Propuesta de Trabajo Fin de Máster

Datos identificativos

Título tentativo
Simulación del caudal en España utilizando redes Long Short-Term Memory
Estudiante
Jesús Casado Rodríguez
Fecha de propuesta/revisión
29 mayo 2023

Breve introducción

Los modelos hidrológicos son algoritmos capaces de reproducir el ciclo del agua. Tradicionalmente, estos modelos están basados en ecuaciones que reproducen de forma aproximada los procesos físicos que tienen lugar en el ciclo hidrológico: precipitación, infiltración, escorrentía, flujo subterráneo del agua, traslación del caudal en los cauces… Los modelos se alimentan de datos meteorológicos (precipitación, temperatura, etc.) y fisiográficos (altitud, pendiente, características del suelo, de la vegetación, etc.) y simulan diversas variables de estado del ciclo hidrológico, siendo la más habitual el caudal, pero pudiendo abarcar otras variables como la cobertura nival, humedad del suelo, nivel freático, etc. Este tipo de modelos siguen siendo los más extendidos en el campo de la hidrología, por ejemplo, para la predicción de avenidas en tiempo real.

A este tipo de modelos pertenece LISFLOOD Open Source [1], [2], desarrollado por el Joint Research Centre de la Comisión Europea. LISFLOOD-OS se utiliza, entre otros, en las simulaciones de EFAS (*European Flood Awareness System*) y de GloFAS (*Global Flood Awareness System*). Para su implementación en GloFAS, el modelo LISFLOOD-OS fue calibrado en cerca de 2000 estaciones de aforo por todo el mundo, partiendo de cerca de 100 mapas con las características físicas de las cuencas y las series temporales de caudal, precipitación y temperatura observadas. La calibración es un proceso muy costoso en cuanto a tiempo en el que los parámetros del modelo físico se ajustan específicamente para cada subcuenca hidrográfica de manera que las simulaciones reproduzcan el caudal observado [3], [4].

Durante los últimos años, con la irrupción de los modelos de aprendizaje automático y aprendizaje profundo, ha aparecido un nuevo tipo de modelos hidrológicos, los puramente basados en datos [5], [6]. Estos modelos utilizan la capacidad que las redes neuronales tienen de extraer patrones de cantidades masivas de datos para simular aspectos del ciclo hidrológico, como puede ser el caudal de los ríos. En los últimos años se ha demostrado la potencialidad de las redes neuronales recurrentes, en concreto las LSTM (Long Short-Term Memory), para la simulación de series temporales de caudal [5],

[7]–[14]. Existen estudios comparativos de estas LSTM y los modelos tradicionales donde se muestra que el uso de redes neuronales supera en rendimiento a los modelos tradicionales. Otras ventajas que se atribuyen a estos métodos son la rapidez tanto en la calibración del modelo (puesto que se calibra un único modelo para toda el área de estudio, en lugar de una calibración para cada subcuenca), y de ejecución (puesto que los modelos tradicionales o físicamente basados deben de resolver ecuaciones en ocasiones complicadas en mallas de cada vez mayor resolución). Por su parte, las desventajas que se atribuyen a este tipo de modelos son principalmente dos: el hecho de que son cajas negras donde no se sabe qué pasa dentro del modelo, y que su aplicación es específica a una variable concreta (por ejemplo, el caudal), es decir, no se simulan el resto de los procesos del ciclo hidrológico.

Con la idea de extraer el potencial de ambos tipos de modelos surgen los modelos híbridos [15]. En ellos se aplican redes neuronales dentro del proceso de modelado, bien en la calibración del modelo físico, bien dentro del propio modelo, bien en el post-procesado de los resultados. En el primero de estos tipos, el que utiliza redes neuronales en la calibración, aparece el concepto de *parameter learning* [16]. La idea principal es que una red neuronal aprenda a estimar los parámetros del modelo físico en función de otras variables, y así poder generar los parámetros adecuados para cualquier ubicación. Un requisito indispensable para poder aplicar *parameter learning* sobre un modelo físico es que sea diferenciable, de modo que pueda aplicarse el algoritmo de retropropagación en su entrenamiento. Un método alternativo es crear una red neuronal que emule el funcionamiento del modelo físico (lo que se conoce como un modelo subrogado). La técnica del *parameter learning* tiene dos ventajas principales. Por un lado, se calibra un único modelo que incluye todas las subcuencas de interés. Por otro lado, como resultado se obtiene la relación entre las variables fisiográficas y los parámetros del modelo físico, lo que permite generar los parámetros del modelo físico en cuencas sin datos de caudal (lo que se llama regionalización de parámetros).

A medio plazo, se quiere explorar la aplicabilidad del *parameter learning* en la calibración de LISFLOOD-OS de cara a su implementación en EFAS y GloFAS. Puesto que LISFLOOD-OS no está escrito en un lenguaje diferenciable, es necesario, como paso previo, crear un modelo subrogado de LISFLOOD. El objetivo de este trabajo es crear una red neuronal recurrente que emule el funcionamiento del modelo hidrológico LISFLOOD-OS. Como datos de partida se dispone de los mapas de características físicas de GloFAS [17], series históricas de caudal y meteorología [18], y los resultados de la simulación de LISFLOOD-OS para dicho periodo histórico. De cara a limitar la carga temporal en la preparación de los datos, se propone limitar el dominio espacial de este trabajo a España. Además, como comparativa del rendimiento entre un modelo híbrido y uno puramente basado en datos, se plantea entrenar un LSTM capaz de simular las series de caudal observado en base a las características físicas de las subcuencas de estudio.

Objetivos

Este trabajo plantea los siguientes objetivos:

- Crear el conjunto de datos necesario para el entrenamiento y validación de las dos redes LSTM. Esto supone la selección de subcuencas de estudio, trazado de las subcuencas y agregación espacial a escala subcuenca los datos distribuidos (en una malla regular) de partida: mapas de atributos de las cuencas, mapas con series temporales de precipitación y temperatura.
  Se planteará la posibilidad de publicar este conjunto de datos dentro de la comunidad hidrológica CARAVAN [19] bajo el nombre CAMELS-ESP (*Catchment Attributes and Meteorology for Large-sample Studies - España*). CAMELS es una iniciativa aparecida en EEUU [20] para generar conjuntos de datos hidrológicos a gran escala para la comparación de modelos globales.
- Desarrollo de una red neuronal que emule el funcionamiento del modelo hidrológico físicamente basado LISFLOOD-OS. Este modelo será seguramente una red neuronal recurrente, más en concreto una LSTM.
- Desarrollo de una red neuronal capaz de simular las series históricas de caudal en las subcuencas de estudio. A diferencia del punto anterior, esta red neuronal no estará condicionada a emular el funcionamiento de LISFLOOD-OS, sino que se pretende estudiar el potencial de un modelo hidrológico puramente basado en datos.

Alcance de los datos y las técnicas a utilizar

Datos

Para el desarrollo de este trabajo se utilizarán las siguientes fuentes de datos:

- **Mapas de atributos y de parámetros** del sistema GloFAS [17]. Estos mapas se denominan estáticos por considerarse invariables en el tiempo; contienen atributos geográficos (altitud, pendiente, uso del suelo, propiedades del suelo, etc.) y los mapas de los parámetros calibrados del modelo LISFLOOD-OS. Son datos públicos desde el mes pasado en el catálogo del Joint Research Centre.
- **Mapas con series meteorológicas históricas** de precipitación y temperatura. Se utilizarán bien los datos de reanálisis del modelo ECMWF-ERA5 o el conjunto de datos EMO (*European Meteorological Observations*) [18]. Ambos datos son accesibles, respectivamente, a través del servicio *Climate Data Store* de Copernicus o desde el repositorio de la Comisión Europea.
- Para el entrenamiento del modelo subrogado de LISFLOOD-OS la variable objetivo son las **series temporales de caudal simulado por LISFLOOD-OS** en el periodo de análisis y en los puntos de cálculo que se definan. Estas series temporales de caudal se pueden descargar del *Climate Data Store* de Copernicus.
- El entrenamiento del modelo LSTM requiere las **series temporales de caudal observado** en el periodo de análisis y en los puntos de cálculo que se definan. Las series a resolución diaria están disponibles en el Anuario de Aforos [21].
- El trazado de las cuencas hidrográficas de cada una de las estaciones de estudio requiere la aplicación de técnicas GIS sobre un **modelo digital del terreno** (MDT).

Estos modelos, a distintas resoluciones espaciales, se pueden descargar del Centro Nacional de Información Geográfica.

Técnicas

El primer bloque de trabajo será la creación del conjunto de datos necesario para el posterior entrenamiento de las redes LSTM. Se partirá del conjunto de estaciones de aforo incluidas en el Anuario de Aforos, de las que se deberá hacer una selección de las ubicaciones de estudio en base a criterios hidrológicos: longitud y calidad de la serie temporal de caudal, tamaño de la cuenca hidrográfica, presencia de alteraciones del régimen natural (embalses, trasvases…). Una vez definidas las estaciones de estudio, se ha de generar una capa GIS de polígonos con la cuenca hidrográfica de cada una de ellas; para ello se utiliza el MDT y las coordenadas de las estaciones. Los polígonos se utilizarán para agregar a nivel de subcuenca hidrográfica los *datasets* espacialmente distribuidos: los mapas de atributos y parámetros, y los mapas de las series meteorológicas. El resultado de este bloque es susceptible de ser publicado bajo el nombre de CAMELS-ESP dentro de la iniciativa CARAVAN, que está generando una comunidad de datos hidrológicos a gran escala en la que poder entrenar y comparar modelos hidrológicos globales. Hasta la fecha España no ha sido incluida en ninguno de estos estudios por desconocerse la disponibilidad en abierto de series diarias de caudal observado. Su inclusión favorecerá la aparición de futuros estudios centrados en España o estudios globales que incluyan a España.

La segunda fase del trabajo será la creación de un LSTM que emule las simulaciones del modelo hidrológico LISFLOOD-OS (Figura 1). Los datos de entrada serán las series meteorológicas, los mapas de atributos y los mapas de parámetros de LISFLOOD-OS, todo ello agregado a nivel de subcuenca. La salida del LSTM es la serie de caudal en cada una de las estaciones de estudio. El modelo se calibrará frente al caudal simulado por LISFLOOD-OS, con lo que aprenderá a reproducir lo que hace este modelo hidrológico, no la realidad. La generación de modelos subrogados de un modelo hidrológico no es una técnica innovadora (se ha aplicado a otros modelos como VIC o HBV), pero nadie ha creado un modelo subrogado de LISFLOOD-OS.

La tercera fase del trabajo será la creación de un LSTM capaz de simular el caudal observado en las estaciones de aforo de España. Dos son las diferencias con el modelo subrogado de LISFLOOD-OS (Figura 1). Por un lado, entre los datos de entrada de la red neuronal no se incluirán los parámetros calibrados de LISFLOOD-OS. Por otro lado, la red se calibrará frente a las series observadas de caudal, no frente a las series simuladas por el modelo hidrológico. El uso de LSTM en la simulación del caudal es una técnica que está cobrando relevancia en el campo de la hidrología y ya aplicada en múltiples estudios; sin embargo, que se sepa, no se ha aplicado a escala nacional sobre las estaciones de aforo de España.

En el entrenamiento de ambas redes neuronales se utilizará como variable objetivo el coeficiente de eficiencia de Kling-Gupta, KGE, [22]. Esto permitirá comparar el rendimiento de la calibración original de LISFLOOD-OS con las dos redes entrenadas en este estudio.

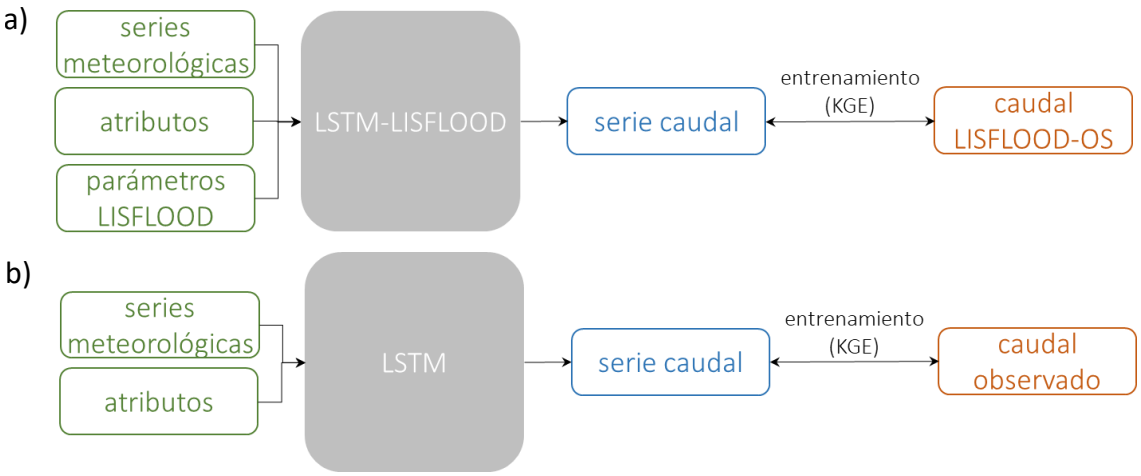

Figura 1. Esquema de las dos redes neuronales: a) modelo subrogado de LISFLOOD OS, b) modelo puramente basado en datos. En verde los datos de entrada del modelo, en azul los datos de salida, y en ocre los datos objetivo a partir de los que se entrenará la red.

Autoevaluación del grado de innovación.

El TFM debe ser original en su totalidad, pero el estudiante puede decidir los aspectos en los que su trabajo es innovador sobre el estado del arte. En la siguiente Tabla el estudiante debe autoevaluar su propuesta en cada uno de los aspectos. No es obligatorio que el TFM sea innovador en todos ellos, pero sí al menos en uno de ellos de manera significativa.

Aspecto	Descripción	Grado de innovación
Datos utilizados	El dataset/datasets es único y no ha sido analizado previamente, y tiene características que lo hacen más interesante que los que se pueden encontrar en estudios precedentes.	Se creará un nuevo *dataset* con datos hidrológicos de España.
Objetivos analíticos	Los objetivos del estudio no se han perseguido anteriormente, y aportan una utilidad, aplicación o perspectiva completamente novedosa.	Es la primera tentativa de generar un modelo subrogado de LISFLOOD-OS útil para la aplicación posterior de *parameter learning*.
Técnicas aplicadas para la preparación de datos	Las técnicas aplicadas a la preparación de los datos son diferentes a las utilizadas en estudios precedentes y aportan una riqueza adicional que puede llevar a mejores análisis o modelos. El impacto de esas técnicas innovadoras frente a su omisión debe formar parte explícita de la evaluación.	
Técnicas aplicadas analíticas	Las técnicas analíticas aplicadas son innovadoras respecto a los estudios existentes, y se incluye la comparación rigurosa con esas técnicas existentes como parte de la evaluación.	
Visualización avanzada	Se utilizan técnicas de visualización dinámicas que aportan una perspectiva no apreciable con técnicas más simples.	

Referencias

[1] Joint Research Centre - European Commission, «Open Source Lisflood», 2023.

[2] J. M. van der Knijff, J. Younis, y A. P. J. de Roo, «LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation», *International Journal of Geographical Information Science*, vol. 24, n.º 2, pp. 189-212, feb. 2010, doi: 10.1080/13658810802549154.

[3] F. A. Hirpa *et al.*, «Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data», *J Hydrol (Amst)*, vol. 566, pp. 595-606, nov. 2018, doi: 10.1016/j.jhydrol.2018.09.052.

[4] S. Grimaldi *et al.*, «GloFAS v4.0: towards hyper-resolution hydrological modelling at global scale», en *European Geoscience Union General Assembly 2023*, Vienna, 2023.

[5] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, y M. Herrnegger, «Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks», *Hydrol Earth Syst Sci*, vol. 22, n.º 11, pp. 6005-6022, nov. 2018, doi: 10.5194/hess-22-6005-2018.

[6] S. Nevo *et al.*, «Flood forecasting with machine learning models in an operational framework», *Hydrol Earth Syst Sci*, vol. 26, n.º 15, pp. 4013-4032, ago. 2022, doi: 10.5194/hess-26-4013-2022.

[7] F. Kratzert, D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, y G. S. Nearing, «Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning», *Water Resour Res*, vol. 55, n.º 12, pp. 11344-11354, dic. 2019, doi: 10.1029/2019WR026065.

[8] D. Feng, K. Fang, y C. Shen, «Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales», *Water Resour Res*, vol. 56, n.º 9, sep. 2020, doi: 10.1029/2019WR026793.

[9] T. Lees *et al.*, «Hydrological concept formation inside long short-term memory (LSTM) networks», *Hydrol Earth Syst Sci*, vol. 26, n.º 12, pp. 3079-3101, 2022, doi: 10.5194/hess-26-3079-2022.

[10] F. Kratzert, D. Klotz, M. Herrnegger, y S. Hochreiter, «A glimpse into the Unobserved: Runoff simulation for ungauged catchments with LSTMs».

[11] K. Yokoo *et al.*, «Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM», *Science of the Total Environment*, vol. 802, ene. 2022, doi: 10.1016/j.scitotenv.2021.149876.

[12] J. Koch y R. Schneider, «Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark», *GEUS Bulletin*, vol. 49, pp. 1-7, 2022, doi: 10.34194/geusb.v49.8292.

[13] M. Gauch, F. Kratzert, D. Klotz, G. Nearing, J. Lin, y S. Hochreiter, «Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network», *Hydrol Earth Syst Sci*, vol. 25, n.º 4, pp. 2045-2062, abr. 2021, doi: 10.5194/hess-25-2045-2021.

[14] G. Mao *et al.*, «Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation», *Physics and Chemistry of the Earth*, vol. 123, oct. 2021, doi: 10.1016/j.pce.2021.103026.

[15] L. J. Slater *et al.*, «Hybrid forecasting: blending climate predictions with AI models», *Hydrol Earth Syst Sci*, vol. 27, n.º 9, pp. 1865-1889, may 2023, doi: 10.5194/hess-27-1865-2023.

[16] W. P. Tsai *et al.*, «From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling», *Nat Commun*, vol. 12, n.º 1, dic. 2021, doi: 10.1038/s41467-021-26107-z.

[17] P. Salamon, S. Grimaldi, J. Disperati, y C. Prudhomme, «LISFLOOD static and parameter maps for GloFAS», 2023.

[18] V. Thiemig *et al.*, «EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe», *Earth Syst Sci Data*, vol. 14, n.º 7, pp. 3249-3272, jul. 2022, doi: 10.5194/essd-14-3249-2022.

[19] F. Kratzert *et al.*, «Caravan - A global community dataset for large-sample hydrology», *Sci Data*, vol. 10, n.º 1, dic. 2023, doi: 10.1038/s41597-023-01975-w.

[20] N. Addor, A. J. Newman, N. Mizukami, y M. P. Clark, «The CAMELS data set: Catchment attributes and meteorology for large-sample studies», *Hydrol Earth Syst Sci*, vol. 21, n.º 10, pp. 5293-5313, oct. 2017, doi: 10.5194/hess-21-5293-2017.

[21] CEDEX, «Anuario de aforos 2018-2019», nov. 2021.

[22] H. V. Gupta, H. Kling, K. K. Yilmaz, y G. F. Martinez, «Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling», *J Hydrol (Amst)*, vol. 377, n.º 1-2, pp. 80-91, oct. 2009, doi: 10.1016/J.JHYDROL.2009.08.003.