 60%.

Estas gráficas parecen indicar que el modelo puede realizar una mejor predicción de la variabilidad temporal en las zonas de baja altitud de áreas con una buena distribución de estaciones, esto es, centro de Tucumán y norte de Salta.

Por otro lado, tanto para la estación de Santiago del Estero, como aquella del centro de Salta, el modelo no presenta buenos resultados. En ambas estaciones, el modelo sobreestima significativamente las precipitaciones, especialmente en para la estación SANTIAGO DEL ESTERO AERO. El modelo parece captar la variabilidad temporal, pero falla en la predicción del valor. Para la estación SALTA AERO la detección de la variabilidad es peor, pues se observa que el modelo no puede mostrar de manera correcta cuáles fueron los años más y menos lluviosos. Para ambas estaciones los errores relativos son altos, con picos de sobreestimación cercanos o superiores al 100%.

Es importante destacar que las mejores predicciones del modelo corresponden a las cuatro primeras estaciones, ubicadas en los dos núcleos de gran disponibilidad de estaciones meteorológicas. Esto indica que el modelo tiene un mejor rendimiento en áreas donde las estaciones son abundantes y se encuentran bien distribuidas. Por lo tanto, no es recomendable utilizar este modelo para realizar predicciones de precipitaciones en áreas con escaso número de estaciones.

5.2.2. KED anisotrópico

Las Figuras 5.9 y 5.10 permiten visualizar el rendimiento del modelo con tratamiento de anisotropía y deriva de altura. Este enfoque considera la dirección preferencial de la variabilidad espacial, lo cual puede impactar de forma diferente en las distintas regiones estudiadas.

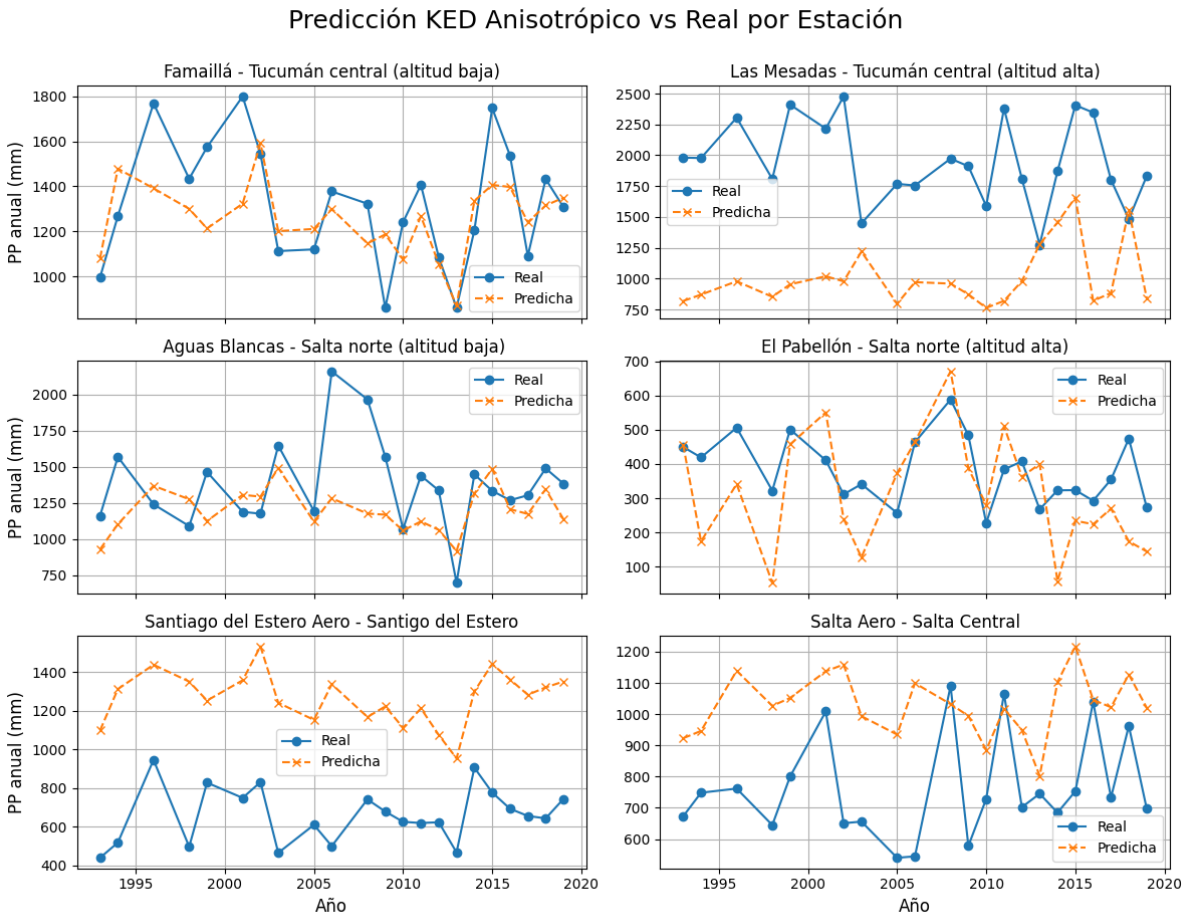

Fig. 5.9. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de KED anisotrópico para cada año.

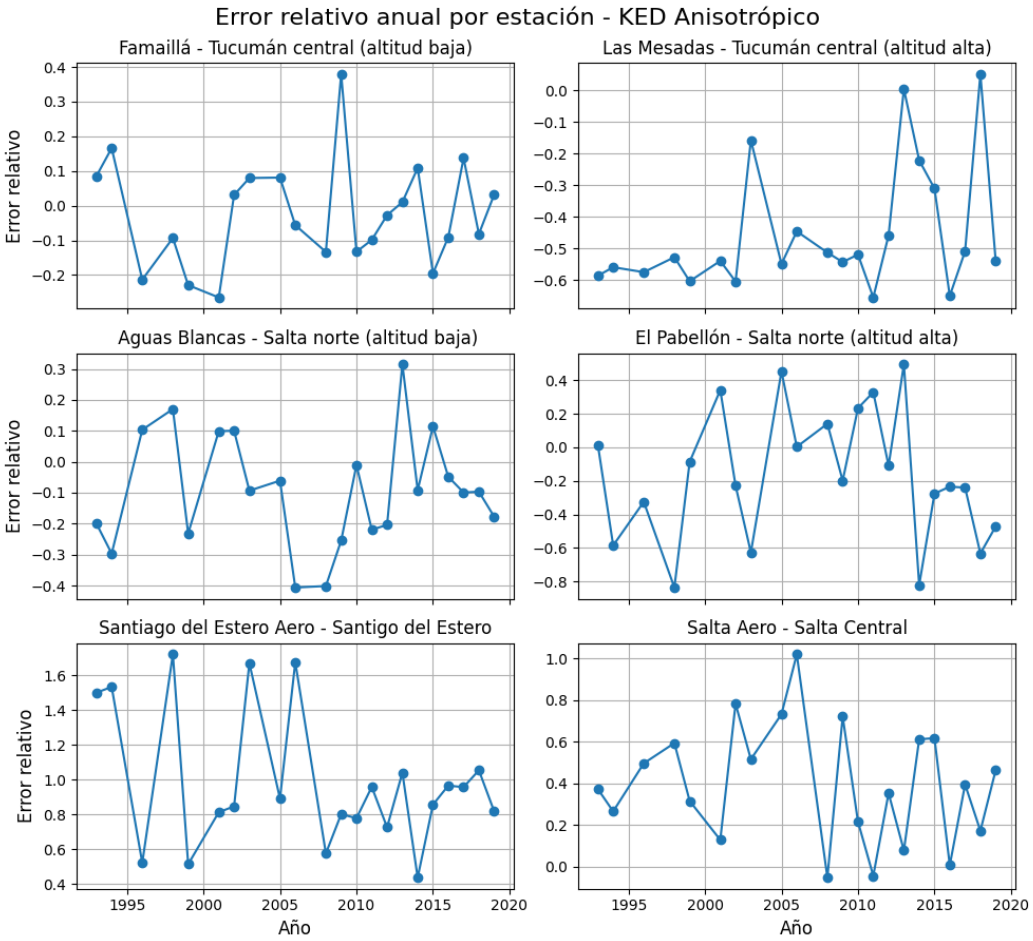

Fig.5.10. Errores relativos por año para el modelo de KED anisotrópico.

Se observa un mejor desempeño de modelo para la estación Famaillá, con errores relativos generalmente entre el 10% y 20%. Este modelo logra captar la variabilidad de las precipitaciones de manera similar al modelo que no considera anisotropía. Sin embargo, los errores relativos que presenta este modelo son mayores.

La inclusión de la anisotropía en este modelo no parece mejorar los resultados obtenidos, en especial en las áreas con menor densidad de estaciones, en donde se observa un mal ajuste de la variabilidad y estimaciones de precipitaciones poco precisas.

Se podría analizar si realizar diferentes modelos para los núcleos de Salta y Tucumán generarían un mejor ajuste con anisotropía, al tener estas regiones una mejor distribución de estaciones. Sin embargo, también se debería analizar si la cantidad de estaciones en ambas regiones es suficiente para modelar correctamente la anisotropía.

5.2.3 OK isotrópico

En este caso, se observa el desempeño del modelo sin considerar la altitud ni el modelado de la anisotropía. Las gráficas permiten evaluar cómo se ajusta este enfoque más simple a los datos medidos, y si logra capturar adecuadamente la variabilidad espacial en las distintas zonas.

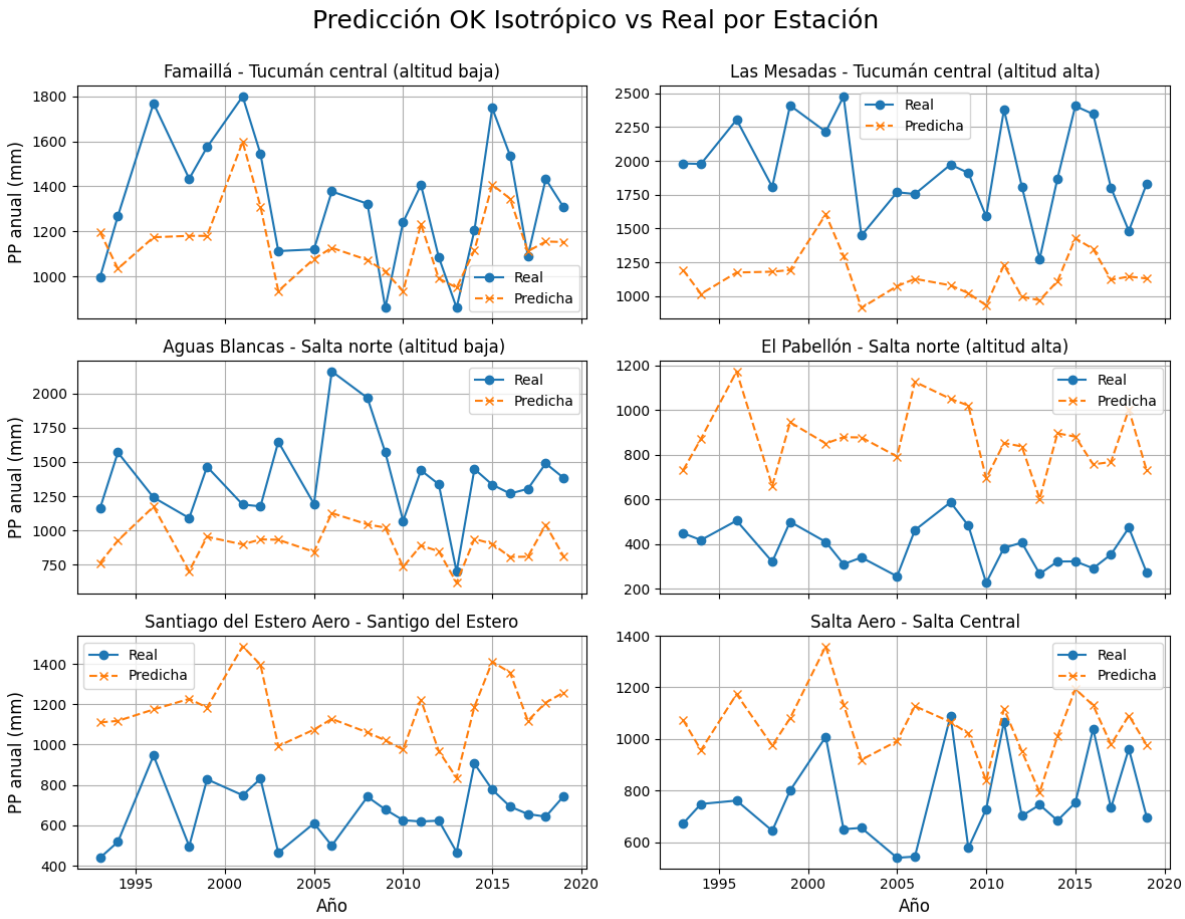

Fig.5.11. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de OK isotrópico para cada año.

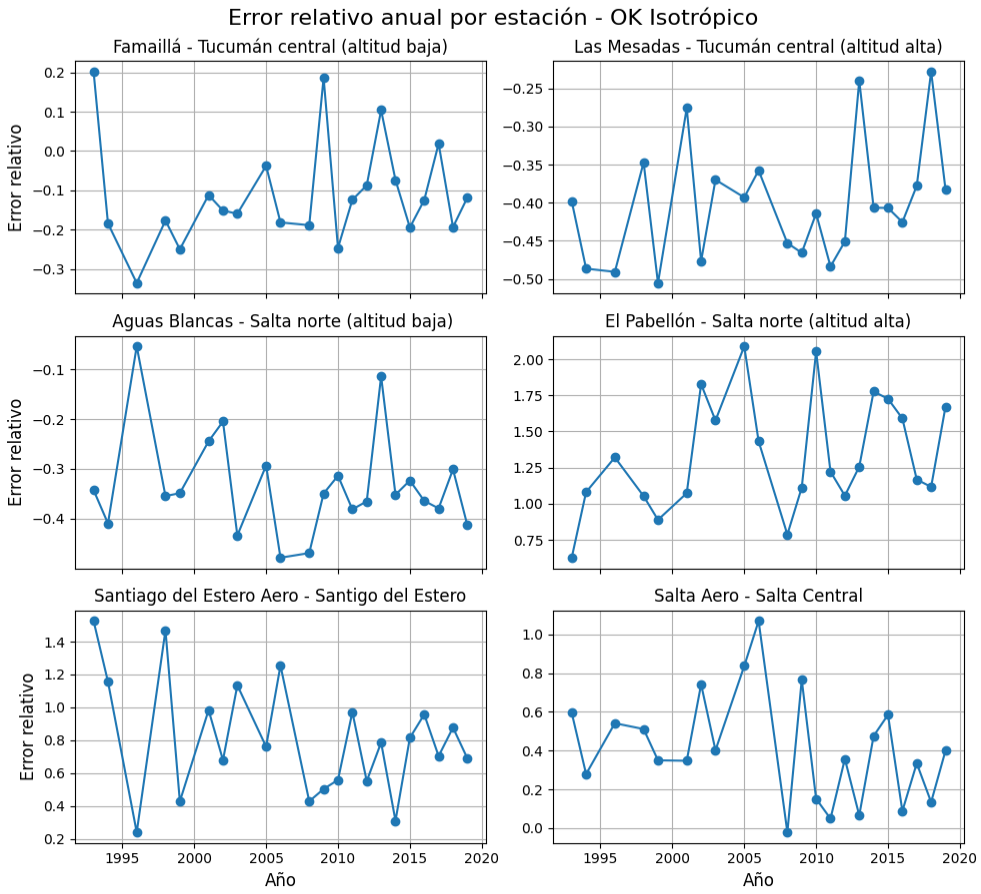

Fig.5.12. Errores relativos por año para el modelo de OK isotrópico.

Se observa un rendimiento inferior del modelo en las zonas de elevada altitud respecto de los modelos que consideran deriva, especialmente en las estaciones del norte de Salta, en las cuales los errores relativos son mayores. Además, se observa que en esas áreas el modelo no puede encontrar patrones correctos de variabilidad temporal.

El peor desempeño del modelo se observa en la estación El Pabellón, correspondiente al área de mayor altitud del norte salteño. En esta estación, el modelo produce una sobreestimación de las precipitaciones significativa, con MRE entre el 60% y 200%. Esto no fue observado en los modelos que incluyen a la altitud en la interpolación.

La comparación visual entre los diferentes modelos, muestran la importancia de tener en cuenta la altitud al realizar las interpolaciones, en especial en regiones caracterizadas por una topografía compleja.

5.2.4. OK anisotrópico

A continuación, se observan los resultados obtenidos al modelar la anisotropía sin incluir la deriva en altura (Figuras 5.13 y 5.14). El siguiente conjunto de gráficas permite evaluar si un modelo que solo considera la variación direccional logra capturar adecuadamente la variabilidad espacial en las distintas zonas.

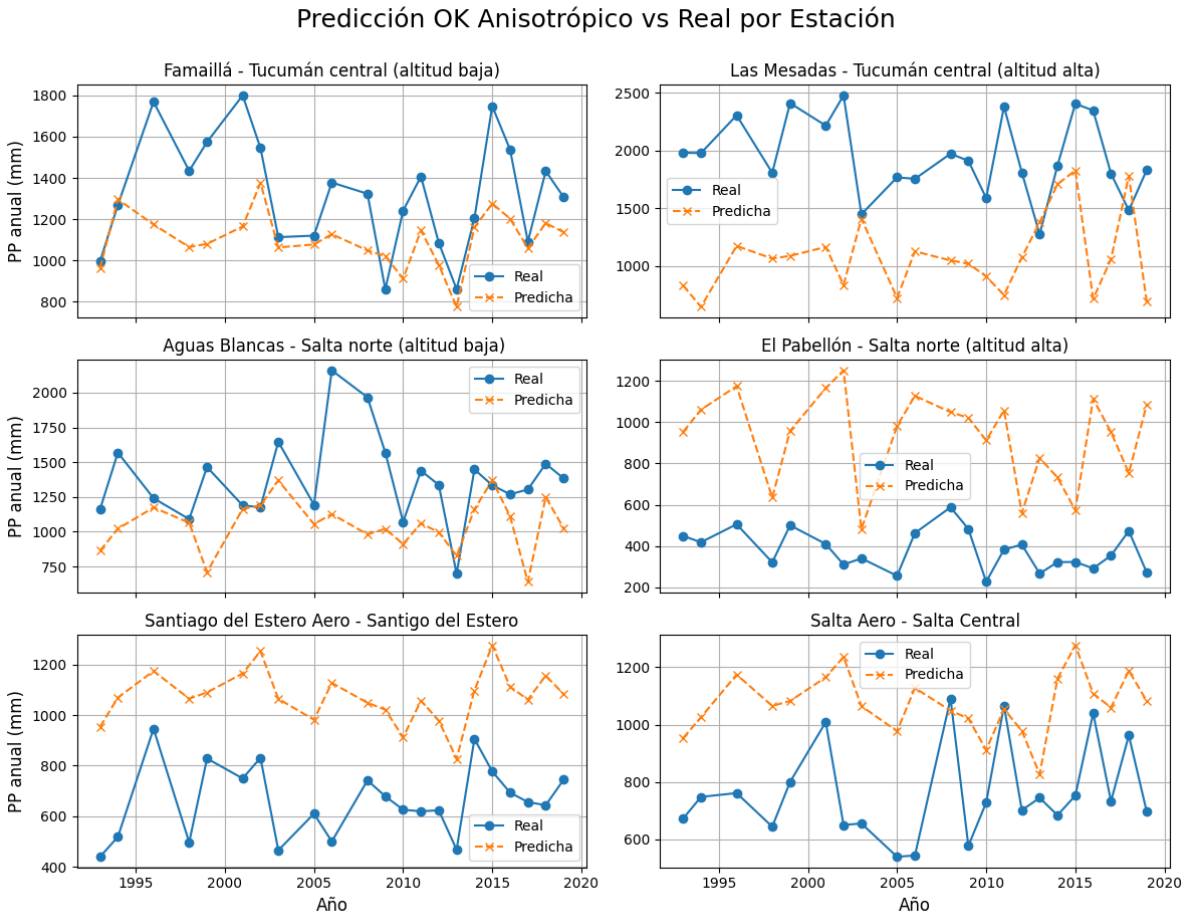

Fig 5.13. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de OK anisotrópico para cada año..

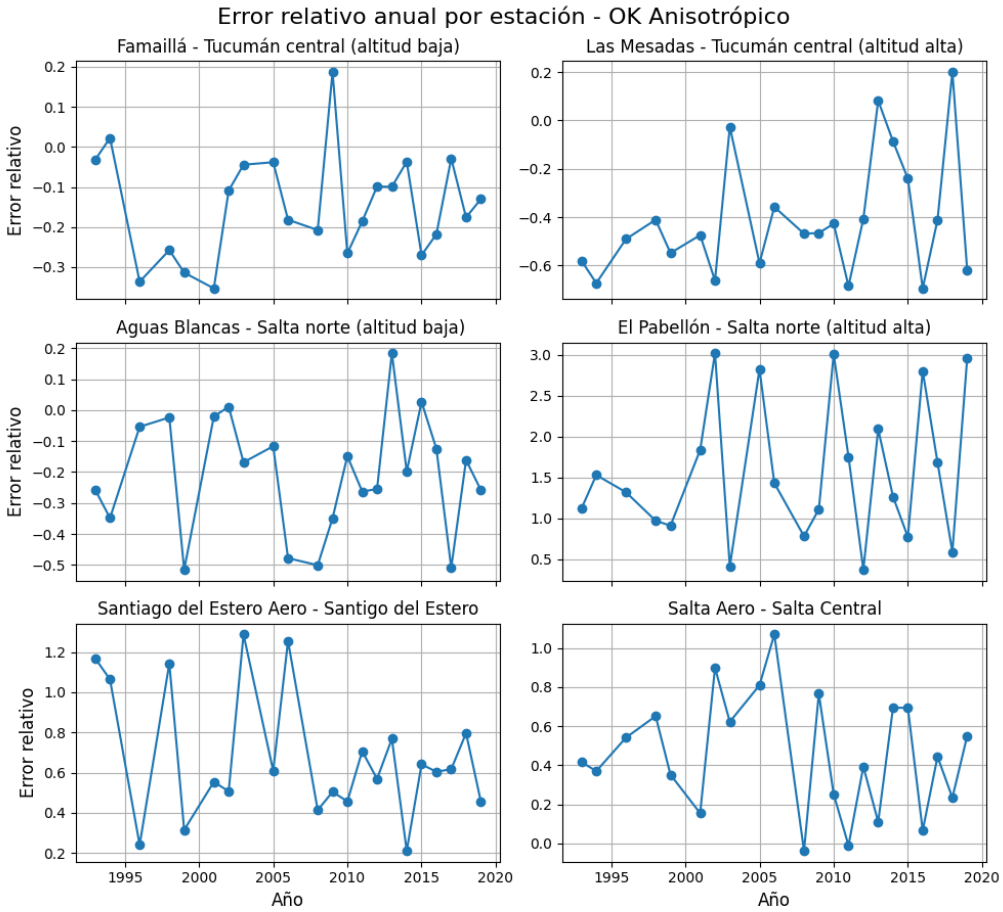

Fig.5.14. Errores relativos por año para el modelo de OK anisotrópico.

Se observa que, en general, el modelo no logra captar adecuadamente los patrones temporales de precipitación en ninguna de las estaciones analizadas, y además presenta valores de MRE elevados. En estaciones como Famaillá, Las Mesadas, Aguas Blancas y SALTA AERO, si bien en algunos años el modelo predice correctamente el valor total de precipitaciones, no consigue reproducir adecuadamente su variabilidad. Por lo tanto, los aciertos en los valores numéricos parecen responder más al azar que a un desempeño efectivo del modelo.

El peor desempeño se observa en la estación El Pabellón, donde no solo se falla en la identificación de patrones temporales, sino que también se observa un sesgo sistemático, manifestado en una constante sobreestimación de las precipitaciones.

Estos resultados destacan, una vez más, la importancia de considerar la altitud en los procesos de interpolación. Además, muestran que la incorporación de anisotropía en el modelo no necesariamente mejora los resultados.

5.2.5. Red neuronal (NN)

Las Figuras 5.15 y 5.16 muestran el desempeño de la red neuronal entrenada para estimar precipitaciones.

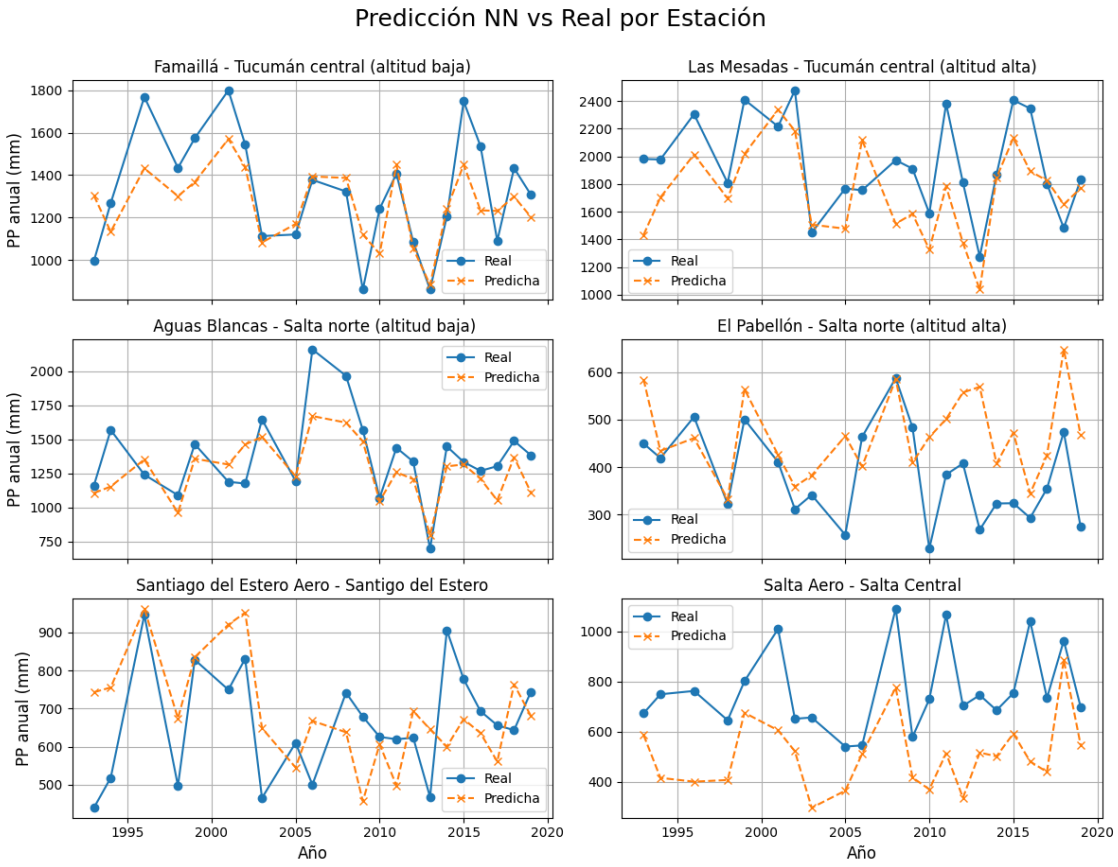

Fig.5.15. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de red neuronal para cada año.

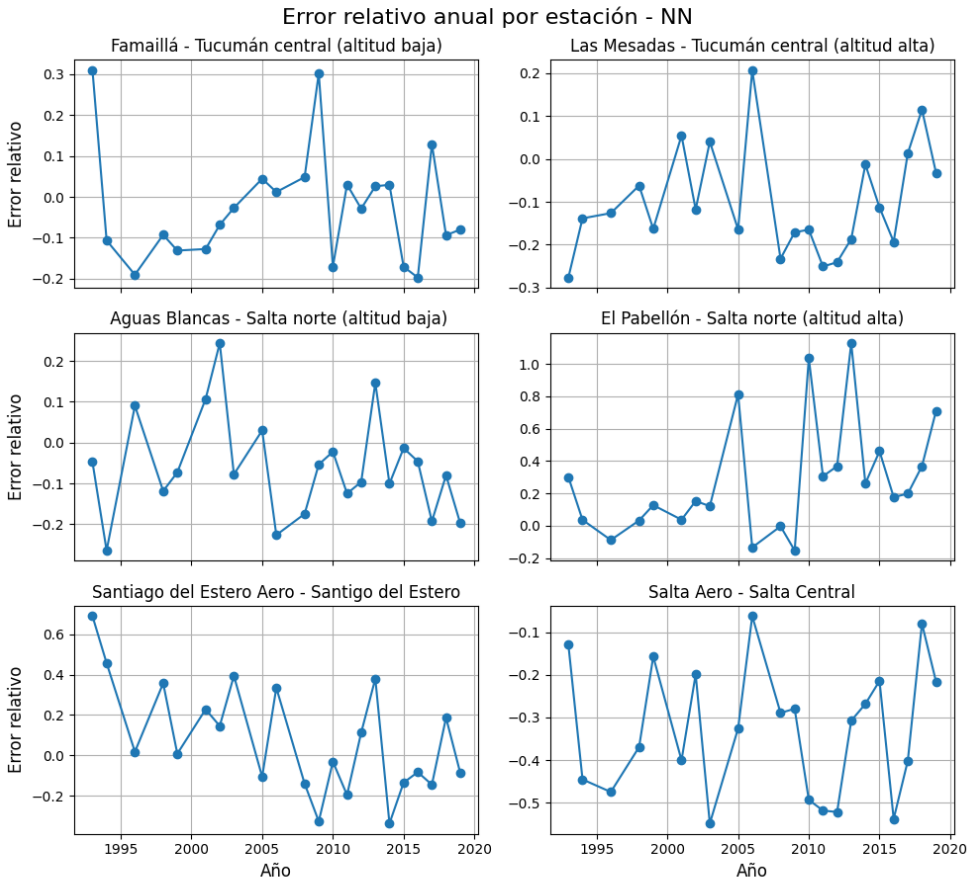

Fig.5.16. Errores relativos por año para el modelo de red neuronal.

Se observa que este modelo es capaz de captar patrones de variabilidad interanual de las precipitaciones y realizar predicciones adecuadas en la mayor parte de las estaciones analizadas. Los valores de MRE son menores en comparación con los modelos analizados anteriormente, a excepción de la estación El Pabellón en donde se evidencia un aumento significativo del error relativo.

El modelo presenta su mejor desempeño en las estaciones pertenecientes al centro tucumano. En ambas estaciones los errores no superan el 30% y en general se encuentran entre el 10% y 20%.

Un resultado a destacar es el obtenido en la estación SALTA AERO, donde los errores relativos son considerablemente menores que los obtenidos con otros métodos. Sin embargo, en esta estación se observa una subestimación sistemática de las precipitaciones.

Además, cabe resaltar el caso de la estación SANTIAGO DEL ESTERO AERO, donde el modelo logra reducir los errores relativos y captar la variabilidad interanual con más precisión

que los modelos analizados previamente. Mientras que estos últimos los errores alcanzaban valores de hasta el 160%, en este modelo no superan el 60%.

Por último, se observa que el error relativo tiende a ser más elevado en áreas con baja densidad de estaciones meteorológicas, como en las estaciones SALTA AERO y SANTIAGO DEL ESTERO AERO, así como en la estación de mayor altitud analizada (estación El Pabellón).

5.2.6. Proceso Gaussiano de Regresión (GPR)

En la Figuras 5.17 y 5.18 se presentan los resultados obtenidos con el modelo de GPR

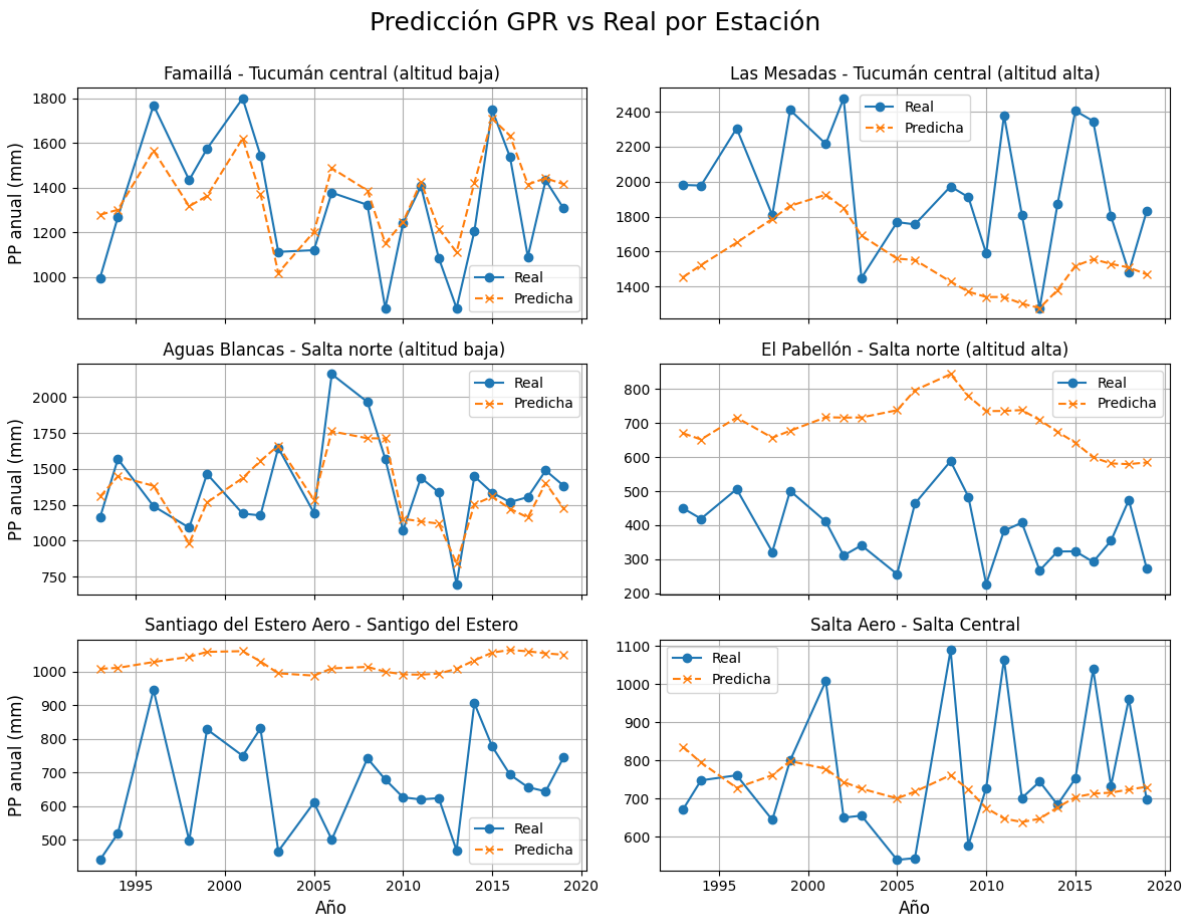

Fig.5.17. Precipitaciones medidas (línea azul) y predichas (línea naranja) por el modelo de GPR para cada año.

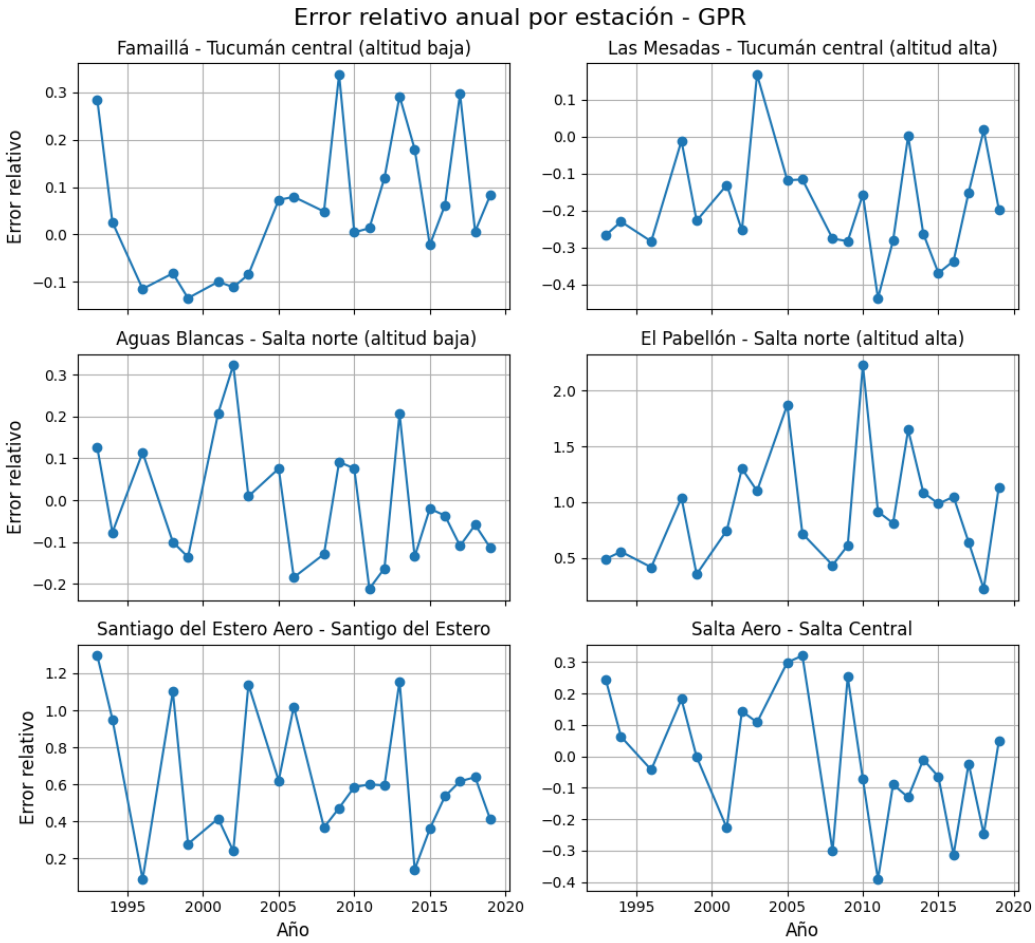

Fig. 5.18. Errores relativos por año para el modelo de GPR.

Se observa que el mejor rendimiento del modelo se obtiene para las estaciones Famaillá y Aguas Blancas, en las cuales los errores relativos no superan el 30% y se captan los patrones de variabilidad interanual. Sin embargo, en las demás estaciones, la variabilidad interanual no es captada, y las predicciones son poco precisas.

En las estaciones SANTIAGO DEL ESTERO y El Pabellón se observa una sobreestimación significativa de las precipitaciones, y los errores relativos superan el 100% en algunos años para ambas estaciones.

Por otro lado, en las estaciones Las Mesadas y SALTA AERO las predicciones coinciden con lo observado en cuanto al valor medio de las precipitaciones, aunque existe cierta subestimación.

Estas gráficas indican que el modelo presenta un mejor desempeño en las áreas de menor altitud de ambos núcleos con gran densidad de estaciones.

5.2.7. ERA5-LAND

Finalmente, se incluyen los gráficos correspondientes a la base de reanálisis ERA5-LAND (Figuras 5.19 y 5.20). Estos resultados permiten comparar el desempeño de los modelos entrenados únicamente con datos locales frente a una fuente de datos ampliamente utilizada a nivel global basada en modelos físicos y observaciones.

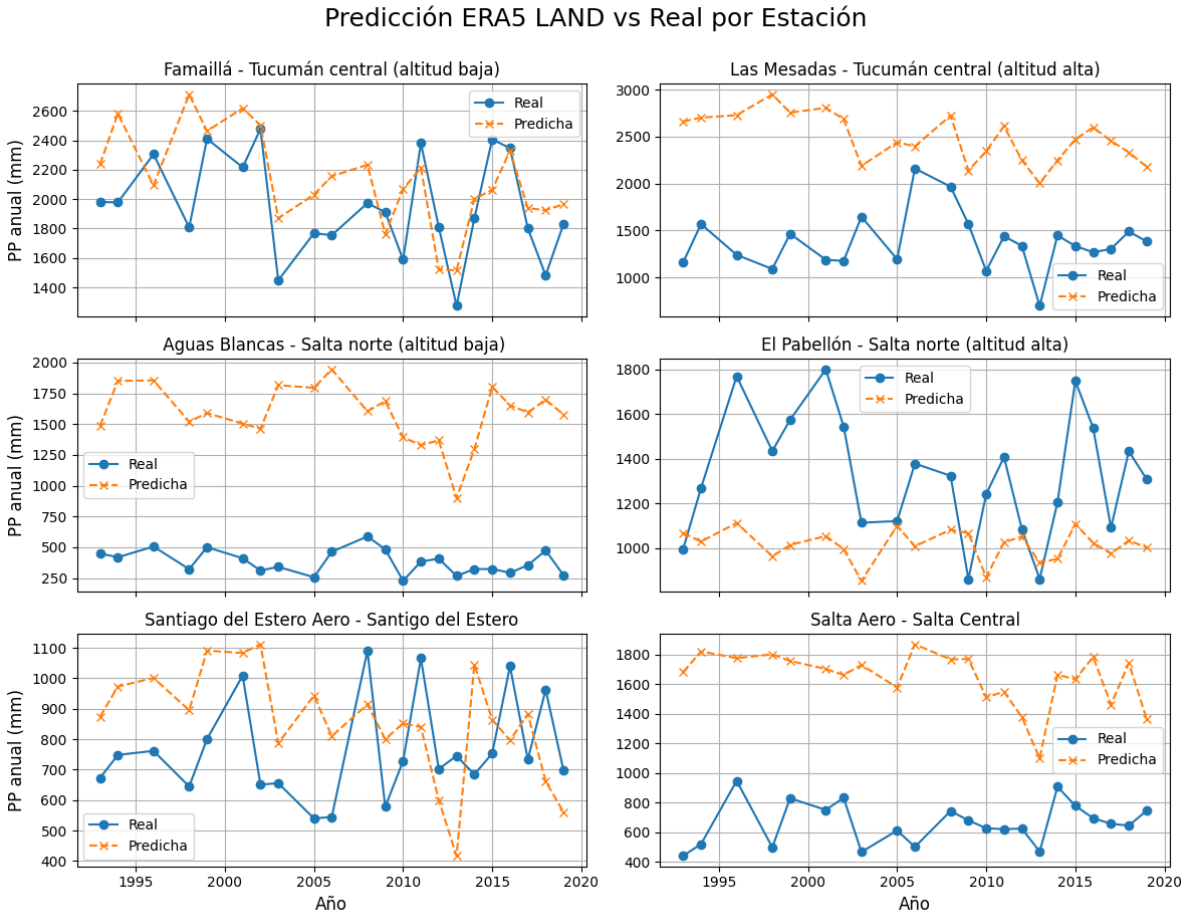

Fig. 5.19. Precipitaciones medidas (línea azul) y predichas (línea naranja) por la base de reanálisis ERA5-LAND para cada año.

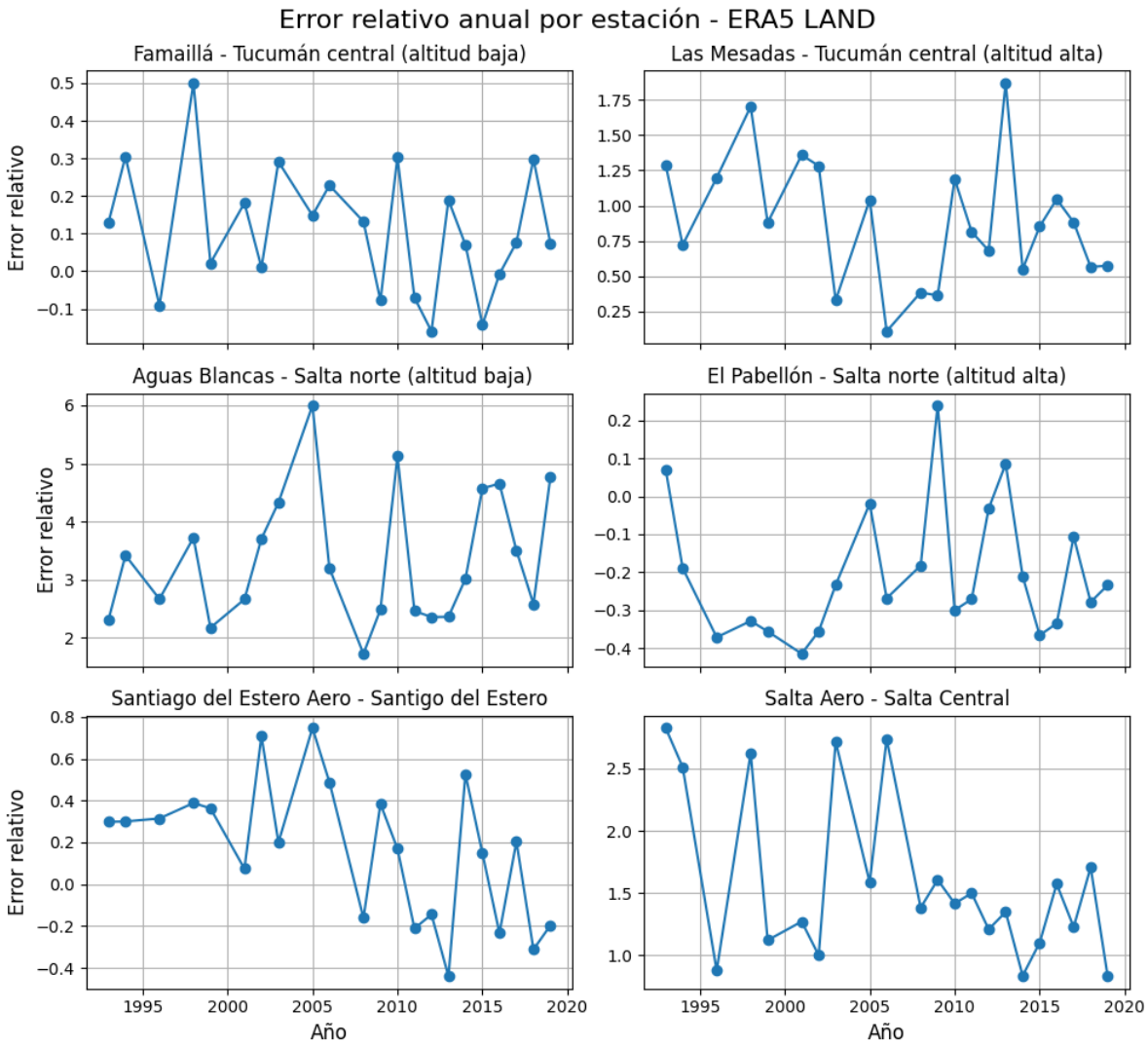

Fig. 5.20. Errores relativos por año para la base de reanálisis ERA5-LAND.

Se observa que los errores relativos para todas las estaciones analizadas son elevados. El mejor rendimiento se observa para la estación Famaillá, con valores de MRE de hasta 50%, mientras que el peor desempeño corresponde a la estación Aguas Blancas, en la cual, los errores relativos se encuentran entre el 200% y 600%. A pesar de que se observa variabilidad interanual en las precipitaciones predichas, la misma muestra poca correlación con las observaciones.

En términos generales, se observa un peor desempeño en todas las estaciones respecto a los modelos analizados anteriormente, en especial si los resultados son comparados con el modelo NN.

Del análisis de los resultados, se puede observar que, en términos generales, las Redes Neuronales presentan el mejor rendimiento en la región. En base a eso, y teniendo en cuenta que el objetivo de este trabajo es estimar valores de lluvias en lugares en donde no se tienen datos, se puede armar un mapa de precipitaciones para toda la región utilizando este modelo. La Fig 5.21 muestra un mapa de interpolaciones de lluvias en el año 2019 usando el modelo desarrollado en este trabajo.

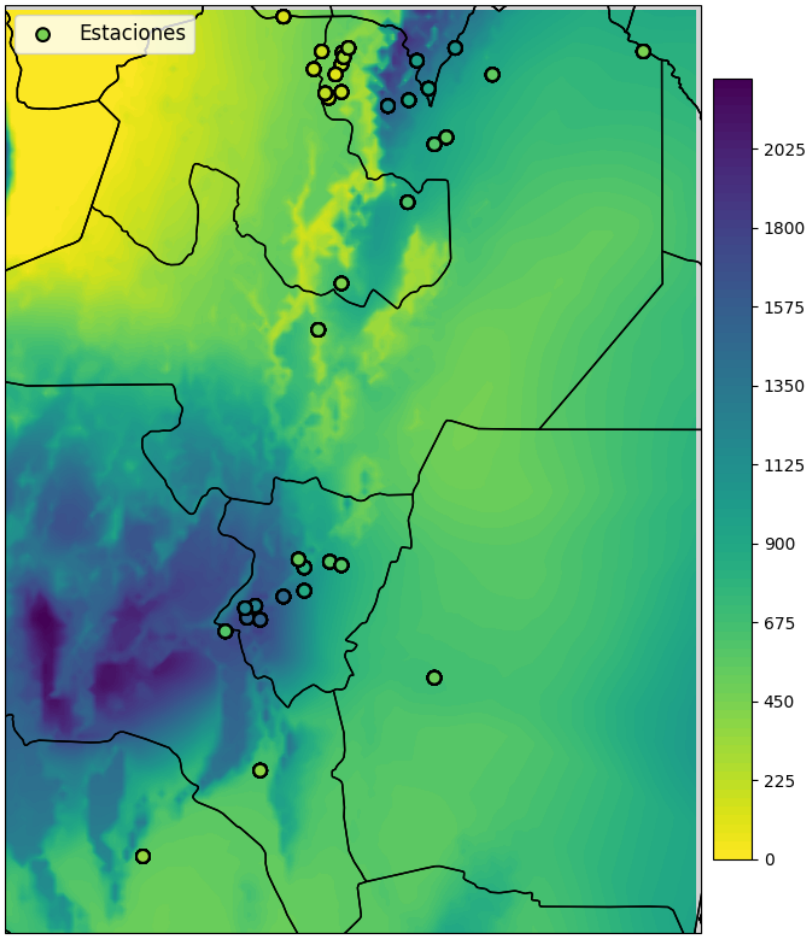

Fig 5.21. Mapa de precipitaciones anuales acumuladas (mm) obtenido con NN para el año 2019 con una resolución de 0.05 × 0.05°. En las estaciones, los colores indican las precipitaciones reales medidas para ese año.

En el mapa se observa una mancha azul sobre Catamarca y el noroeste de Tucumán (zona de los Valles Calchaquíes), que podría corresponder a un error en la predicción, ya que se trata de una región árida donde las precipitaciones reales son muy bajas. Este resultado debe interpretarse con precaución, dado que en esa zona no existen estaciones meteorológicas, por lo que las predicciones no son confiables. De todos modos, no se puede descartar que, en años

particulares, como 2019, se hayan registrado lluvias más intensas, por lo que habría que verificarlo con observaciones.

Como aspecto positivo, el mapa reproduce adecuadamente el gradiente espacial de precipitaciones en el norte: muestra una zona lluviosa en el cúmulo de estaciones del norte de Salta y una transición progresiva hacia condiciones más secas al oeste, sobre Jujuy y el norte de Chile, así como hacia el oeste de Chaco, donde la ausencia de relieve montañoso explica la marcada disminución de las lluvias. Además, los valores observados (puntos coloreados en Figura 5.21) muestran una buena concordancia con las predicciones del modelo de NN en sus alrededores.

Se observa un comportamiento similar del modelo para los demás años, como muestran las Figuras 5.22 y 5.23.

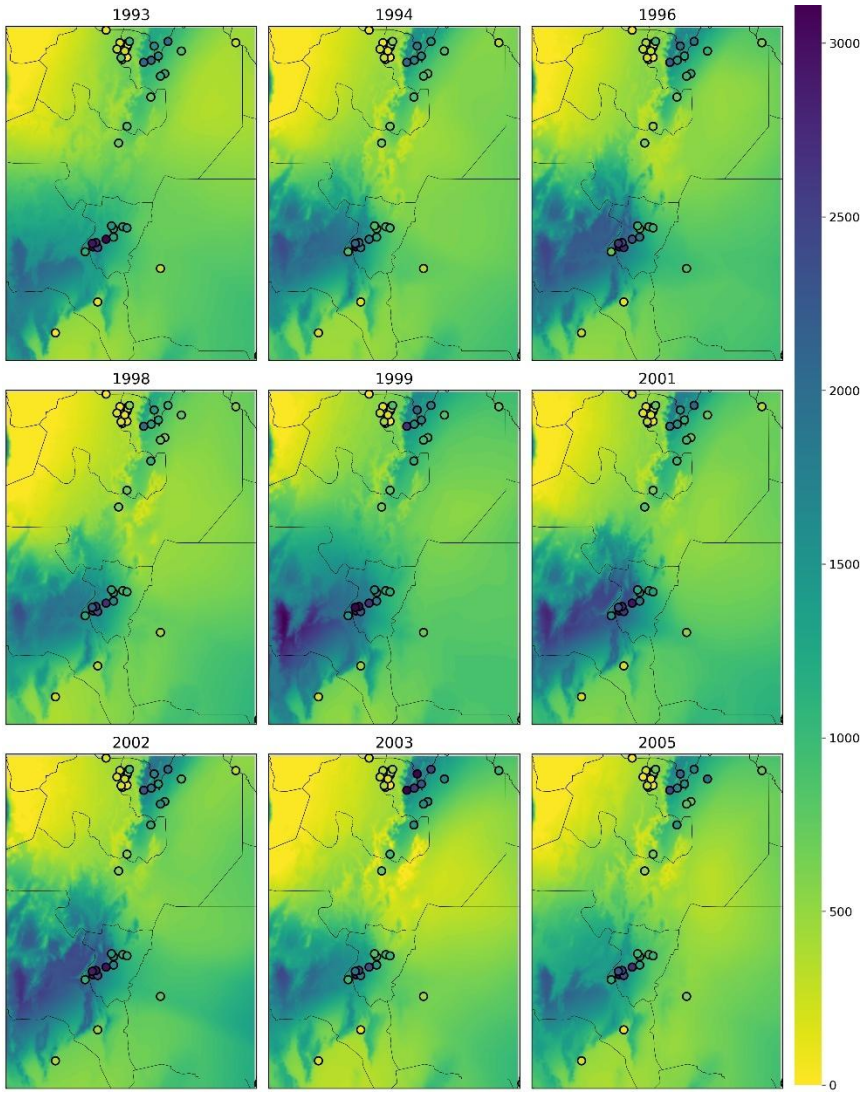

Fig 5.22. Mapas de precipitaciones anuales acumuladas (mm) obtenido con NN para el periodo 1993-2005 con una resolución de 0.05 × 0.05°. En las estaciones, los colores indican las precipitaciones reales medidas para ese año.

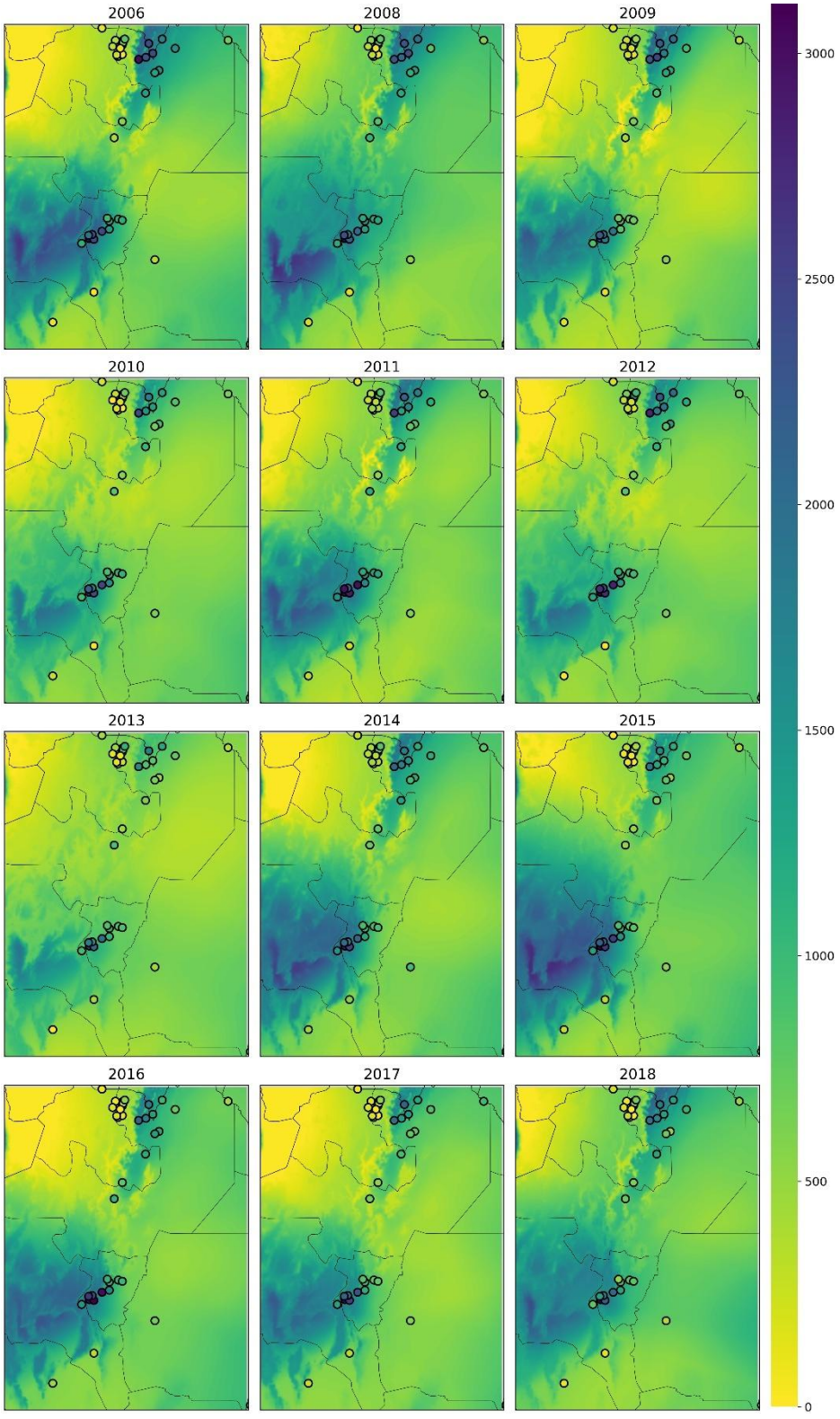

Fig 5.23. Mapas de precipitaciones anuales acumuladas (mm) obtenido con NN para el periodo 2006-2018 con una resolución de 0.05 × 0.05°. En las estaciones, los colores indican las precipitaciones reales medidas para ese año.

Capítulo 6: Discusiones

La disponibilidad de información espacialmente continua es esencial para la planificación, la gestión de riesgos y la gestión ambiental (Li et al., 2011). Sin embargo, la obtención de este tipo de datos puede ser difícil, especialmente en regiones montañosas. En este sentido, los conjuntos de datos climáticos espacialmente completos representan una fuente de información fundamental para la modelización, el monitoreo y la gestión de los recursos hídricos (Hiebl y Frei, 2017). En particular, el conocimiento de la distribución espacial de las precipitaciones es clave para la gestión de recursos hídricos, la agricultura y la prevención de desastres naturales.

La interpolación espacial de las precipitaciones representa un desafío importante en regiones donde la red de estaciones meteorológicas o pluviómetros es escasa y/o muy dispersa, ya que las mediciones en una estación solo son representativas hasta cierta distancia (Long et al., 2016). En el Noroeste Argentino (NOA), este problema es particularmente relevante, debido a que grandes áreas carecen completamente de estaciones meteorológicas, lo que genera faltantes en la información y dificulta la estimación espacial continua de las precipitaciones.

A esto se suma la complejidad topográfica del NOA, ya que la presencia de la cordillera de los Andes genera un marcado gradiente de precipitaciones en dirección Este-Oeste. Las zonas más lluviosas se encuentran entre los 300 y 3000 m de altitud, mientras que por encima de los 3000 m predominan las zonas secas (Castino et al., 2017). Además, las máximas precipitaciones anuales se observan en alturas cercanas a los 1300 m. Esta fuerte relación entre precipitación y elevación justifica la necesidad de utilizar métodos que incorporen esta variable auxiliar en los modelos de interpolación.

En este contexto, este trabajo se centra en comparar métodos geoestadísticos tradicionales con métodos de aprendizaje automático, evaluando su rendimiento mediante diferentes métricas estadísticas. Esta comparación es de gran importancia, ya que a menudo los métodos que se basan en relaciones lineales tienden a subestimar los valores de las variables climáticas. Las precipitaciones, al ser variables complejas, no muestran una relación lineal con respecto a la elevación (Bonsoms y Ninyerola, 2023).

6.1. Comparación de los métodos de interpolación

Los métodos tradicionales de geoestadística, como las diferentes variantes de Kriging, son una opción de interpolación espacial muy utilizada. Sin embargo, estos métodos requieren asumir un variograma teórico y, en áreas con poca disponibilidad de datos o datos muy

espaciados, pueden sobreestimar la correlación espacial entre los diferentes puntos, con lo que los resultados pueden no ser precisos (Long et al., 2016). En cambio, métodos no paramétricos que no se basen en supuestos espaciales pueden generar mejores resultados.

Estos métodos son ampliamente utilizados como herramientas de "mejor predicción lineal insesgada" (Best Unbiased Linear Prediction, BULP) para el análisis de datos espaciales (Hengl et al., 2018). En su forma más básica, el Kriging Ordinario (OK) combina linealmente los valores medidos, asignando pesos determinados por la correlación espacial entre los puntos, descrita por el variograma elegido (Sluiter, 2009). Este método asume una estacionariedad intrínseca, aunque en la práctica las variables meteorológicas rara vez lo son (Hengl et al., 2018).

Por otro lado, el Kriging con deriva externa (KED) es una extensión del Kriging tradicional que permite incorporar información auxiliar, como la altitud, en la interpolación. Este método utiliza la información adicional para derivar la media local de la variable principal y luego realiza Kriging sobre los residuales (Goovaerts, 2000). En KED la relación entre la altura y las precipitaciones se evalúa localmente, lo que permite tener en cuenta los cambios en esta relación a lo largo del área analizada, algo especialmente relevante en regiones como el NOA. Este enfoque es especialmente útil en regiones montañosas, donde la relación entre precipitación y altura varía espacialmente. Este método asume una relación lineal entre la variable principal y la variable secundaria (Fouedjio y Klump, 2019).

A pesar de la gran popularidad de los métodos geoestadísticos tradicionales, en las últimas décadas ha crecido el interés por emplear métodos de aprendizaje automático o *machine learning* (ML), que, aunque son más exigentes computacionalmente, no requieren supuestos estadísticos estrictos sobre la distribución o la estacionariedad de la variable a predecir (Hengl et al., 2018). Entre ellos, las Redes Neuronales (*Neural Networks*, NN) se destacan por su capacidad de modelar relaciones no lineales complejas entre múltiples variables predictoras, con lo que pueden generar resultados más precisos que los métodos tradicionales al analizar relaciones de este tipo (Sluiter, 2009; Lupi et al., 2023). Su naturaleza adaptativa, en la que el modelo "aprende" durante el proceso de entrenamiento, las convierte en herramientas populares para el modelado de procesos de alta dimensionalidad, como es el caso de los meteorológicos (Sluiter, 2009; Bottaro, 2021). Las Redes Neuronales han sido ampliamente aplicadas en las ciencias de la atmósfera y meteorología, tanto en interpolación espacial como en pronóstico temporal (eg. Bonsoms y Ninyerola, 2023; Lupi et al., 2023; Chen et al., 2015).

Hengl et al. (2018) analizan las diferencias entre el Kriging y los métodos de aprendizaje automático, en su caso, Random Forests, y destacan que, si bien las predicciones obtenidas con

ambos enfoques fueron similares, el uso de Random Forests presenta ventajas importantes. En primer lugar, no requiere definir ni ajustar un variograma, lo que simplifica notablemente el proceso. Además, evita la necesidad de asumir supuestos estadísticos estrictos acerca de la distribución y la estacionariedad de la variable objetivo.

Otra ventaja señalada por los autores es la flexibilidad para incorporar, combinar y extender covariables de distinto tipo, una característica muy importante ya que la mayoría de los métodos de Kriging suponen una relación lineal entre la variable principal y las covariables o variables auxiliares. Estas mismas ventajas pueden observarse en este trabajo al emplear NN, ya que con este método tampoco se deben realizar supuestos y el mejor desempeño obtenido con este modelo podría atribuirse a su capacidad para capturar relaciones no lineales y para incorporar múltiples variables auxiliares.

En este trabajo, las NN mostraron un rendimiento superior y más equilibrado que los demás métodos, con errores bajos y buena capacidad para capturar la variabilidad temporal. Estos resultados concuerdan con los de Bonsoms y Ninyerola (2023), quienes compararon métodos tradicionales de interpolación lineal con diferentes métodos aprendizaje automático para la interpolación de precipitaciones anuales en los Pirineos, obteniendo un R^2 de 0.85 con redes neuronales. Este valor es mayor que los logrados en este trabajo. La superioridad de los métodos de aprendizaje automático se atribuye a la posibilidad de representar relaciones no lineales entre la precipitación y variables como la altitud, algo que los modelos lineales, como KED, no pueden realizar.

Los problemas de sesgo observados en los resultados obtenidos con KED pueden deberse a que se asume una relación lineal entre las precipitaciones y la altura. Fouedjio y Klump (2019) observaron que los métodos de ML superan a KED cuando la relación entre la variable objetivo y la de deriva no es lineal, mientras que KED solo muestra ventajas cuando esta relación es estrictamente lineal. En su trabajo, estos autores compararon KED con Quantile Regression Forest (QRF), un método de ML, al igual que las NN, no asume formas funcionales específicas entre las variables. El mejor desempeño de las NN en esta tesis sugiere que la relación entre las precipitaciones y la altitud (variable auxiliar) no es lineal.

6.2. Kriging: inclusión de altura y modelado de la anisotropía

El concepto de incorporar la elevación en la interpolación de la precipitación es un enfoque estadístico establecido, como se observa en Goovaerts (2000). El Kriging con deriva externa (KED) es un método de interpolación geoestadístico ampliamente utilizado en climatología

debido a su flexibilidad y capacidad para integrar variables auxiliares, como la altura, incorporando información espacial adicional (Hiebl y Frei, 2017).

Goovaerts (2000) analizó diferentes métodos de interpolación espacial para las precipitaciones anuales y mensuales en la región de Algarve, Portugal. En su trabajo, comparó métodos que incorporan la altitud como variable auxiliar, entre ellos el Kriging con deriva externa (KED), con métodos univariantes que no consideran información adicional, incluyendo el Kriging Ordinario. Encontró que los algoritmos multivariados, entre ellos KED, superaron a los demás métodos, destacando la importancia de considerar la correlación espacial y la elevación local. Estos resultados coinciden con lo observado en este trabajo, donde el método KED presentó mejores resultados que OK en casi todas las subregiones del NOA, especialmente en aquellas con mayor altitud.

Por otro lado, Masson y Frei (2014) evaluaron diferentes extensiones de los métodos clásicos de relación precipitación-altura, utilizando principalmente el KED en una sección transversal de los Alpes Europeos. Observaron que el KED presentaba menor error de interpolación que una regresión lineal, y esto se logró con la incorporación de una única variable auxiliar (la altitud local). Por lo tanto, como concluyen Hiebl y Frei (2017), la inclusión de información adicional espacial mejoró el rendimiento de su modelo. Esto se encuentra en concordancia con lo observado en esta tesis, donde la inclusión de la altura en general mejoró las estimaciones.

Sin embargo, la inclusión de una variable auxiliar no siempre mejora los resultados respecto a métodos sin información adicional (Li et al., 2011). Esto puede deberse a asumir una relación lineal entre la variable principal y la auxiliar cuando la relación no es de este tipo. Esto concuerda con lo observado en la región de Santiago del Estero, donde la inclusión de la altura como variable auxiliar en Kriging (KED) no mejoró los resultados, observándose un menor MAE para los modelos de OK.

Con respecto a la decisión de elegir modelar o no la anisotropía, Oliver y Webster (2015) indican que para estimar un variograma isotrópico confiable se requieren al menos entre 100 y 150 puntos, y un número aún mayor cuando existe variación direccional. Esto respalda lo observado en este trabajo, donde los modelos anisotrópicos en general presentaron un rendimiento inferior que los isotrópicos, lo que puede deberse a la escasez de estaciones meteorológicas en la región (solo 48 puntos disponibles en toda la grilla). También se observa un sesgo importante en los modelos de Kriging isotrópico que puede deberse a que la escasez de estaciones generó problemas para ajustar un variograma correctamente en la región.

6.3. Influencia de la densidad de estaciones

La densidad de estaciones meteorológicas constituye un factor clave que condiciona la calidad de las interpolaciones. Herrera et al. (2019) analizaron la influencia de tres fuentes de incertidumbre: densidad de estaciones, método de interpolación y resolución espacial para la interpolación de precipitaciones en Polonia y España. En su trabajo concluyeron que la densidad de estaciones fue el factor de incertidumbre más influyente, explicando más del 60% de la varianza en España y más del 50% en Polonia. Además, encontraron que la baja densidad de estaciones amplifica la dependencia con respecto del método de interpolación utilizado, observándose que las áreas con mayor dependencia del método de interpolación correspondían a regiones con estaciones dispersas. Esto concuerda con lo observado en este trabajo: las zonas con menor densidad de estaciones mostraron errores más altos y mayores diferencias entre los distintos métodos aplicados.

Resultados similares fueron reportados por Henn et al. (2017), quienes compararon seis bases de datos de alta resolución de precipitaciones mensuales y diarias en terrenos complejos. En su trabajo hallaron que las mayores diferencias absolutas respecto de los valores de precipitaciones reales se encontraban en áreas elevadas, con errores relativos entre el 5% y el 60%. Henn et al. (2017) explican que el sesgo en áreas montañosas puede deberse a que las redes de pluviómetros en estas zonas son más dispersas y al disminuir la densidad de estaciones con la elevación, el modelo debe extrapolar el valor de las precipitaciones a partir de los valores observados en zonas bajas. Esta insuficiencia de datos observacionales puede dificultar una interpolación espacial robusta. En este trabajo se observó un patrón similar, a pesar de que los núcleos de estaciones no se concentraban exclusivamente en zonas bajas, los modelos presentaron un rendimiento inferior en regiones con baja densidad de estaciones y también en áreas de mayor altitud. Esto sugiere que la baja densidad y la mala distribución de las estaciones son una de las principales fuentes de error en la interpolación de precipitaciones.

6.4. Comparación entre Kriging y GPR

En este análisis cabe destacar que las predicciones mediante Procesos Gaussianos en la geoestadística corresponden a lo que se denomina Kriging. Sin embargo, Kriging generalmente se aplica a espacios de entrada de dos o tres dimensiones (Rasmussen y Williams, 2006), donde las variables predictoras son las coordenadas espaciales. En otras palabras, se podría decir que Kriging es una especialización de los Procesos Gaussianos, adaptada a la geoestadística. En los dos modelos de Kriging utilizados, el espacio de entrada es bidimensional, donde las variables

de entrada son la latitud y la longitud. En cambio, en el Proceso Gaussiano se puede trabajar con espacios de cualquier dimensionalidad. Por lo tanto, la principal diferencia entre los modelos realizados con Kriging y con GPR sería su implementación: en GPR se utilizó como variables de entrada el año, la latitud, longitud y la altura, mientras que en Kriging se utilizó latitud y longitud, y en KED se incorporó a la altura, pero como variable de deriva.

Por otro lado, los modelos de GPR no asumen una relación funcional entre el espacio de entrada y el de salida. En este caso, la dependencia entre entradas y salidas se modela a través de la función de covarianza. En las variantes de Kriging utilizadas, el espacio de entrada es bidimensional, con latitud y longitud como variables predictoras. Al incluir la altura en el KED como variable auxiliar, se asume una relación lineal entre ésta y las precipitaciones, realizando luego el Kriging sobre los residuales. Por lo tanto, KED establece explícitamente una relación funcional entre altura y precipitaciones.

Esta diferencia motivó el uso combinado de modelos de Kriging y GPR en este trabajo. Los modelos de Kriging, implementados con la librería GSTools, permiten modelar con facilidad la anisotropía espacial, incorporando el ángulo y la razón de anisotropía, además de incluir una variable auxiliar como la altura al utilizar Kriging con deriva externa (KED). Estos modelos resultan especialmente útiles para analizar cómo la inclusión de la altura y el modelado de la anisotropía afectan a la estimación espacial de las precipitaciones.

Por otro lado, los modelos de GPR implementados con Scikit-learn permiten incorporar múltiples variables de entrada, lo que posibilita estudiar cómo la incorporación de nuevas variables predictoras impacta en las interpolaciones espaciales. Se debe destacar que la eficiencia de los modelos de GPR disminuye en espacios de alta dimensión (cuando el número variables supera unas cuantas docenas) pero aun así permite manejar un número mayor de variables que los modelos de Kriging tradicionales.

En conjunto, la utilización de ambas aproximaciones, Kriging y GPR, permitió realizar un análisis integral, comparando los efectos de modelar explícitamente la anisotropía y la altura con los efectos de incorporar múltiples variables adicionales en el proceso de interpolación espacial.

6.5. Consideraciones epistemológicas sobre ML y métodos tradicionales

En las últimas décadas el avance de la computación y la disponibilidad masiva de datos hicieron que el aprendizaje automático (ML) pasara de ser una herramienta auxiliar a convertirse en un actor central en Ciencias Naturales y Ciencias de la Tierra en general. El uso creciente del ML en este ámbito plantea cuestiones epistemológicas y filosóficas relevantes

sobre la naturaleza del conocimiento científico. Tradicionalmente, la ciencia ha buscado formular leyes y modelos basados en principios físicos comprensibles, donde cada variable y parámetro posee un significado dentro de un marco teórico coherente. En cambio, los modelos de ML se sustentan en relaciones estadísticas complejas y en procesos de optimización interna que, si bien son altamente precisos, muchas veces operan como una "caja negra", sin ofrecer una interpretación física directa de sus resultados.

Como señala Radicella (2023), el ML representa una nueva forma de modelado "no físico" pero empírico, capaz de capturar regularidades ocultas en grandes volúmenes de datos sin requerir hipótesis explícitas sobre los mecanismos subyacentes. Este enfoque busca complementar, más que reemplazar, a los modelos tradicionales basados en la física. La sinergia entre ambos paradigmas, el empírico y el teórico, podría marcar un cambio de paradigma en la forma en que se entiende y predice el comportamiento de una variable o parámetro.

La controversia surgida tras el Premio Nobel de Física 2024, otorgado a John Hopfield y Geoffrey Hinton por sus contribuciones al desarrollo de la inteligencia artificial, refleja precisamente este debate entre la explicación y la predicción. Algunos sectores de la comunidad científica han cuestionado si el ML, centrado en la optimización estadística, puede considerarse una "contribución física" en sentido estricto. Sin embargo, otros sostienen que la capacidad del ML para descubrir patrones emergentes en sistemas complejos, como la atmósfera, constituye una nueva forma de conocimiento científico, donde la comprensión se construye a partir de los datos más que de los principios fundamentales.

En este contexto, el ML no debe verse como un reemplazo del método científico clásico, sino como una herramienta complementaria que amplía los límites de lo que puede observarse, modelarse y predecirse. Su valor filosófico radica, por lo tanto, en cuestionar cómo definimos "comprender" un fenómeno: ¿es necesario conocer las causas para predecir con precisión, o puede la predicción misma ser una forma válida de conocimiento?

Radicella (2023) enfatiza que muchas disciplinas naturales están dejando el exclusivo enfoque "modelo físico → predicción" hacia enfoques que aceptan la complejidad y la ayuda de la computación y el ML. Una diferencia filosófica central es objetivo: la ciencia tradicional prioriza la explicación y comprensión de la causalidad mientras que el ML prioriza la predicción eficaz. En meteorología esto plantea preguntas prácticas, como:

- ¿Nos alcanza con predicciones más precisas (por ejemplo, interpolaciones de precipitación) si no entendemos los mecanismos físicos subyacentes?

- ¿Puede el ML descubrir relaciones físicas ocultas, o solo correlaciones sin explicación causal?
- ¿Cómo evaluar la capacidad de generalización de un modelo de ML frente a cambios climáticos o condiciones fuera del rango de entrenamiento?

Aceptar la complejidad no implica renunciar a la búsqueda de explicaciones; por eso surgen métodos interpretables y enfoques híbridos que permiten mantener una narrativa científica clara y reproducible. Radicella (2023) discute el problema de la "caja negra": modelos potentes pero opacos. En meteorología esto tiene consecuencias prácticas y éticas:

- Reproducibilidad**:** los modelos de ML pueden ser sensibles al preprocesamiento de los datos de entrada, a la elección de hiperparámetros o a las semillas aleatorias. Para la ciencia, esto exige protocolos claros (versiones de datos, semillas, códigos).
- Responsabilidad**:** decisiones operativas basadas en modelos de ML (como alertas o manejo de recursos) requieren saber cómo fallan y cuáles son sus sesgos.

Un tema recurrente tanto en esta tesis como en Radicella (2023) es la limitación observacional. En meteorología regional, las redes de observaciones suelen ser heterogéneas y sesgadas espacialmente. Esto tiene las siguientes implicaciones:

- Conjuntos de entrenamiento sesgados (zonas densamente muestreadas) pueden llevar a que los modelos ML sobreajusten a patrones locales y fallen en zonas con pocas observaciones;
- Es fundamental documentar y cuantificar incertidumbres (validación espacial robusta, tests fuera de muestra);
- Integrar conocimiento físico, como altitud, orografía, patrones orográficos de precipitación, como covariables mejora la generalización, tal como se observa en los análisis realizados de KED y de redes neuronales.

En resumen, el creciente desarrollo del ML ha impulsado un cambio epistemológico en las ciencias de la Tierra: de enfoques centrados en la explicación a partir de modelos físicos a marcos híbridos que combinan complejidad, datos y computación. Siguiendo las reflexiones de Radicella (2023) sobre la ionosfera, es posible aplicar la misma mirada a la meteorología: muchos procesos troposféricos son altamente no lineales y multiescala, y las herramientas de ML aportan gran capacidad predictiva cuando la física a pequeña escala es incompleta o las observaciones son escasas. No obstante, surgen dilemas filosóficos y prácticos vinculados a la relación entre predicción, comprensión causal, reproducibilidad y sesgos en los datos. En este escenario, la respuesta práctica pasa por modelos híbridos (gray-box), protocolos estrictos de

reproducibilidad, validación espacial rigurosa y herramientas de interpretabilidad que permitan confiar en las predicciones para aplicaciones operativas. Es necesario integrar la potencia predictiva del ML con los valores tradicionales de la ciencia: explicación, transparencia y responsabilidad.

Capítulo 7: Conclusiones y Perspectivas a Futuro

7.1. Conclusiones generales

En este trabajo se compararon diferentes métodos de interpolación espacial de precipitaciones anuales en la región del Noroeste Argentino (NOA), incluyendo cuatro variantes de Kriging, Redes Neuronales (NN), Proceso Gaussiano de Regresión (GPR) y la base de reanálisis ERA5-LAND. Los resultados muestran que el rendimiento de los modelos depende en gran medida de la subregión analizada. En la tabla 7.1 se resumen las principales conclusiones de este trabajo.

Modelos	Desempeño	Comentarios/Conclusiones
KED Isotrópico	Buen desempeño en Norte de Salta	El desempeño mejora al incluir la altitud como deriva. Destaca la importancia de incluirla en la interpolación. Sin embargo, presenta sesgo en regiones con pocas estaciones.
Modelos Anisotrópicos de Kriging	Rendimiento inferior que para los modelos isotrópico en mayor parte de las regiones	No mejora el desempeño incluir la anisotropía; reduce R y aumenta MAE por la escasez y mala distribución de estaciones.
NN	Método más equilibrado, tiene un buen desempeño en todo el NOA	Presenta un mejor rendimiento en zonas con alta densidad de estaciones, mientras que su desempeño es menor en zonas de alta altitud o con pocas estaciones
GPR	Buen desempeño en zonas bajas con buena cobertura de estaciones	Baja capacidad para reproducir la variabilidad temporal
ERA5-LAND	Mal desempeño en casi todas las regiones	MAE altos y baja capacidad para representar la variabilidad temporal

Tabla 7.1. Resumen del desempeño de los modelos.

En términos generales, las Redes Neuronales (NN) presentan un buen desempeño en todo el NOA, con bajos valores de MAE y MRE y valores de R relativamente altos. Esto indica una buena capacidad para captar la variabilidad temporal de las precipitaciones, además de realizar estimaciones precisas. Por estas razones, el modelo de NN es el más equilibrado en la región. Sin embargo, muestra menor desempeño en las estaciones con mayor altitud y en las regiones de baja densidad de estaciones, aunque esta es una limitación general en todos los métodos.

En la Figura 5.21. se presenta el mapa de precipitación anual estimada mediante el modelo de redes neuronales para el año 2019, con una resolución espacial de $0.05° \times 0.05°$. Si bien los resultados en general son consistentes con los patrones espaciales esperados, es importante destacar que, en las regiones ubicadas al sudoeste del NOA, particularmente en las zonas altas de Catamarca, las predicciones son poco confiables debido a la ausencia de estaciones meteorológicas. En estas áreas, el modelo muestra valores de precipitación superiores a 1000 mm en zonas desérticas, lo cual carece de consistencia física. Sin embargo, en el norte del NOA los resultados presentan una mayor consistencia física ya que representan, al menos cualitativamente, el gradiente de lluvias esperado en la dirección este-oeste.

El modelo creado en esta tesis en base a NN tiene un desempeño mucho mejor que el de la base global ERA5-LAND, lo cual denota la importancia del uso de datos observados localmente para obtener estimaciones precisas en el NOA. Como desventaja general de este modelo basado en NN, es que el mismo requiere de una buena cantidad de observaciones de precipitación, lo cual no siempre está disponible en diferentes regiones del mundo. En este último caso, la base ERA5-LAND puede ser superior y/o más conveniente. Sin embargo, una ventaja importante del modelo basado en NN, es que el mismo es computacionalmente mucho menos costoso en comparación al ERA5-LAND, y logra mejores resultados en el NOA. Esto indica que no necesariamente un modelo basado en la física del sistema climático va a generar mejores resultados que un modelo estadístico basado en ML.

Por otro lado, otro modelo que mostró un rendimiento aceptable es el modelo de KED isotrópico, especialmente en la región norte de Salta. Sin embargo, este modelo tiende a sobreestimar fuertemente los valores de lluvia en algunas regiones y subestimarlos en otras. Se observa que su desempeño empeora en regiones con poca densidad de estaciones.

El método de Proceso Gaussiano de Regresión (GPR) mostró un MAE bajo en algunas estaciones, particularmente en zonas bajas y con gran cantidad de estaciones, pero su capacidad para reproducir la variabilidad temporal fue limitada (R bajas). En consecuencia, solo se recomienda su uso en las zonas de baja altura y alta densidad de estaciones.

Por último, la base de reanálisis ERA5-LAND, presentó un rendimiento consistentemente inferior a los modelos desarrollados. Sus valores de MAE son altos y no logra reproducir la variabilidad temporal de la lluvia anual en la región. Su desempeño es aceptable únicamente en zonas llanas, aunque sin superar a los métodos de interpolación desarrollados en esta tesis.

Se debe destacar que la elección del modelo dependerá de los objetivos de la interpolación. Si se prioriza la precisión en la estimación de las precipitaciones, entonces se deberá utilizar un modelo que presente valores de MAE y MRE bajos, mientras que no importará el valor de R. En cambio, si el objetivo es capturar la variabilidad temporal, se priorizará un valor alto de R. De forma general, las Redes Neuronales constituyen el modelo más equilibrado en toda la región, ya que combinan precisión de las estimaciones y capacidad de captar la variabilidad temporal, teniendo un desempeño muy bueno o aceptable en todas las subregiones.

7.2. Análisis por subregión

Este análisis permitió determinar cuál es el modelo más adecuado para cada subregión del NOA:

- Norte de Salta: en las zonas de mayor altitud, el KED isotrópico es el método más adecuado mientras que, en las zonas más bajas, las Redes Neuronales presentan mejores resultados. En toda la región, el ERA5-LAND, un modelo basado en la física del sistema climático presenta un rendimiento notablemente inferior.
- Tucumán central: en las zonas de mayor altitud, el modelo de Redes Neuronales muestra el mejor desempeño, mientras que en la parte baja cualquiera de los modelos desarrollados produce estimaciones aceptables. En ambos casos, los modelos desarrollados superan ampliamente al ERA5-LAND.
- Santiago del Estero: tanto las Redes Neuronales como el GPR presentan los mejores resultados respecto a los demás modelos, aunque los errores son mayores que en otras subregiones. El ERA5-LAND produce errores medios absolutos similares en magnitud, pero no reproduce correctamente la variabilidad temporal, lo que limita su utilidad.
- Salta central: el modelo de Redes Neuronales proporciona un desempeño aceptable, siendo el modelo más equilibrado entre los desarrollados para esta subregión. Por otro lado, se observa que la base de reanálisis ERA5-LAND presenta un mal rendimiento tanto en los valores estimados como en la variabilidad temporal.

7.3. Importancia de la inclusión de la altura y el modelado de la anisotropía

Los resultados muestran que incluir la altitud como variable de deriva mejora de manera significativa el rendimiento del Kriging en zonas montañosas (por ejemplo, el norte de Salta y Tucumán), ya que permite representar mejor los cambios en las precipitaciones asociados a las diferencias de altura en la región. Sin embargo, en regiones de baja altura como Santiago del Estero, esta incorporación puede afectar a la calidad de las estimaciones, ya que la altitud deja de ser una variable relevante.

Por otro lado, el modelado de la anisotropía no mejora el rendimiento de los modelos. En varios casos, los modelos anisotrópicos mostraron valores más bajos de R, lo que sugiere una menor capacidad para capturar la variabilidad temporal, y valores más altos de MAE, indicando una menor precisión en las estimaciones. Esto puede deberse a la distribución irregular y escasa de las estaciones meteorológicas, que dificulta una estimación confiable de la dependencia direccional de las precipitaciones. Por lo tanto, los modelos isotrópicos presentan mejores resultados en la región del NOA, caracterizada por una red de estaciones escasa y mal distribuida.

7.4. Perspectivas a futuro

Los resultados de este trabajo sugieren líneas futuras de investigación que permitirían seguir Los resultados de este trabajo sugieren líneas futuras de investigación que permitirían seguir mejorando la estimación espacial y temporal de precipitaciones en el NOA. Entre las principales perspectivas a futuro se pueden mencionar:

1. Aumentar las variables de entrada: se puede incorporar nuevas variables meteorológicas y climáticas, como temperatura, humedad o presión atmosférica lo que aportaría nueva información sobre la variabilidad climática en la región y mejoraría la interpolación espacial de las precipitaciones en el NOA.
2. Ampliar la base de datos: posteriormente a la finalización de la tesis, se obtuvieron las coordenadas de nuevas estaciones meteorológicas. La incorporación de nuevas estaciones meteorológicas en la región permitirá aumentar la densidad espacial de los datos y, con ello, mejorar la precisión de los modelos.
3. Mejorar el criterio de filtrado de datos: en este trabajo se aplicó un criterio estricto, descartando cualquier año en el que alguna estación tuviera más de 35 datos faltantes. Esto garantizó la comparabilidad entre métodos en todos los años, pero limitó la cantidad de años

que se pudieron utilizar. Se podría flexibilizar este umbral, por ejemplo, diferenciando entre datos faltantes en la estación seca y en la estación lluviosa.

4. Extender el análisis a otras escalas temporales: se pueden aplicar los métodos a la precipitación diaria, mensual y estacional, lo que permitirá evaluar su capacidad para describir la variabilidad de corto plazo. Además, se podrían utilizar los métodos de aprendizaje automático para las predicciones en diferentes escalas temporales si se incorporan más variables.

5. Continuar con el análisis comparativo de los modelos obtenidos en esta tesis con otras bases globales de alta resolución (por ejemplo, IMERG, CHIRPS, MERRA, etc.).

Bibliografía

Bonsoms, J., & Ninyerola, M., 2023. Comparison of linear, generalized additive models and machine learning algorithms for spatial climate interpolation. *Theoretical and Applied Climatology.* https://doi.org/10.1007/s00704-023-04725-5

Bottaro, F., 2021. Spatialization on grid of meteorological variables with machine learning methods. *Master Thesis*, Universita degli Studi di Padova (Universidad de Padua).

Cai, W., McPhaden, M.J., Grimm, A.M. et al., 2020. Climate impacts of the El Niño–Southern Oscillation on South America. *Nat Rev Earth Environ* 1, 215–231. https://doi.org/10.1038/s43017-020-0040-3

Castino, F., Bookhagen, B., Strecker, M. R., 2017. Rainfall variability and trends of the past six decades (1950–2014) in the subtropical NW Argentine Andes. *Climate Dynamics*, 48(3), 1049–1067. https://doi.org/10.1007/s00382-016-3127-2

Chen, L.; Han, B.; Wang, X.; Zhao, J.; Yang, W.; Yang, Z., 2023. Machine Learning Methods in Weather and Climate Applications: A Survey. *Appl. Sci.*, 13, 12019. https://doi.org/10.3390/app132112019

Colamonaco, S., 2024. *Deep Learning Models for Downscaling of Meteorological Variables.* Master Thesis, University of Bologna.

Ferrero, M.A., & Villalba, R., 2019. Interannual and Long-Term Precipitation Variability Along the Subtropical Mountains and Adjacent Chaco, 22–29° S: in Argentina. *Front. Earth Sci.*, 7, 148. https://doi.org/10.3389/feart.2019.00148

Fogt, R.L., & Marshall, G.J., 2020. The Southern Annular Mode: Variability, trends, and climate impacts across the Southern Hemisphere. *WIREs Clim Change*, 11:e652. https://doi.org/10.1002/wcc.652

Fouedjio, F., & Klump, J., 2019. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental Earth Sciences*, 78, 655. https://doi.org/10.1007/s12665-018-8032-z

Fowler, H.J., Lenderink, G., Prein, A.F. et al., 2021. Anthropogenic intensification of short-duration rainfall extremes. *Nat Rev Earth Environ* 2, 107–122. https://doi.org/10.1038/s43017-020-00128-6

Funk, C., Peterson, P., Landsfeld, M. et al., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci Data* 2, 150066. https://doi.org/10.1038/sdata.2015.66

Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation.* Oxford University Press, Applied Geostatistics Series, Oxford, UK, 483 p. ISBN 0-19-511538-4.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228, 113–129. https://doi.org/10.1016/S0022-1694(00)00144-X

Gupta, V., Jain, M.K., Singh, P.K., & Singh, V., 2019. An assessment of global satellite-based precipitation datasets in capturing precipitation extremes: a comparison with observed precipitation dataset in India. *Int. J. Climatol.* https://doi.org/10.1002/joc.6419

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518. https://doi.org/10.7717/peerj.5518

Henn, B., Newman, A.J., Livneh, B., Daly, C., & Lundquist, J.D., 2017. An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology.* http://dx.doi.org/10.1016/j.jhydrol.2017.03.008

Herrera, S., Kotlarski, S., Soares, P. M. M., Cardoso, R. M., Jaczewski, A., Gutiérrez, J. M., & Maraun, D., 2019. Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution. *International Journal of Climatology*, 39, 3717–3729. https://doi.org/10.1002/joc.5878

Hiebl, J., & Frei, C., 2017. Daily precipitation grids for Austria since 1961—development and evaluation of a spatial dataset for hydroclimatic monitoring and modelling. *Theoretical and Applied Climatology*, 132, 327–345. https://doi.org/10.1007/s00704-017-2093-x

Li, J., Heap, A. D., Potter, A., & Daniell, J. J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12), 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B., 2022. *Machine Learning: A First Course for Engineers and Scientists.* Cambridge University Press. https://doi.org/10.1017/9781009097058

Long, Y., Zhang, Y., & Ma, Q., 2016. A merging framework for rainfall estimation at high spatiotemporal resolution for distributed hydrological modeling in a data-scarce area. *Remote Sensing*, 8(7), 599. https://doi.org/10.3390/rs8070599

Lupi, A., Luppichini, M., Barsanti, M., Bini, M., & Giannecchini, R., 2023. Machine learning models to complete rainfall time series databases affected by missing or anomalous data. *Earth Science Informatics.* https://doi.org/10.1007/s12145-023-01122-4

Mantua, N., & Hare, S., 2002. The Pacific decadal oscillation. *Journal of Oceanography*, 58, 35–44.

Masson, D., & Frei, C., 2014. Spatial analysis of precipitation in a high-mountain region: exploring methods with multi-scale topographic predictors and circulation types. *Hydrol. Earth Syst. Sci.*, 18, 4543–4563. https://doi.org/10.5194/hess-18-4543-2014

Medina, F., Zossi, B., Bossolasco, A., & Elias, A., 2023. Performance of CHIRPS dataset for monthly and annual rainfall-indices in Northern Argentina. *Atmospheric Research*, 283, 106545. https://doi.org/10.1016/j.atmosres.2022.106545

Medina, F.D., Zossi, B.S., & Elias, A.G., 2025. Changes on the summer Rx1-ENSO and Rx1-SAM relations over Northern Argentina driven by PDO phases. *Theor Appl Climatol*, 156, 95. https://doi.org/10.1007/s00704-024-05350-6

Minetti, J. L., 2009. *El Clima del Noroeste Argentino.* Laboratorio Climatológico Sudamericano.

Muñoz Sabater, J., 2019. ERA5-Land monthly averaged data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS).* https://doi.org/10.24381/cds.68d2bb30

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N., 2021. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

O'Neill, B., van Aalst, M., Zaiton Ibrahim, Z., Berrang Ford, L., Bhadwal, S., Buhaug, H. et al., 2022. Key Risks Across Sectors and Regions. In: *Climate Change 2022: Impacts, Adaptation and Vulnerability.* Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, et al. (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 2411–2538. https://doi.org/10.1017/9781009325844.025

Oliver, M. A., & Webster, R., 2015. *Basic Steps in Geostatistics: The Variogram and Kriging.* SpringerBriefs in Agriculture. https://doi.org/10.1007/978-3-319-15865-5

Organización Meteorológica Mundial -World Meteorological Organization (WMO)-, 2021. *WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019).* WMO-No. 1267. https://wmo.int/publication-series/atlas-of-mortality-and-economic-losses-from-weather-climate-and-water-related-hazards-1970-2021

Radicella, S.M., 2023. New Ways to Modelling and Predicting Ionosphere Variables. Atmosphere, 14, 1788. https://doi.org/ 10.3390/atmos14121788

Rasmussen, C. E., & Williams, C. K. I., 2006. *Gaussian Processes for Machine Learning.* MIT Press. ISBN 026218253X.

Shrestha, D., Sharma, S., Hamal, K., Jadoon, U.K., & Dawadi, B., 2021. Spatial Distribution of Extreme Precipitation Events and Its Trend in Nepal. *Applied Ecology and Environmental Sciences*, 9(1), 58–66. http://pubs.sciepub.com/aees/9/1/8

Skansi, M.d.l.M., Brunet, M., Sigro, J., Aguilar, E., et al., 2013. Warming and wetting signals emerging from analysis of changes in climate extreme indices over South America. *Global Planet. Change*, 100, 295–307. https://doi.org/10.1016/j.gloplacha.2012.11.004

Sluiter, R., 2009. *Interpolation methods for climate data.* KNMI Technical Report; Royal Netherlands Meteorological Institute, De Bilt, The Netherlands.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L., 2018. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, 56, 79–107. https://doi.org/10.1002/2017RG000574