SISTEMAS CIBER-FÍSICOS PARA LA GESTIÓN DE LOS RECURSOS HÍDRICOS

Federico Vilaseca

Tutor
Alberto Castro

Co-tutora
Angela Gorgoglione

Índice

1. Introducción

El Sistema Cíber-Físico estudiado en este trabajo es el vinculado a la gestión de los recursos hídricos. El mismo está desarrollado en torno al ciclo natural del agua, e involucra a los elementos de medición de sus distintas componentes, así como también a las herramientas técnicas y organismos de decisión cuyo fin es la gestión sustentable del recurso. En la Figura 1, se representa el mismo en un esquema, que está basado en el de Wang et al. (2015).

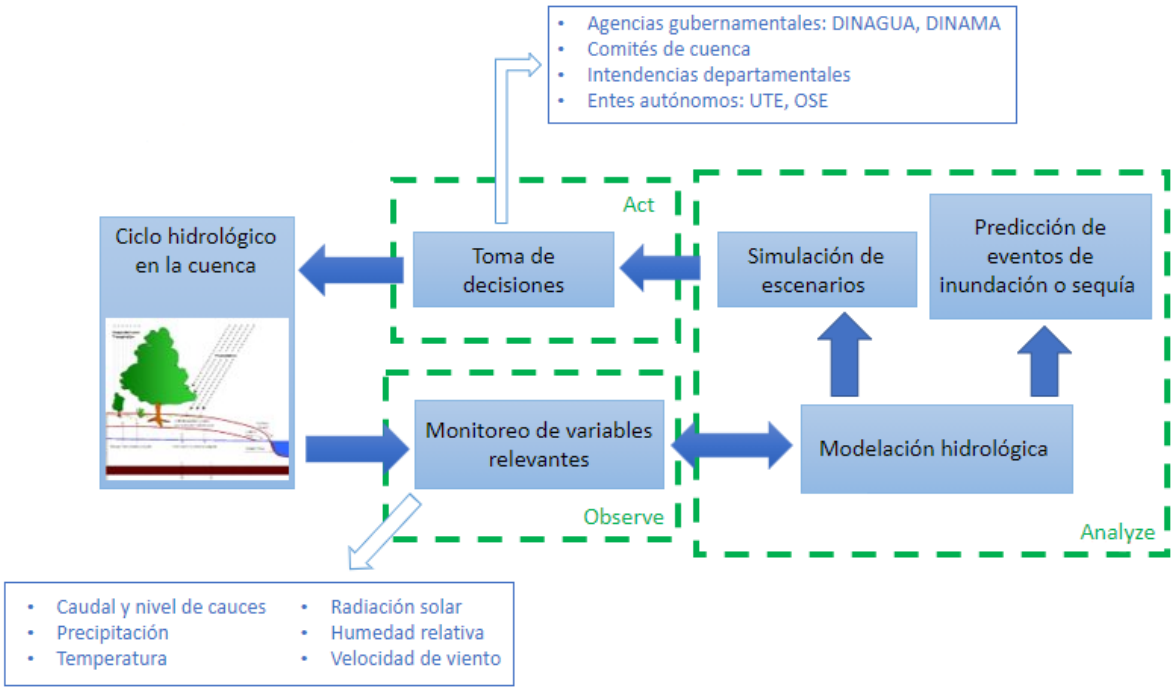

Figura 1.- Esquema del Sistema Ciber-Físico de la gestión de recursos hídricos.

El sistema se subdivide en tres componentes: de observación ("observe"), de análisis ("analyze") y de acción ("act"). La componente de observación incluye el monitoreo de variables relevantes para el estudio del ciclo hidrológico. Estas incluyen la cuantificación de volúmenes o flujos de agua en sus diferentes estados, así como también de otras variables que brindan información sobre los mismos. Como ejemplo se tienen: el caudal y nivel en cauces; variables atmosféricas como temperatura, humedad o radiación solar; la precipitación y el contenido de humedad en el suelo. A menudo estas mediciones se realizan mediante sensores conectados a estaciones telemétricas, que envían los datos registrados a servidores comunes y permiten el monitoreo en tiempo real.

La componente de análisis hace referencia al procesamiento de dichos datos, con el fin de traducirlos en información relevante para la gestión del agua. Es habitual que en esta etapa se implementen modelos de simulación o pronóstico, como herramientas de soporte para la predicción de escenarios o para sistemas de alerta.

La componente final involucra la toma de decisiones y la acción directa sobre el sistema. El proceso de toma de decisiones se apoya en los resultados obtenidos en la etapa de análisis, y lleva al diseño de estrategias de intervención sobre el ciclo hidrológico. Estas intervenciones pueden ser, por ejemplo: toma de agua de fuentes superficiales o subterráneas, implantación de obras hidráulicas, cambios en el uso y manejo de suelos, acciones territoriales para prevención de inundaciones, etc. Por lo general, en esta etapa actúan agencias gubernamentales o empresas de servicios asociados al recurso hídrico.

En particular, este trabajo se enmarca en la etapa de análisis e involucra la modelación del proceso precipitación-escorrentía. Este es un subproceso del ciclo hidrológico, que abarca desde la precipitación sobre el continente hasta la salida en forma de flujo superficial, pasando por etapas intermedias que tienen lugar en los distintos estratos del suelo. En la Figura 2, se presenta un esquema conceptual del ciclo hidrológico en su conjunto, mientras que en la Figura 3.- Esquema conceptual del proceso precipitación-escorrentía. Extraído de Neitsch et al. (2011)., se puede ver específicamente el proceso precipitación-escorrentía.

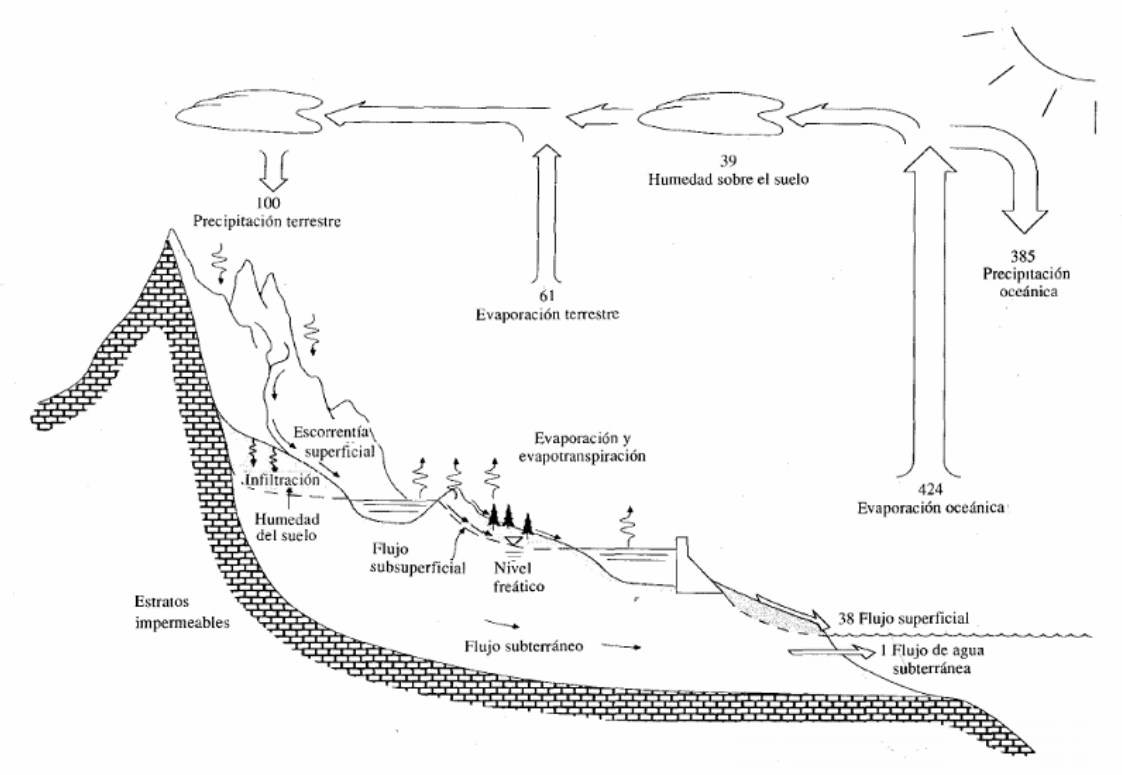

Figura 2.- Esquema conceptual del ciclo hidrológico. Extraído de Chow et al. (1994).

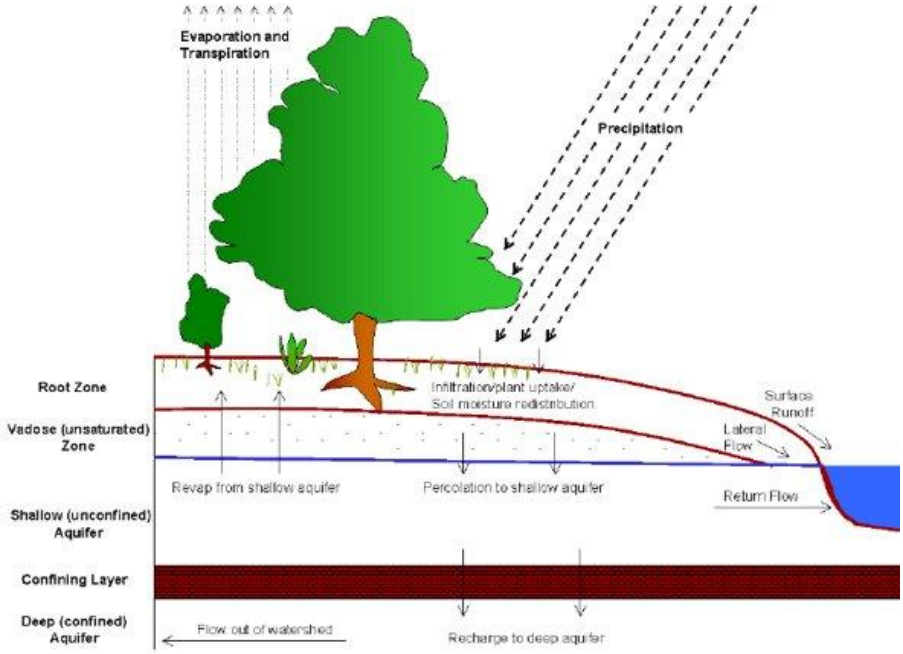

Figura 3.- Esquema conceptual del proceso precipitación-escorrentía. Extraído de Neitsch et al. (2011).

En ambos esquemas es posible observar que, entre la precipitación y la escorrentía se dan varios procesos, entre los que cabe mencionar: evaporación, transpiración (por parte de plantas), infiltración, percolación, flujo lateral (subsuperficial), flujo de retorno (subterráneo), flujo superficial, etc. Debido a la complejidad de cada uno (no linealidad, aleatoriedad y no estacionariedad), la modelación de estos procesos no resulta sencilla y es un tema relevante de investigación. Existen actualmente dos enfoques para la modelación del proceso precipitación-escorrentía: la modelación de base física y la modelación basada en datos.

La modelación basada en procesos físicos utiliza modelos matemáticos para representar las distintas etapas del proceso y cuantificar los flujos y volúmenes retenidos en etapas intermedias del mismo. Es la herramienta que se utiliza tradicionalmente en modelación hidrológica. A lo largo de los años se han desarrollado un gran número de modelos de este tipo, con diferentes niveles de complejidad. Para su implementación a un caso de estudio particular se suele seguir un procedimiento que consta de dos etapas: calibración y validación. En la primera se realiza un ajuste de los parámetros del modelo matemático, de modo que las series temporales de salida del modelo se ajusten lo mejor posibles a los datos observados, para un período de tiempo determinado. En la segunda se verifica que el modelo, con los parámetros ajustados, se comporta adecuadamente para otros períodos de tiempo. De esa forma se asegura su funcionalidad para simulación de nuevos escenarios.

La modelación basada en datos es de desarrollo más reciente, y fue posibilitada por los avances realizados en las últimas décadas en el área de la computación. Se basa en algoritmos capaces de detectar y modelar correlaciones complejas entre las variables de entrada y de salida, sin necesariamente representar los procesos físicos involucrados. Su implementación requiere de una etapa adicional, que es el entrenamiento del modelo, durante la cual el algoritmo “aprende” el comportamiento de los datos de entrada y los vincula a los de salida.

Ambos tipos de modelos poseen sus ventajas y desventajas respecto al otro. Los de base física suelen ser más robustos a la hora de simular escenarios fuera del período de calibración, ya que son capaces de caracterizar los procesos en cada caso de estudio. En cambio, tienen la desventaja de que no

siempre logran representarlos en toda su complejidad, ya que los modelos matemáticos que se utilizan suelen ser relativamente simples. Los basados en datos suelen tener mayor capacidad para ajustar los datos observados y presentan ventajas desde el punto de vista del tiempo de cálculo, pero están limitados por la cantidad de datos disponibles para su implementación. Además, se suele cuestionar su aplicabilidad fuera del período de datos para el cual fueron entrenados, por no representar los procesos físicos involucrados.

Resulta claro que la aplicabilidad de ambos tipos de modelos es complementaria, y depende en gran medida del caso de estudio. Los avances en modelación basada en datos deben estar enfocados en investigar su capacidad para representar los procesos físicos más relevantes dentro de su estructura de aprendizaje. En este marco, este trabajo se plantea los siguientes objetivos:

1. Implementación de un modelo basado en procesos físicos a un caso de estudio de relevancia nacional.
2. Implementación de un modelo basado en datos al mismo caso de estudio.
3. Comparación de los resultados de ambos modelos, con foco en la representación de procesos físicos.

2. Revisión bibliográfica

En el marco de este trabajo se llevó a cabo una revisión bibliográfica con el fin de conocer las tendencias actuales en cuanto a modelación basada en datos del proceso precipitación-escorrentía, la interpretación física de sus procesos de aprendizaje y resultados, y su comparación con modelos de base física. Existe una numerosa cantidad de trabajos publicados en los últimos años en los que se utilizan técnicas de aprendizaje automático para el entrenamiento de modelos hidrológicos. Entre ellos hay modelos de paso mensual (Shortridge et al., 2016; White, 2017), diario (Fu et al., 2020; Kan et al., 2017) y horario (Yaseen et al., 2020) aplicados a cuencas individuales (Kratzert, Herrnegger, et al., 2019; Tongal & Booij, 2018) o a estudios regionales (Kratzert et al., 2018; Shortridge et al., 2016; White, 2017) en los cuales son entrenados para pronosticar salidas de varias cuencas en simultáneo.

Dentro de las técnicas de aprendizaje automático utilizadas es muy común encontrar aplicaciones en torno a Redes Neuronales Artificiales y Random Forest, incluyendo de comparación de estos con otras técnicas y de combinación de distintos algoritmos (Kan et al., 2017; Kim et al., 2019; Shortridge et al., 2016; Tongal & Booij, 2018; White, 2017). En particular, toman relevancia las diferentes variantes de Redes Neuronales (Kratzert et al., 2019; Yaseen et al., 2018). De entre ellas, la que parece ser más adecuada para la simulación de series temporales es la Red Neuronal Recurrente del tipo Long Short-Term Memory (LSTM) por su capacidad de incorporar los datos antecedentes dentro del proceso de aprendizaje, dotando a los modelos de una noción nativa de temporalidad (Fu et al., 2020; Kratzert et al., 2018).

Si bien hay varios estudios cuyo fin es evaluar el desempeño de nuevas técnicas de aprendizaje automático para simular procesos hidrológicos (Fu et al., 2020; Yaseen et al., 2018), los que resultan de mayor interés para los objetivos de este trabajo son aquellos que le dan importancia a la interpretación física de los resultados y procesos de aprendizaje de dichos modelos. Entre ellos se destacan los trabajos de: Kratzert et al. (2019) donde se realizan interpretaciones físicas del proceso interno de aprendizaje de las LSTM analizando la respuesta de los nodos internos ante distintos escenarios de simulación; y el de (Tongal & Booij, 2018) en el cual se estudia la relevancia de la separación de flujo base en modelación basada en datos y se discute sobre la efectividad de incorporar modelos híbridos con este fin.

Otro punto de relevancia es la selección de las variables de input de los modelos. En ese sentido se tienen casos como los estudios de Yaseen et al. (2018) y Fu et al. (2020), en los cuales se utilizan únicamente series de caudal antecedente con desfasaje temporal, que no parecen tener aplicación fuera de la regresión y no permiten la interpretación de procesos físicos; y otros como los de White (2017) y Tongal & Booij (2018) en los cuales se plantea una discusión sobre la relevancia de incorporar otras variables climáticas y el desfasaje que se debe incorporar a las mismas para representar adecuadamente la temporalidad de los procesos. Dentro de las variables incorporadas en estos trabajos se encuentran la precipitación, nieve, temperatura y evapotranspiración potencial. En los casos en que se estudian modelos regionales, toman relevancia además características físicas de la cuenca (Kratzert et al., 2018; Shortridge et al., 2016; White, 2017).

A efectos de la implementación en este trabajo se destaca la aplicación de Random Forest (Brieman, 2001) para modelación hidrológica, debido a la facilidad que presenta para interpretación de sus resultados, además de su efectividad para problemas de regresión. En cuanto a la selección de variables de entrada, se valora la importancia de incorporar variables climáticas al análisis y de considerar las mismas con un desfasaje adecuado, acorde a las características físicas del caso de estudio.

3. Zona de estudio

La cuenca seleccionada para el estudio es la del río Santa Lucía Chico, con cierre en la represa de Paso Severino. La misma está comprendida mayormente dentro del departamento de Florida, cubriendo un área en planta de 2478 km^2. Es una cuenca de relevancia a nivel nacional debido a la intensa actividad agrícola que se lleva a cabo en su territorio. Además, es una de las fuentes que alimenta el sistema de potabilización de Aguas Corrientes, mediante la reserva de agua en Paso Severino. En la Figura 4, se puede observar su ubicación en Uruguay.

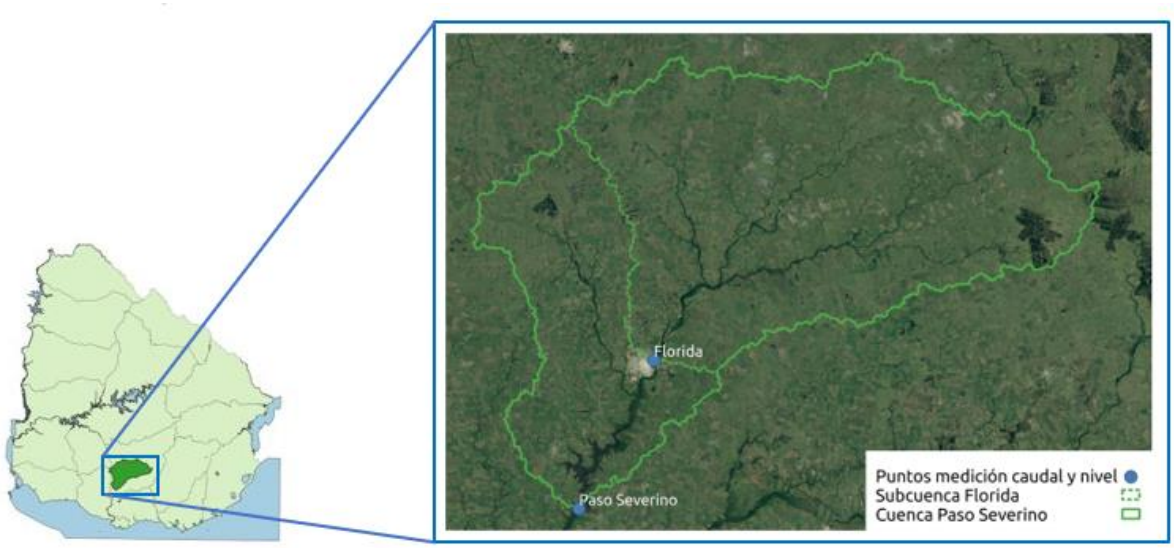

Figura 4.- Ubicación en Uruguay de la cuenca del río Santa Lucía Chico, con cierre en Paso Severino.

Dentro de la misma se lleva a cabo el monitoreo sistemático de variables vinculadas al ciclo hidrológico. Entre ellas, se destaca la medición de nivel de la superficie libre y caudal del río en dos sitios: el puente de Ruta 5 en la ciudad de Florida (estación hidrométrica 53.1 de DINAGUA) y en el vertedero de la represa de Paso Severino (administrada por OSE). La ubicación de ambas se muestra en la Figura 4. A su vez, en su entorno, se ubican varias estaciones pluviométricas de INUMET, que registran acumulados diarios de precipitación. Para este proyecto fueron consideradas 8 de ellas,

ubicadas en las localidades de: Cerro Colorado, Florida, La Cruz, Mendoza, Reboledo, San Gabriel, Sarandí Grande y Villa 25 de Agosto. Su ubicación respecto a la cuenca puede observarse en la Figura 5.

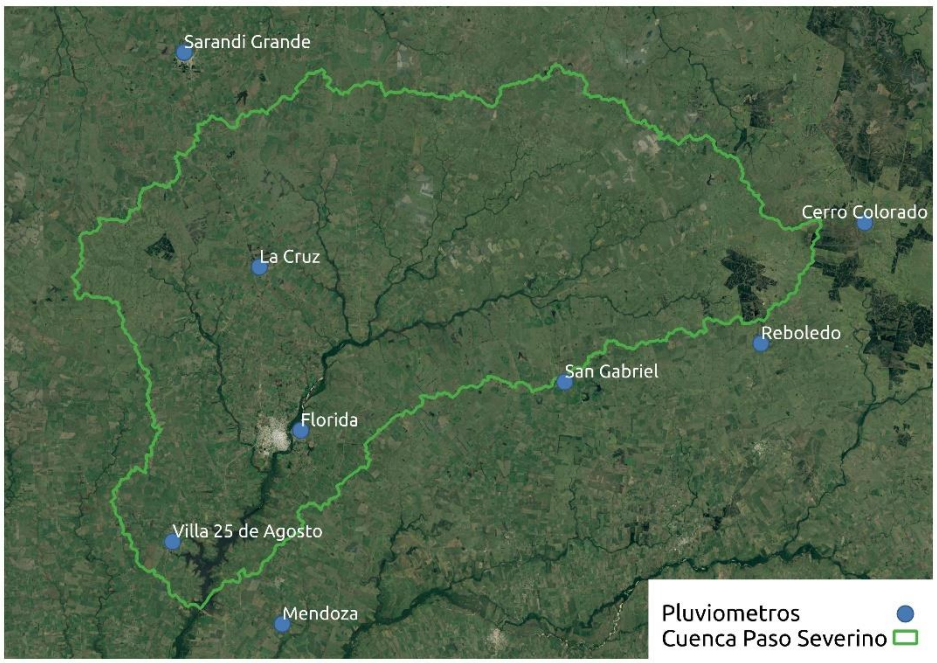

Figura 5.- Ubicación de estaciones pluviométricas respecto a la cuenca.

4. Descripción de los datos

A continuación, se describen brevemente las variables que se tuvieron en cuenta para el estudio y se detallan las características de las series temporales obtenidas.

4.1 Caudal y nivel del río

Como se menciona anteriormente, se obtuvieron series temporales de caudal promedio diario en dos puntos del cauce: Paso Severino (vertedero de la presa) y Florida (estación de aforo 53.1 de DINAGUA, ubicada sobre el puente de Ruta 5). El caudal o descarga del río ("Q") es una de las variables más relevantes para este proyecto, ya que es la salida principal del proceso precipitación-escorrentía, que se pretende simular. En ambos casos, lo que se mide sistemáticamente es el nivel de la superficie del agua en la sección ("H"), y el caudal se infiere a partir de una función matemática. Esto se debe a que la medición de caudal en forma continua resulta difícil y económicamente costosa, mientras que el nivel suele ser más sencillo y barato de medir. Por lo tanto, junto con las series de caudal, se tienen las series de nivel correspondientes, que también forman parte del set de datos del proyecto.

Existen diversas formas de construir las funciones que vinculan nivel y caudal, conocidas habitualmente como "Curvas H-Q". A continuación, se describen dos métodos, los cuales resultan de interés para este trabajo ya que de ellos provienen los datos de caudal utilizados. Estos son: *leyes de vertido* de estructuras hidráulicas y *curvas de aforo*.

Una *sección de control* es aquella donde existe una relación biunívoca entre el caudal y el tirante del flujo. Es habitual generarlas artificialmente mediante estructuras de vertido con diversos fines, entre ellos, la medición de caudal. La función matemática que representa dicha relación biunívoca se

denomina “Ley de vertido”. En el caso del vertedero de una represa, suele representarse mediante la siguiente ecuación:

$$Q = C\ H^{3/2}$$

El coeficiente *C*, depende de la geometría particular de cada vertedero. Utilizando esta función, se puede inferir el caudal a partir del nivel aguas arriba de la estructura.

En secciones naturales sin estructuras de control es habitual el uso de curvas de aforo para la estimación del caudal. Estas funciones se ajustan a partir de un número finito de mediciones (simultáneas) de nivel y caudal. Dicho ajuste suele realizarse mediante una función potencial del tipo:

$$Q = A\ (H - H_0)^b$$

Cabe destacar que normalmente el ajuste no consiste en una única función, sino que se realiza por tramos de niveles, de modo de tener en cuenta la variación en vertical de la geometría en la sección transversal. Además, la validez de estas curvas está condicionada por los cambios en la geometría de la sección que ocurren a través de los años, por lo que las mismas se actualizan periódicamente.

A continuación, se describen las dos series temporales de caudal utilizadas.

4.1.1 Florida

En Uruguay, se cuenta con una gran cantidad de puntos de monitoreo con curva de aforo, las cuales son generadas y mantenidas por la Dirección Nacional de Agua (DINAGUA), cubriendo los ríos y arroyos más importantes del país. Para mayor practicidad en la medición, estos suelen encontrarse en secciones atravesadas por puentes. En la mayor parte de los puntos de monitoreo, se cuenta con estaciones telemétricas de medición de nivel, las cuales generan un registro cada 30 minutos que se actualiza cada 6 horas y se encuentra disponible online. Para acceder a información de caudales, es necesario realizar un pedido de información a DINAGUA o contar con los parámetros de la curva de aforo en la sección de interés.

Dentro de la cuenca de estudio, se encuentra la estación 53.1 de DINAGUA, asociada a la sección del río Santa Lucía Chico atravesada por el puente de Ruta 5. Se realizó una solicitud para obtener niveles y caudales medios diarios en dicha estación para el período comprendido entre el 1/1/1980 hasta el 8/6/2018.

A su vez se cuenta con los parámetros de la curva de aforo actualmente asociada a dicha sección, por lo que se descargó de la web la serie de niveles entre el 9/6/2018 y el 30/6/2020, y se estimaron los caudales correspondientes a ese período usando la curva. Cabe destacar que esta serie complementaria tiene datos cada 30 minutos.

4.1.2 Paso Severino

Sobre el punto de cierre de la cuenca de Santa Lucía Chico se ubica la represa de Paso Severino, construida en la década de 1980 con el fin de generar un reservorio de agua para abastecer de agua potable a Montevideo y su área metropolitana. La misma es administrada por las Obras Sanitarias del Estado (OSE).

En represas como la de Paso Severino, cuyo fin es la reserva de agua para potabilización, es habitual encontrar obras de toma, que consisten en tuberías que conectan el lago con el cauce aguas abajo de la estructura. Éstas cuentan con compuertas que se abren y cierran en función de la necesidad de agua que haya en la planta de tratamiento aguas abajo, en nuestro caso la de Aguas Corrientes. Por lo que

el total del caudal que atraviesa la presa surge de la suma del que cae por el vertedero y el que atraviesa la toma.

Se realizó una solicitud de información a OSE, mediante la cual se pudo obtener la serie de caudales medios diarios erogados por la presa desde el 1/1/1990 (puesta en funcionamiento de la represa) hasta el 3/5/2016. Se tiene registro por separado de la componente de caudal que vierte por encima de la presa, así como también de la que se extrae mediante la obra de toma. A su vez se cuenta con el promedio diario del nivel de la superficie libre del agua en el lago inmediatamente aguas arriba de la misma, para el período mencionado.

4.2 Precipitación

La precipitación es el input fundamental del proceso que se quiere simular, y por lo tanto requerido en ambos modelos. En Uruguay, el organismo encargado del registro de información pluviométrica es el Instituto Uruguayo de Meteorología (INUMET), cuya red de estaciones de medición cubre todo el territorio nacional. El instrumento habitualmente utilizado para estas mediciones es el pluviómetro, el cual consiste en un recipiente que capta el agua de la precipitación y registra el volumen que ingresa al mismo en función del tiempo.

Para acceder a información pluviométrica, se debe realizar una solicitud de información a INUMET. En este caso, se solicitó el registro de acumulados diarios de precipitación (mm/día) en 8 estaciones ubicadas en torno a la cuenca, cubriendo el período entre el 1/1/1980 y el 30/6/2020. Las estaciones seleccionadas fueron: Cerro Colorado, Florida, La Cruz, Mendoza, Reboledo, San Gabriel, Sarandí Grande y Villa 25 de Mayo. El contar con datos de varias estaciones cercanas a la cuenca permite tener en cuenta la variabilidad espacial de la precipitación, la cual adquiere importancia en el proceso de generación de escorrentía en cuencas extensas como la que se está estudiando.

Cabe destacar que, además de INUMET, otra institución que lleva registros de información pluviométrica es el Instituto Nacional de Investigación Agropecuaria (INIA). En este caso se utilizó únicamente la de INUMET, ya que ninguna de las 5 estaciones de INIA se encuentra lo suficientemente cerca de la cuenca como para aportar información adicional.

4.3 Variables climáticas

Dentro del proceso físico de precipitación-escorrentía que se intenta modelar, toma relevancia el proceso de evapotranspiración. El mismo resulta de la adición de dos procesos: el de evaporación del agua de lluvia retenida en la cuenca (en las capas superficiales del suelo, depresiones del terreno, vegetación, etc.) y el de transpiración de las plantas y árboles a través de sus hojas. La evapotranspiración genera la transferencia de agua desde la cuenca, en estado líquido, hacia la atmósfera, en estado de vapor. Este proceso depende en gran medida de variables atmosféricas como: temperatura, radiación solar, humedad relativa y velocidad del viento.

En Uruguay, estas variables son registradas por INIA en sus 5 estaciones, y se encuentran disponibles online. Para este estudio, se descargaron datos correspondientes a la estación más cercana a la cuenca, ubicada en la localidad de Las Brujas, departamento de Canelones cubriendo el período 1/1/1980 - 30/6/2020. Las variables seleccionadas fueron: humedad relativa media (%), radiación solar acumulada (cal/cm^2/día), temperatura máxima diaria (°C), temperatura mínima diaria (°C) y velocidad del viento media a 2m de altura (km/24hs). Si bien se podían haber incorporado más (o menos) variables al análisis, se seleccionaron estas por ser las requeridas por el modelo de base física seleccionado (modelo SWAT, Arnold et al., 1998, el cual se describe en la sección 6.1.1) para simular el proceso precipitación-escorrentía. Al modelo basado en datos se incorporaron las mismas, de forma de poder comparar ambos adecuadamente.

4.4 Resumen

En la Tabla 1 se presenta un resumen de las características de los datos que se recopilaron para su uso en ambas etapas de modelación.

Tabla 1.- Resumen descriptivo de las series temporales utilizadas.

Variable	Fuente	Ubicación estación	Unidad	Período	Frecuencia
Caudal	OSE	Paso Severino - vetedero	m^3/s	1/1/1990 - 3/5/2016	Promedio diario
Caudal	OSE	Paso Severino - toma	m^3/s	1/1/1990 - 3/5/2016	Promedio diario
Nivel	OSE	Paso Severino	m	1/1/1990 - 3/5/2016	Promedio diario
Caudal	DINAGUA	Florida R5, Est. 53.1	m^3/s	1/1/1980 - 8/6/2018	Promedio diario
Caudal	DINAGUA	Florida R5, Est. 53.1	m^3/s	9/6/2018 - 30/6/2020	30 minutos
Nivel	DINAGUA	Florida R5, Est. 53.1	m	1/1/1980 - 8/6/2018	Promedio diario
Nivel	DINAGUA	Florida R5, Est. 53.1	m	9/6/2018 - 30/6/2020	30 minutos
Precipitación	INUMET	Cerro Colorado	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	Florida	mm	1/5/1989 - 30/6/2020	Acumulado diario
Precipitación	INUMET	La Cruz	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	Mendoza	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	Reboledo	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	San Gabriel	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	Sarandí Grande	mm	1/1/1980 - 30/6/2020	Acumulado diario
Precipitación	INUMET	Villa 25 de Mayo	mm	1/1/1980 - 30/6/2020	Acumulado diario
Humedad relativa	INIA	Las Brujas	%	1/1/1980 - 30/6/2020	Promedio diario
Radiación solar	INIA	Las Brujas	$cal/cm^2/día$	1/1/1980 - 30/6/2020	Acumulado diario
Temperatura	INIA	Las Brujas	ºC	1/1/1980 - 30/6/2020	Mínimo diario
Temperatura	INIA	Las Brujas	ºC	1/1/1980 - 30/6/2020	Máximo diario
Velocidad de viento a 2 m de altura	INIA	Las Brujas	Km/24 hs	1/1/1980 - 30/6/2020	Promedio diario

5. Análisis de los datos

Para poder proceder a los trabajos de modelación se realizó un análisis de los datos, de modo de conocer sus características y determinar posibles anomalías o datos faltantes. En este apartado, se presentan los resultados de ese proceso de análisis. En primer lugar, se realizó un procesamiento previo de las series temporales. En particular, se uniformizó el paso temporal de las series de caudales, para tener valores promedio diarios y se ajustaron las unidades de las variables climáticas, para adecuarlas a los requerimientos del modelo SWAT. En segundo lugar, se llevó a cabo un análisis estadístico exploratorio de los datos. A partir de este análisis, no fueron detectadas anomalías que requirieran estudios en mayor detalle, por lo tanto, no se realizaron recortes ni modificaciones a los datos.

5.1 Preprocesamiento

Se promediaron los datos de nivel de superficie libre con frecuencia cada 15 minutos, de modo de obtener una serie de datos medios diarios correspondiente al período del 9/6/2018 hasta el 30/6/2020. De ese modo, se igualó el paso de tiempo con el de la serie del período 1/1/1980 - 8/6/2018. Luego, utilizando la curva de aforo, dichos niveles se convirtieron a caudales. Finalmente, se unieron todos los datos, para generar series de nivel y caudal que cubran el período entre el 1/1/1980 y el 30/6/2020.

Los valores de radiación solar, humedad relativa y velocidad de viento se ajustaron para adaptarlos a las unidades que utiliza el modelo SWAT: MJ/ m^2/día; fracción adimensional (en lugar de porcentaje) y m/s respectivamente.

5.2 Análisis exploratorio

El análisis exploratorio se llevó a cabo mediante la generación de varios gráficos descriptivos, los cuales se presentan y comentan a continuación. En la Tabla 2 se introduce la nomenclatura que se usa de aquí en adelante para las variables.

Tabla 2.- Nomenclatura de variables.

Nombre	Variable	Nombre	Variable
PP_CCol	Precipitación en Cerro Colorado	TMax	Temperatura máxima
PP_Flor	Precipitación en Florida	TMin	Temperatura mínima
PP_LCru	Precipitación en La Cruz	Wnd	Velocidad de viento
PP_Mend	Precipitación en Mendoza	ETP	Evapotranspiración potencial
PP_Rebo	Precipitación en Reboledo	Q	Caudal en Florida
PP_SGab	Precipitación en San Gabriel	H	Nivel en Florida
PP_SGra	Precipitación en Sarandí Grande	H_VERT	Nivel en P. Severino
PP_VMay	Precipitación en Villa 25 de Mayo	Q_VERT	Caudal de vertido en P. Severino
RHum	Humedad relativa	Q_TOMA	Caudal de toma en P. Severino
Slr	Radiación solar	Q_TOT	Caudal total en P. Severino

5.2.1 Distribución de valores

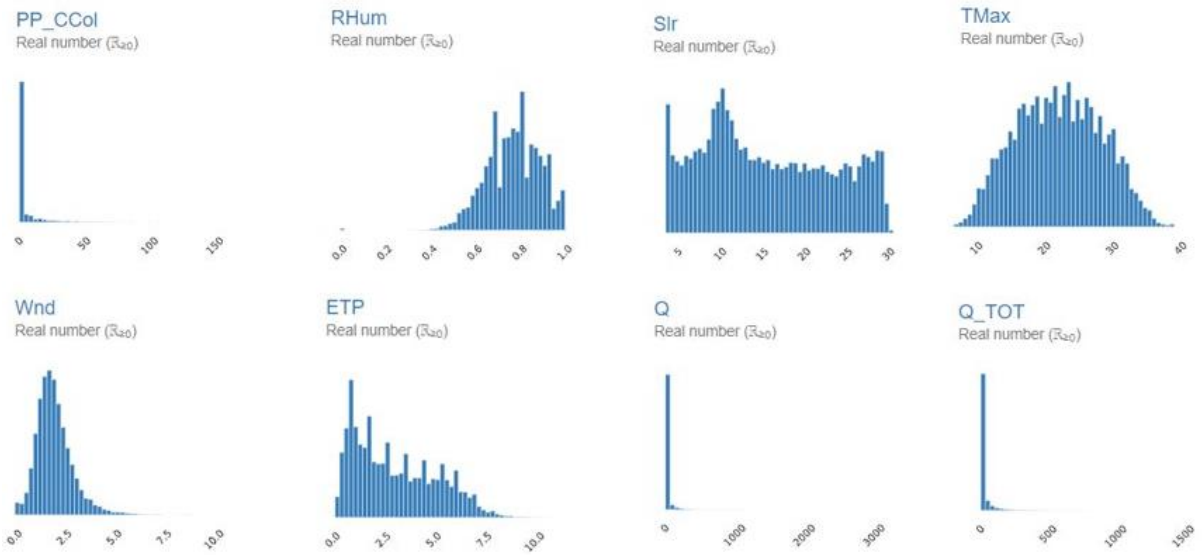

Figura 6.- Distribución de valores de las variables.

En los gráficos de la Figura 6, se puede observar la distribución de valores de las variables consideradas. Cabe mencionar que se incluye únicamente la precipitación en Cerro Colorado ("PP_CCol") como representante de las restantes, ya que los gráficos son similares en todos los casos. A su vez, se excluyen las series de nivel, por ser similares a las de caudal. Del mismo modo, se muestra solamente la temperatura máxima, por tener una distribución muy parecida a la de la mínima.

Se observa que tanto precipitaciones como caudales acumulan la mayor cantidad de valores en el primer cuantil. En el caso de las precipitaciones esto se debe a la alta cantidad de valores nulos dados por los días sin lluvia (del entorno del 70%). En los caudales se deben a la alta permanencia del flujo base, que representa valores bajos, cercanos al cero. A su vez, en ambos casos hay poca permanencia de caudales altos, correspondientes a eventos extremos de precipitación y crecida. El resto de las

variables presentan una distribución más uniforme, mostrando las particularidades propias del fenómeno que representan.

5.2.2 Matriz de correlación

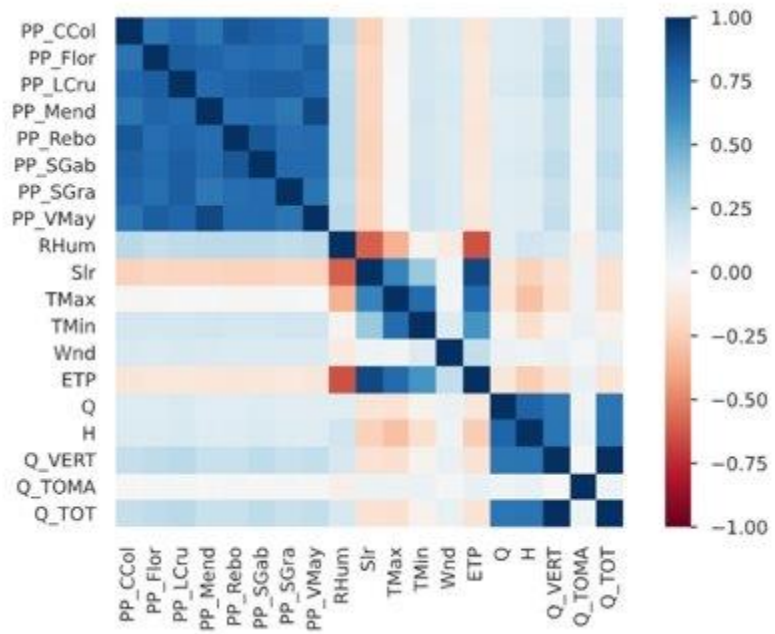

Figura 7.- Matriz de correlación entre las variables.

La matriz de correlación de la Figura 7 da una idea de cómo se interrelacionan entre sí las variables. Se puede observar que las precipitaciones están altamente correlacionadas entre sí, aunque esta correlación varía debido a la distribución espacial de la lluvia sobre la cuenca. Lo mismo se observa para los caudales y niveles, destacándose la alta correlación entre los registros de Florida y los de Paso Severino. Esto demuestra que el efecto de laminación de la presa no es significativo, ya que de otro modo produciría variaciones que alteren esta correlación. Por último, se puede apreciar también que las variables climáticas se encuentran correlacionadas, destacándose la alta correlación negativa entre algunas de ellas.

5.2.3 Datos faltantes

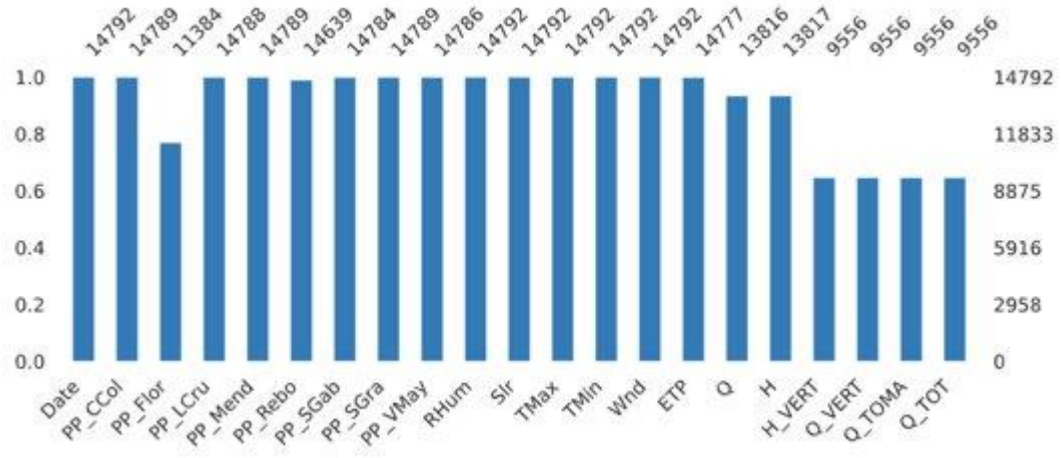

Figura 8.- Cantidad de datos válidos para cada variable.

En el período de estudio hay un total de 14792 días, por lo que, al estar trabajando con valores diarios, una serie temporal completa debería tener esa misma cantidad de datos. El gráfico de la Figura 8 muestra la cantidad de datos por variable. Se destaca la buena calidad en este sentido, ya que la mayor parte de las variables tienen menos del 10% de faltantes en el registro. Las excepciones son el pluviómetro de Florida y las series hidrométricas de Paso Severino. En ambos casos, esto se debe a que las series disponibles no cubren el total del período: en el caso de la precipitación en Florida, la serie abarca desde el 1/5/1989 al 30/6/2020, y, en Paso Severino, las series de nivel y caudal van desde el 1/1/1990 al 3/5/2016 (ver Tabla 1).

6. Metodología

6.1 Modelación basada en procesos físicos

6.1.1 Descripción del modelo

Se utilizó el modelo SWAT (Arnold et al., 1998) cuya aplicación es extendida en todo el mundo. Se trata de un modelo distribuido. Esto quiere decir que se subdivide la cuenca para la realización de los cálculos hidrológicos y luego se combinan los resultados de cada división para obtener el hidrograma de respuesta en el punto de cierre. Este esquema permite tener en cuenta la distribución espacial de las características físicas del terreno, así como también de las variables meteorológicas de input. En la Figura 9, se presenta un esquema de su funcionamiento.

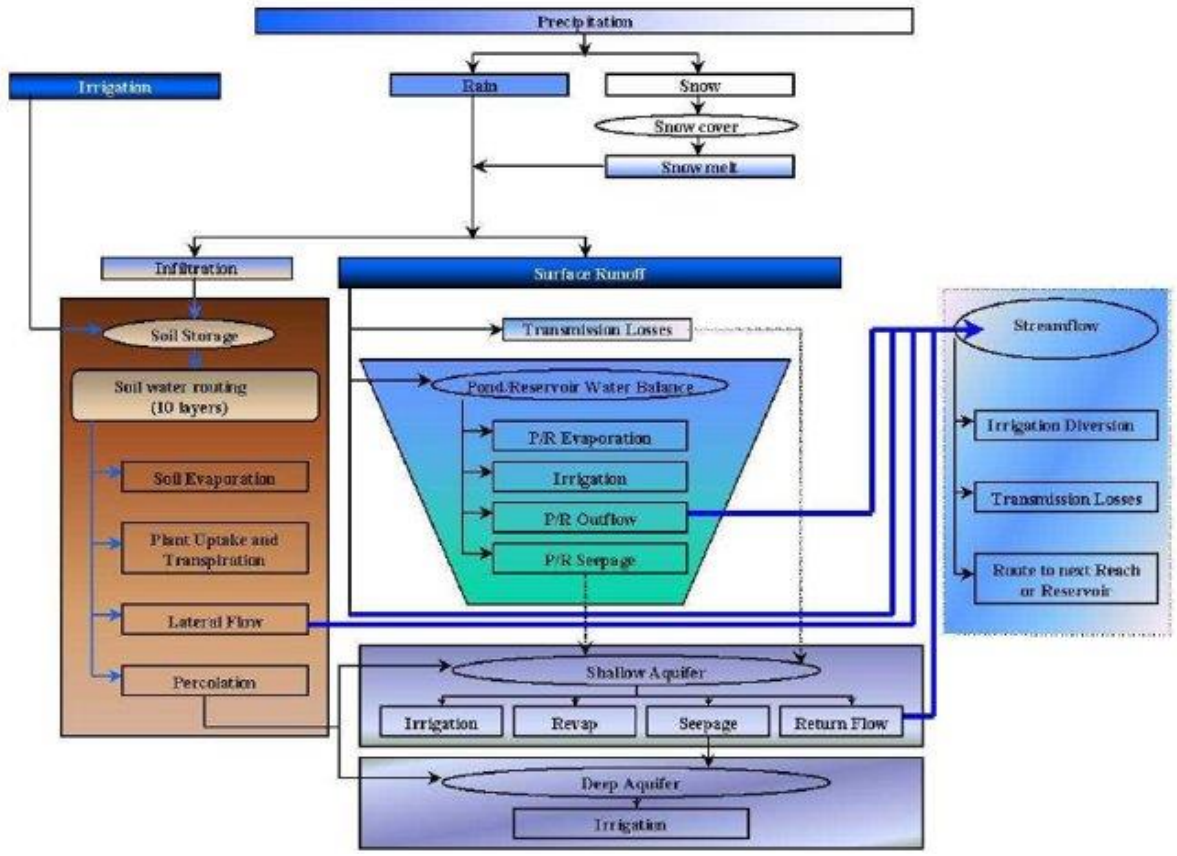

Figura 9.- Esquema conceptual del funcionamiento del modelo SWAT, extraído de Neitsch et al. (2011).

El modelo requiere como inputs series temporales de: precipitación, humedad relativa, radiación solar, temperatura máxima, temperatura mínima y velocidad de viento. También requiere información espacial de topografía, tipo de suelos y usos del suelo, las cuales utiliza para la subdivisión de la cuenca. Esta se realiza en dos niveles: la cuenca total se divide en subcuencas de acuerdo con la

hidrografía y luego cada subcuenca en varias unidades de respuesta hidrológica o HRU por sus siglas en inglés, que se obtienen intersectando tipos de suelo, usos del suelo y pendiente de la cuenca. Los cálculos que se observan en el esquema de la Figura 9 se realizan para cada una de estas HRU, luego se suman para cada subcuenca y por último se calcula el tránsito agregado del caudal a través de la cuenca, para determinar el hidrograma en su salida. En la Figura 10, se presenta esquemáticamente la información espacial utilizada y la división resultante en subcuencas y HRUs.

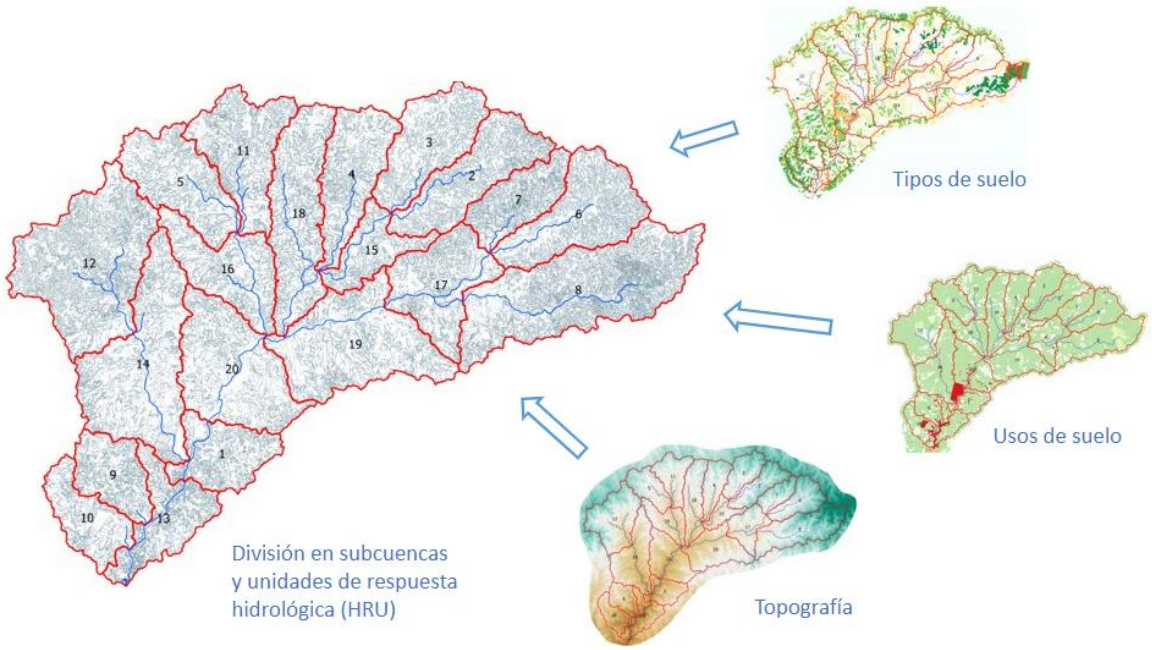

Figura 10.- Subdivisión espacial resultante durante la implementación de SWAT en la cuenca.

Además del trabajo con información espacial, la implementación de SWAT requiere la determinación inicial de los parámetros del modelo. Estos ascienden al orden de decenas, siendo algunos representantes de características generales de toda la cuenca, pero otros específicos de cada subcuenca o HRU. Este proceso de ajuste inicial resulta fundamental ya que el ingreso de valores correctos es la mejor forma para asegurar una representación adecuada de los procesos físicos que se dan en la cuenca. Además, este primer estado del modelo es la base sobre la cual se lleva a cabo el proceso de calibración, que se describe más adelante; si el mismo no es adecuado, la calibración no resulta efectiva.

6.1.2 Calibración

La calibración del modelo consiste en el ajuste de parámetros con el fin de asemejar lo más posible la serie temporal de caudales resultante a la de observados, para un período de tiempo dado. Este proceso puede realizarse manualmente, en base a experiencia del modelador, o con la ayuda de algoritmos de optimización. En el último caso, se requiere la definición de una función objetivo, cuyo valor se quiere optimizar. En modelación hidrológica es habitual el uso del coeficiente de Nash-Sutcliffe como función objetivo:

$$NS = 1 - \frac{\sum_i (Q_{mod} - Q_{obs})_i^2}{\sum_i \left(Q_{obs,i} - \overline{Q_{obs}}\right)^2} \qquad (Ecuación\ 1)$$

Siendo:

- NS : el coeficiente de Nash-Sutcliffe

- Q_{mod} : el caudal resultante del modelo
- Q_{obs} : el caudal observado

En este trabajo, la calibración se llevó a cabo mediante el software SWAT-CUP (Abbaspour et al., 2007), que es una herramienta especialmente diseñada para la calibración del modelo SWAT y permite utilizar varios algoritmos de optimización. En este caso, se utilizó el algoritmo SUFI-2, (Abbaspour et al., 2004; Abbaspour et al., 2007) por ser el de aplicación más extendida.

Se seleccionaron 24 parámetros para optimizar, vinculados directamente con los procesos físicos relevantes en la cuenca. En la Tabla 3 se presentan estos parámetros junto con una breve descripción de su significado físico.

Tabla 3.- Resumen de parámetros considerados durante la calibración.

Parámetro	Interpretación	Parámetro	Interpretación
CN2.mgt	Número de curva asociado a la cobertura del suelo	**CH_K1.sub**	Conductividad hidráulica de canales (dentro de subcuencas)
ALPHA_BF.gw	Constante de recesión del flujo base	**CH_N1.sub**	Coeficiente de rugosidad de Manning de canales (dentro de subcuencas)
GW_DELAY.gw	Tiempo de retardo del flujo subterráneo	**CH_S1.sub**	Pendiente media de canales (dentro de subcuencas)
CH_K2.rte	Conductividad hidráulica de canales (fuera de subcuencas)	**CH_S2.rte**	Pendiente media de canales (fuera de subcuencas)
CH_N2.rte	Coeficiente de rugosidad de Manning de canales (fuera de subcuencas)	**EPCO.hru**	Factor de compensación de toma de agua por parte de cultivos
OV_N.hru	Coeficiente de rugosidad de Manning de flujo no concentrado	**HRU_SLP.hru**	Pendiente de la HRU
ESCO.hru	Factor de compensación de evaporación de agua en el suelo	**SLSUBBSN.hru**	Longitud de la pendiente
GW_REVAP.gw	Coeficiente de retorno del agua subterránea a la zona insaturada	**SOL_Z.sol**	Profundidad del suelo
SOL_K.sol	Conductividad hidráulica saturada del suelo	**SURLAG.bsn**	Coeficiente de retardo del flujo superficial
SOL_AWC.sol	Agua Disponible del suelo	**EVRCH.bsn**	Factor de ajuste de evaporación
GWQMN.gw	Contenido de agua límite en reservorio subterráneo para que haya flujo de retorno	**RES_RR.res**	Descarga media diaria de la represa
CANMX.hru_	Capacidad máxima de intercepción en las hojas de árboles y cultivos	**RES_K.res**	Conductividad hidráulica del fondo de la represa

El proceso de calibración da como resultado un rango óptimo de parámetros, así como el conjunto de ellos que maximiza la función objetivo. Una vez finalizado el proceso, se mide el desempeño del modelo resultante a través de los siguientes estadísticos:

$$Coeficiente\ de\ determinación: \qquad R^2 = \frac{\left[\sum_i \left(Q_{mod,i} - \overline{Q_{mod}}\right)\left(Q_{obs,i} - \overline{Q_{obs}}\right)\right]^2}{\sum_i \left(Q_{mod,i} - \overline{Q_{mod}}\right)^2 \sum_i \left(Q_{obs,i} - \overline{Q_{obs}}\right)^2} \qquad (Ecuación\ 2)$$

$$Raíz\ del\ error\ medio\ cuadrático: \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{mod} - Q_{obs})_i^2} \qquad (Ecuación\ 3)$$

$$Eficiencia\ de\ Kling - Gupta: \qquad KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \qquad (Ecuación\ 4)$$

$$\alpha = \frac{\sigma_{mod}}{\sigma_{obs}} \qquad \beta = \frac{\mu_{mod}}{\mu_{obs}} \qquad r = coef.\,de\ regresión\ lineal\ entre\ mod.\,y\ obs.$$

Siendo:

- Q_{mod} : el caudal resultante del modelo
- Q_{obs} : el caudal observado
- N : la cantidad total de datos de las series comparadas
- σ_{mod} : la desviación estándar de la serie resultante del modelo
- σ_{obs} : la desviación estándar de la serie de caudales observados
- μ_{mod} : la media de la serie resultante del modelo
- μ_{obs} : la media de la serie de caudales observados

6.1.3 Validación

La validación consiste en la aplicación del modelo resultante a una serie temporal diferente a la utilizada durante la calibración. De ese modo, se verifica su validez fuera del rango de condiciones para el cual se ajustaron los parámetros.

En este caso, para ello se utilizó el software SWAT-CUP, para aplicar el rango de parámetros obtenido durante la calibración a los datos de un período de tiempo diferente. Como estadísticos de comparación se utilizaron todos los antes mencionados (Ecuaciones 1, 2, 3 y 4).

6.2 Modelación basada en datos

Para la modelación basada en datos, se optó por utilizar un método que permita extraer la mayor cantidad posible de información acerca del proceso de entrenamiento (de "caja blanca"). Por este motivo, el algoritmo de aprendizaje seleccionado fue Random Forest (Brieman, 2001), el cual además cuenta con gran cantidad de antecedentes en modelación hidrológica.

6.2.1 Entrenamiento y análisis de sensibilidad

Durante el entrenamiento, el algoritmo procesa la matriz de datos de entrada para encontrar relaciones entre las variables y "aprender" el comportamiento de la variable de salida. El desempeño del modelo no se mide sobre la serie temporal con la que se entrenan los datos, sino sobre otro conjunto, llamado de validación. Este procedimiento se realiza para evitar el sobre ajuste del algoritmo

al conjunto de entrenamiento. Este proceso se repite, ajustando los parámetros del modelo o las series de datos de entrada, hasta lograr una performance adecuada sobre el conjunto de validación.

En nuestro caso, la variable de salida es el caudal en Paso Severino ("Q_TOT"), y las variables de entrada son todas las que se describen en la Sección 3 a excepción de las series hidrométricas en Florida y los niveles en Paso Severino. A esto se agrega la variable de la fecha, que indica la temporalidad que caracteriza el problema estudiado.

Para llevar a cabo un proceso de entrenamiento efectivo, se realizaron algunas modificaciones previas al conjunto de datos de entrada, habituales para este tipo de problemas. En primer lugar, se normalizaron las variables continuas (todas menos la fecha) utilizando un esquema "Min-Max". Se optó por este por sobre la normalización Gaussiana para mantener el sentido físico de las variables involucradas, las cuales adquieren valores no-nulos. En segundo lugar, se aplicó codificación "One-Hot" a las variables categóricas. Por último, se aplicó interpolación lineal para el relleno de datos faltantes, ya que se trata de una cantidad menor y no hay períodos largos con ausencia de datos.

El proceso precipitación-escorrentía, que se quiere modelar, se caracteriza por tener una fuerte dependencia con las condiciones de humedad antecedentes en la cuenca, debido a la retención de humedad en el suelo y en reservorios subterráneos. Para modelarlo correctamente, fue necesario generar variables de ingreso artificiales (a partir de las existentes) que permitieran al algoritmo asimilar esta dependencia compleja. En la Tabla 4 se describen las variables generadas.

Tabla 4.- Variables artificiales introducidas para representar la temporalidad de los procesos físicos involucrados.

Variables modificadas	Sufijo	Descripción
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay	_acum7	Valor acumulado en los 7 días previos (última semana)
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay	_acum14	Valor acumulado entre los 8 y 14 días previos (penúltima semana)
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_lag7	Valor de la variable 7 días atrás.
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_lag14	Valor de la variable 14 días atrás.
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_mean 7 _mean 14	Promedio en la última semana y en la penúltima.
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_median7 _median14	Mediana de la última y semana y de la penúltima
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_max7 _max14	Valor máximo de la última semana y de la penúltima
PP_CCol, PP_Flor, PP_LCru, PP_Mend, PP_ Rebo, PP_SGab, PP_SGra, PP_VMay, RHum, Slr, TMax, TMin, Wnd, ETP	_min7 _min14	Valor mínimo de la última semana y de la penúltima
Fecha	-	Desglose en: año, mes, día, estación, día del año, día del mes, día de la semana, semana del año, semana del mes, elapsed.

Esto motivó la realización de un análisis de sensibilidad que permitió evaluar la respuesta del entrenamiento a las diferentes variables, con el fin de obtener información sobre el aprendizaje de los procesos más relevantes. Para esto, se entrenaron 11 modelos diferentes, cada uno con un conjunto diferente de datos de entrada, evaluándose su performance sobre el conjunto de validación. En todos los casos se utilizaron 1000 árboles completos, sin establecer límites al crecimiento del árbol, ni a la cantidad de variables u observaciones utilizadas en cada uno. Cabe mencionar que se utilizaron las mismas medidas de desempeño que para la modelación de base física: NS, R^2, RMSE y KGE (Ecuaciones 1, 2, 3 y 4 respectivamente). En la Tabla 5, se presenta la selección de variables para cada uno.

Tabla 5.- Modelos entrenados durante análisis de sensibilidad a variables de input.

	Variables
Modelo 0	Todas
Modelo 1	Fecha desglosada, variables originales
Modelo 2	Fecha desglosada, variables originales, _lag14
Modelo 3	Fecha desglosada, variables originales, _lag7
Modelo 4	Fecha desglosada, variables originales, _lag7, _lag14
Modelo 5	Fecha desglosada, variables originales, _lag7, _lag14, _acum7, _acum14
Modelo 6	Fecha desglosada, variables originales, _acum7, _acum14
Modelo 7	Fecha desglosada, variables originales, _lag7, _lag14, _acum7, _acum14, _mean7, _mean14
Modelo 8	Fecha desglosada, variables originales, _lag7, _lag14, _acum7, _acum14, _mean7, _mean14, _min7, _min14, _max7, _max14
Modelo 9	Fecha desglosada, variables originales, _lag7, _lag14, _mean7, _mean14, _min7, _min14, _max7, _max14
Modelo 10	Solo fecha “elapsed”, todas las variables

7. Resultados

7.1 Modelación basada en procesos físicos

A continuación, se presentan los resultados de la calibración y validación del modelo basado en procesos físicos. El período considerado para calibración fue entre el 1/1/1993 y el 31/12/2002, mientras que para validación entre el 1/1/2003 y el 31/12/2006.

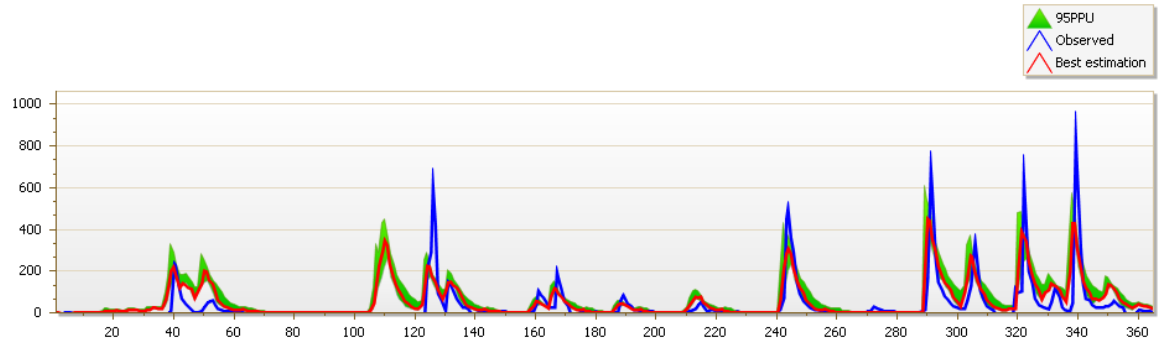

Figura 11.- Resultado de calibración para modelo SWAT. Serie temporal correspondiente a Paso Severino, para el año 1993.

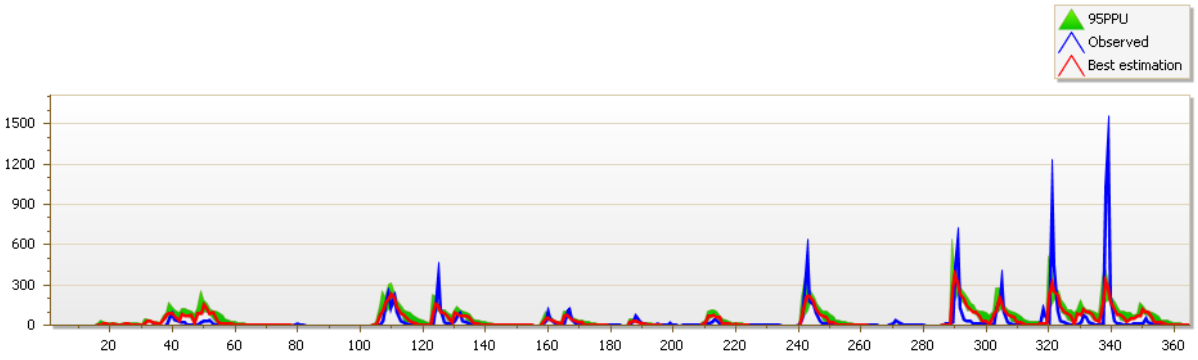

Figura 12.- Resultado de calibración para el modelo SWAT. Serie temporal correspondiente a Florida, para el año 1993.

Tabla 6.- Parámetros resultantes de la calibración del modelo SWAT.

Parámetro	Valor Paso Severino	Valor Florida
NS	0.58	0.51
R^2	0.62	0.53
RMSE	64.8	78.7
KGE	0.44	0.49

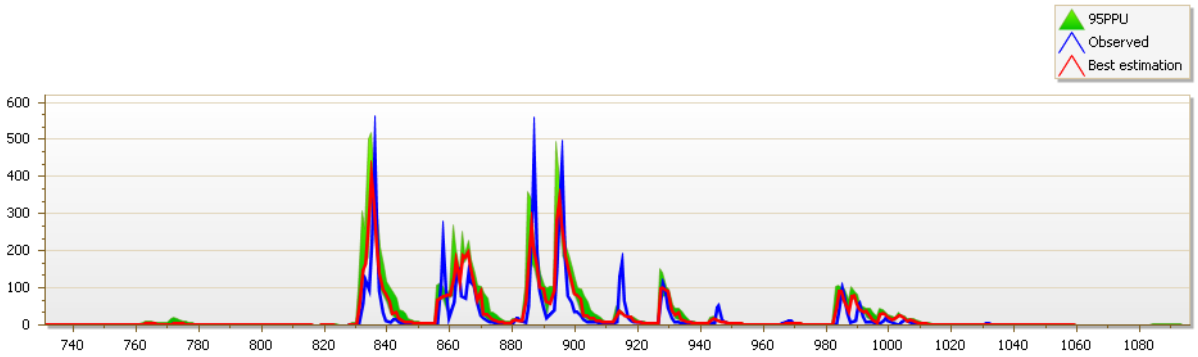

Figura 13.- Resultado de validación para el modelo SWAT. Serie temporal correspondiente a Paso Severino, para el año 2005.

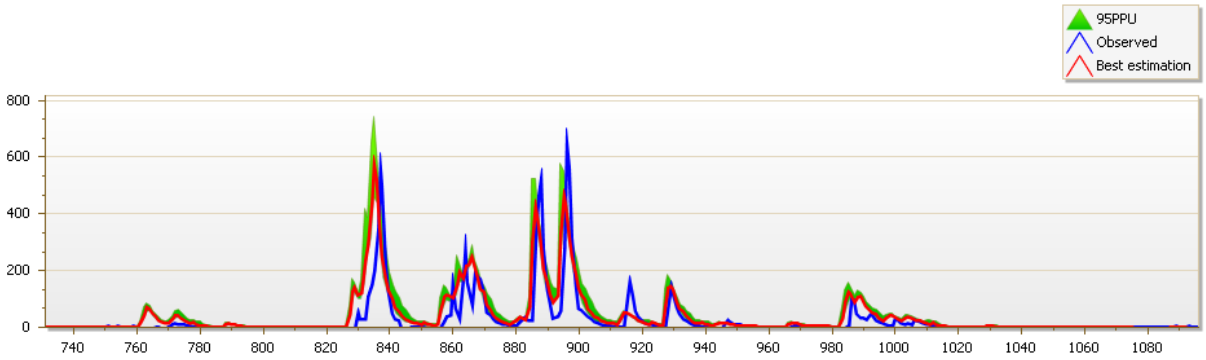

Figura 14.- Resultado de validación para el modelo SWAT. Serie temporal correspondiente a Florida, para el año 2005.

Tabla 7.- Parámetros resultantes de la validación del modelo SWAT.

Parámetro	Valor Paso Severino	Valor Florida
NS	0.60	0.63
R^2	0.64	0.64

RMSE	57.4	43.6
KGE	0.57	0.62

7.2 Modelación basada en datos

A continuación, se presentan los resultados de los 11 modelos que se entrenaron durante el análisis de sensibilidad a las variables de entrada para el modelo basado en datos. Para el conjunto de entrenamiento se consideró el período entre el 1/1/1993 y el 31/12/2002, mientras que para el de validación entre el 1/1/2003 y el 31/12/2006. De ese modo, ambos modelos se implementaron para el mismo período de tiempo y por lo tanto sus resultados son comparables.

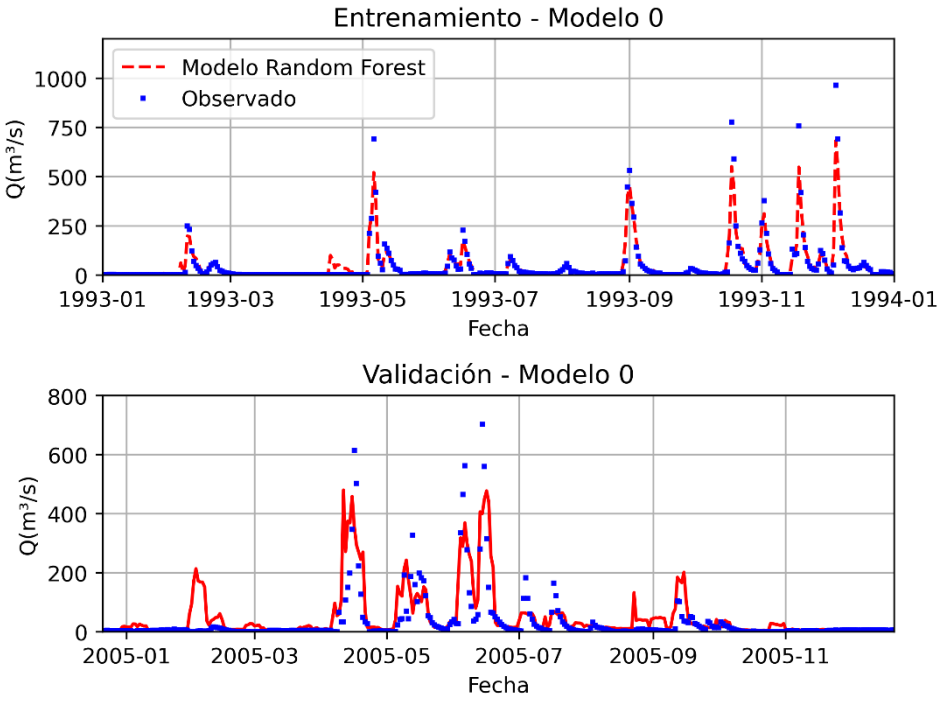

Figura 15.- Resultados de entrenamiento del Modelo 0.

Tabla 8.- Resultados de entrenamiento del Modelo 0.

Parámetro	**Entrenamiento**	**Validación**
NS	0.94	0.35
R^2	0.94	0.35
RMSE	25.0	75.6
KGE	0.71	0.43

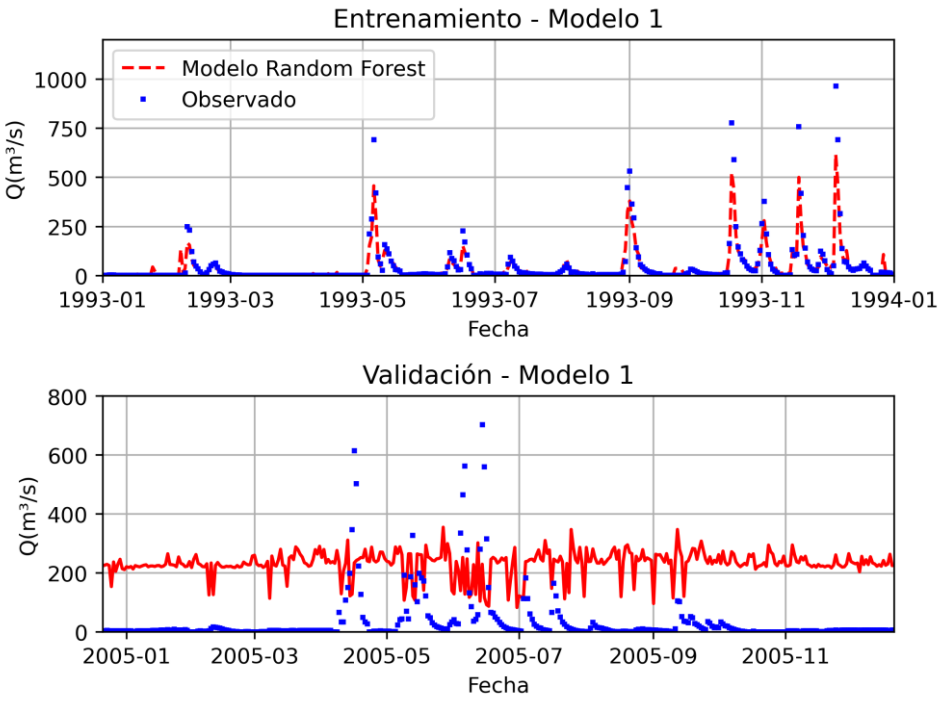

Figura 16.- Resultados de entrenamiento del Modelo 1.

Tabla 9.- Resultados de entrenamiento del Modelo 1.

Parámetro	Entrenamiento	Validación
NS	0.89	-5.42
R^2	0.89	-5.42
RMSE	33.6	229.1
KGE	0.5	-0.62

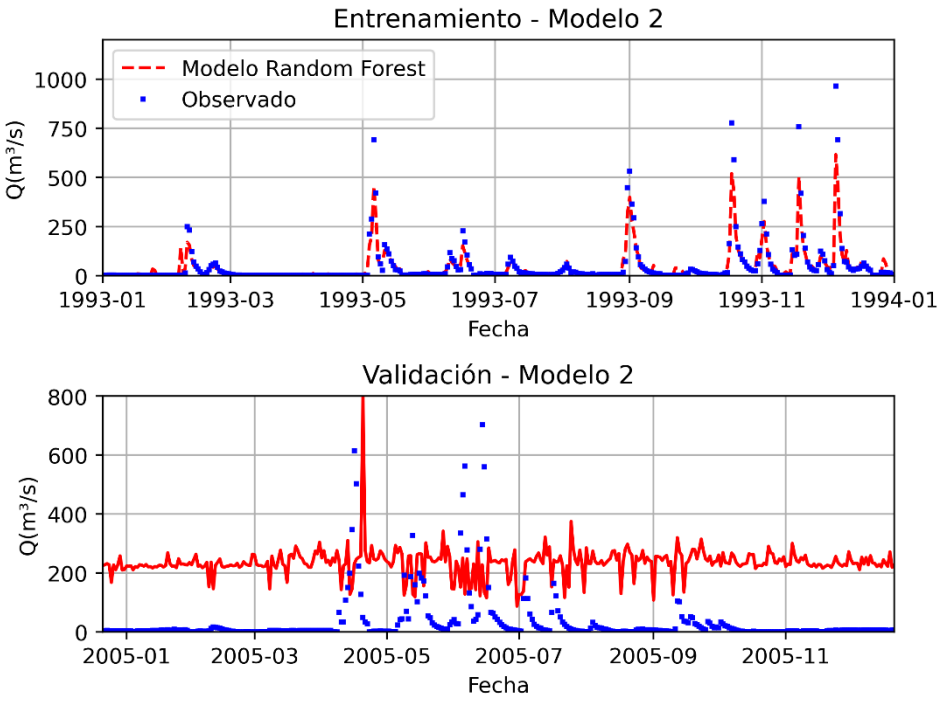

Figura 17.- Resultados de entrenamiento del Modelo 2.

Tabla 10.- Resultados de entrenamiento del Modelo 2.

Parámetro	Entrenamiento	Validación
NS	0.88	-5.53
R^2	0.88	-5.53
RMSE	34.7	230.9
KGE	0.5	-0.57

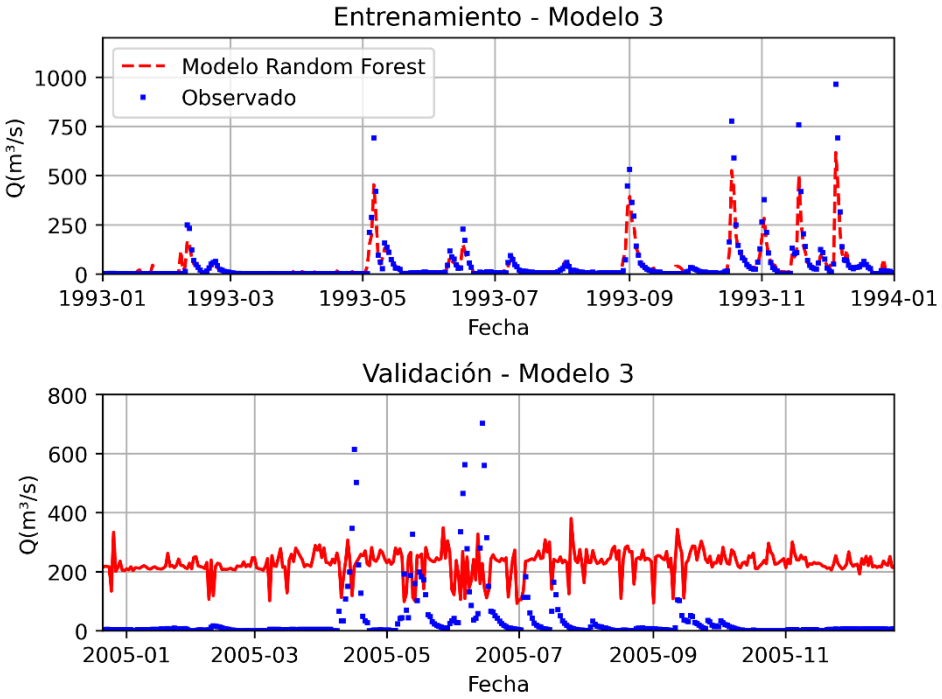

Figura 18.- Resultados de entrenamiento del Modelo 3.

Tabla 11.- Resultados de entrenamiento del Modelo 3.

Parámetro	Entrenamiento	Validación
NS	0.88	-5.23
R^2	0.88	-5.23
RMSE	34.2	225.5
KGE	0.5	-0.63

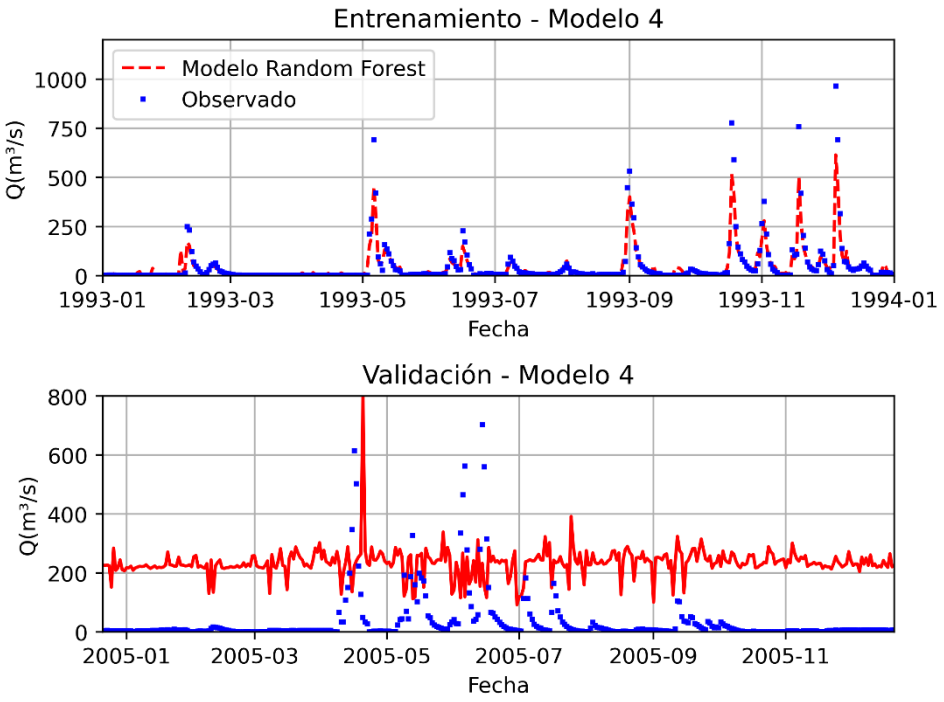

Figura 19.- Resultados de entrenamiento del Modelo 4.

Tabla 12.- Resultados de entrenamiento del Modelo 4.

Parámetro	Entrenamiento	Validación
NS	0.88	-5.5
R^2	0.88	-5.5
RMSE	35.0	230.4
KGE	0.5	-0.57

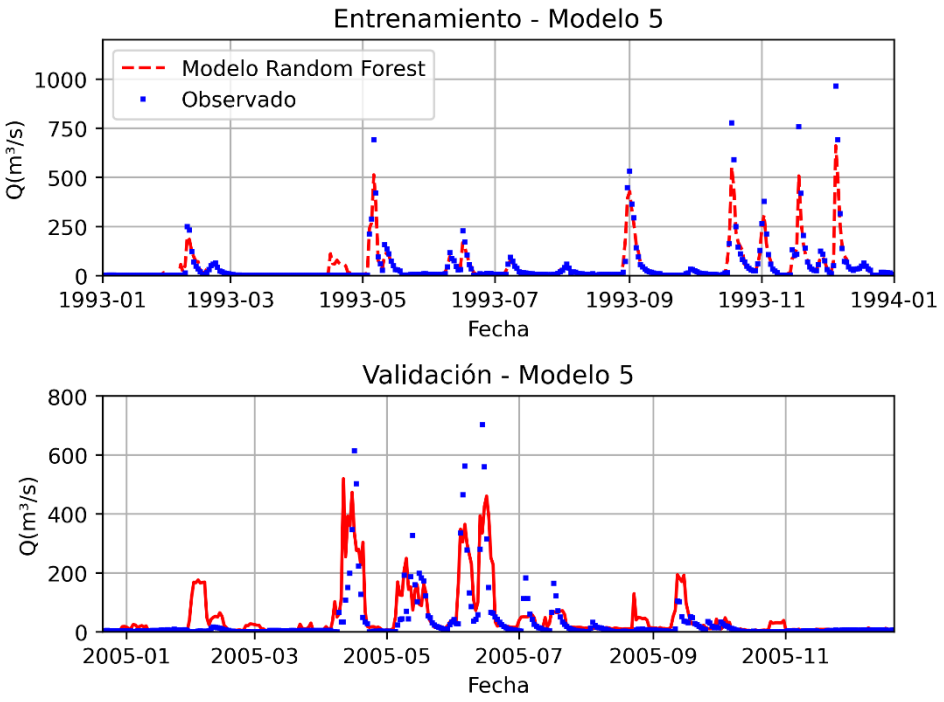

Figura 20.- Resultados de entrenamiento del Modelo 5.

Tabla 13.- Resultados de entrenamiento del Modelo 5.

Parámetro	Entrenamiento	Validación
NS	0.93	0.31
R^2	0.93	0.31
RMSE	26.7	75.1
KGE	0.69	0.43

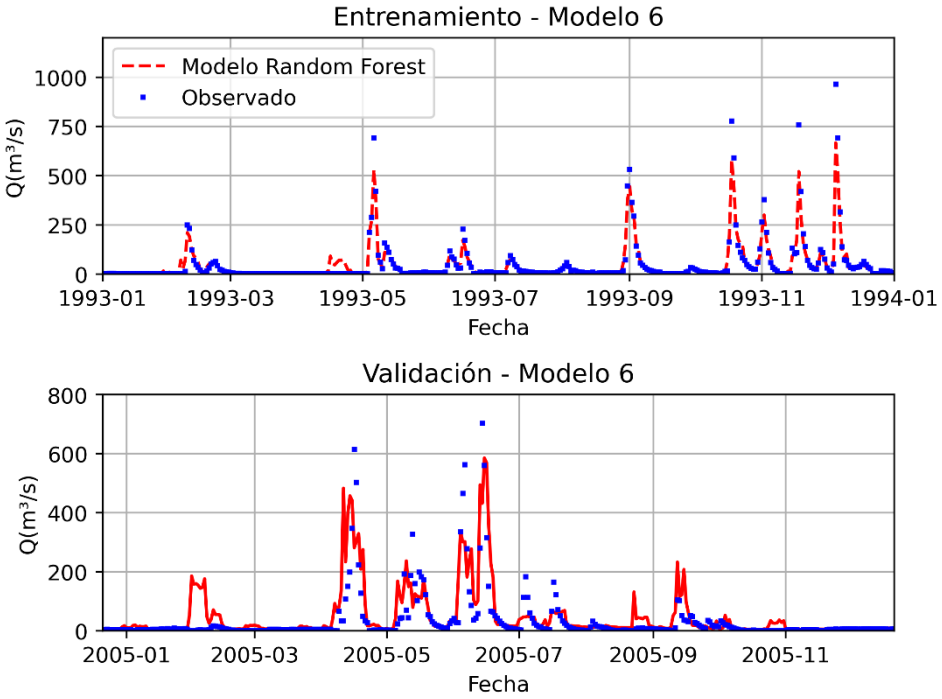

Figura 21.- Resultados de entrenamiento del Modelo 6.

Tabla 14.- Resultados de entrenamiento del Modelo 6.

Parámetro	Entrenamiento	Validación
NS	0.93	0.26
R^2	0.93	0.26
RMSE	26.5	77.6
KGE	0.68	0.42

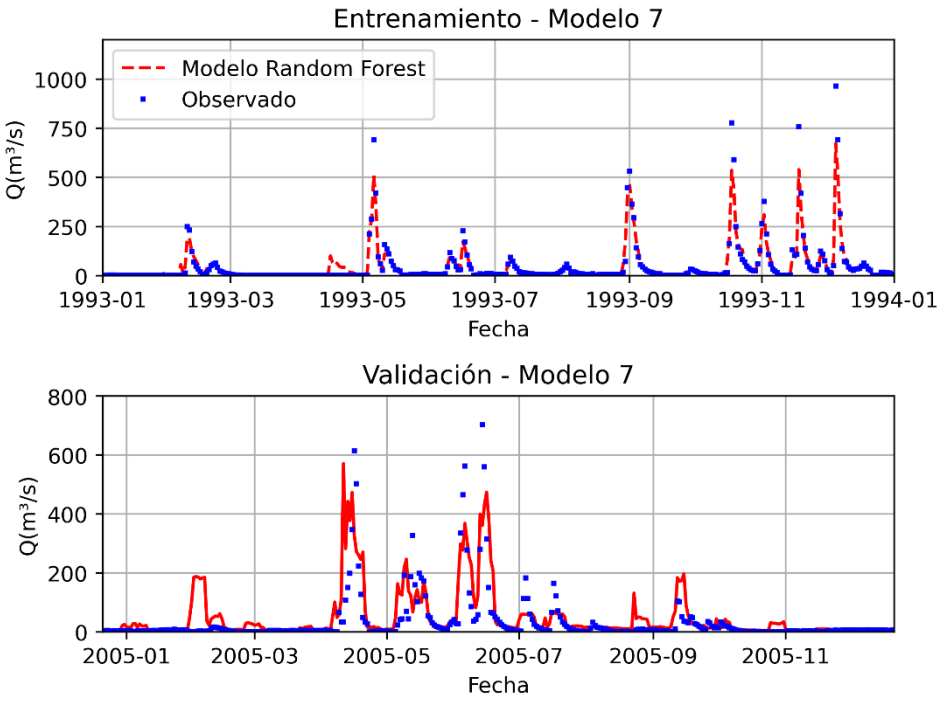

Figura 22.- Resultados de entrenamiento del Modelo 7.

Tabla 15.- Resultados de entrenamiento del Modelo 7.

Parámetro	Entrenamiento	Validación
NS	0.93	0.32
R^2	0.93	0.32
RMSE	25.5	74.6
KGE	0.7	0.42

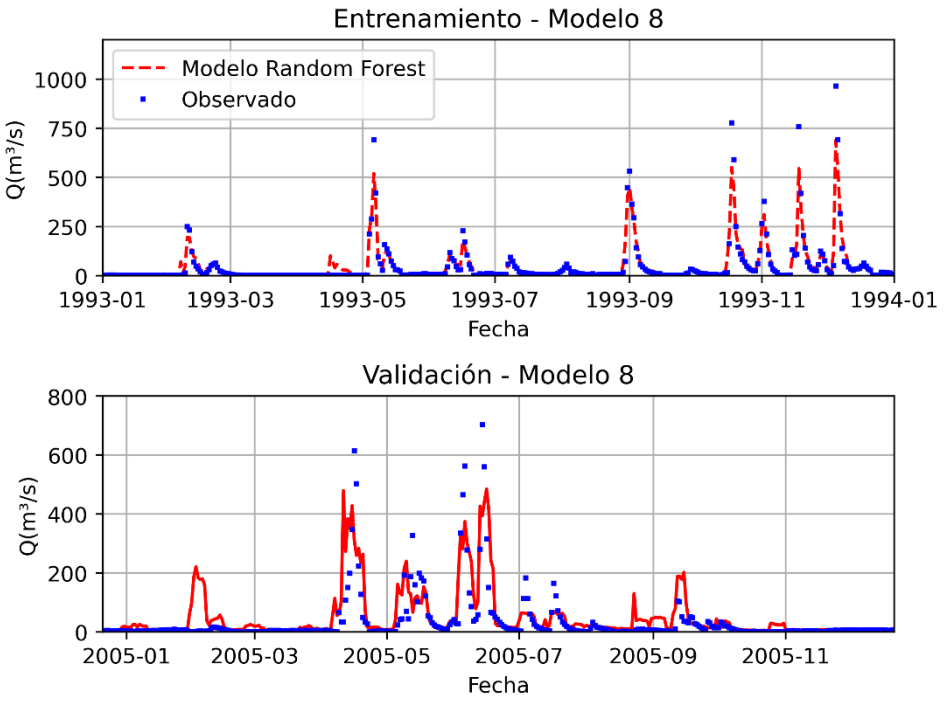

Figura 23.- Resultados de entrenamiento del Modelo 8.

Tabla 16.- Resultados de entrenamiento del Modelo 8.

Parámetro	Entrenamiento	Validación
NS	0.94	0.35
R^2	0.94	0.35
RMSE	25.1	72.9
KGE	0.71	0.42

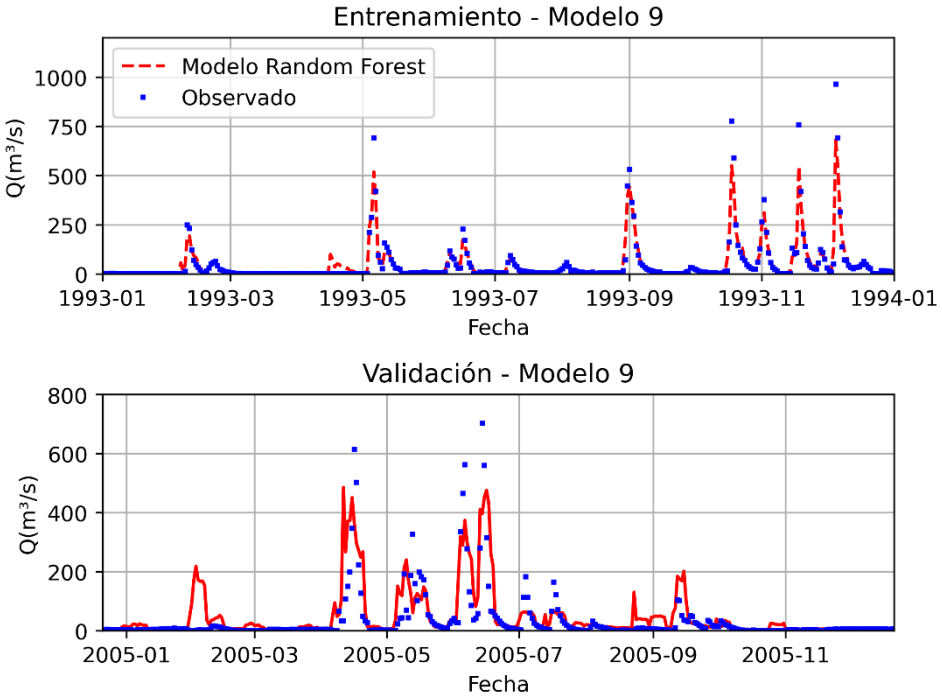

Figura 24.- Resultados de entrenamiento del Modelo 9.

Tabla 17.- Resultados de entrenamiento del Modelo 9.

Parámetro	Entrenamiento	Validación
NS	0.94	0.36
R^2	0.94	0.36
RMSE	25.1	72.3
KGE	0.71	0.44

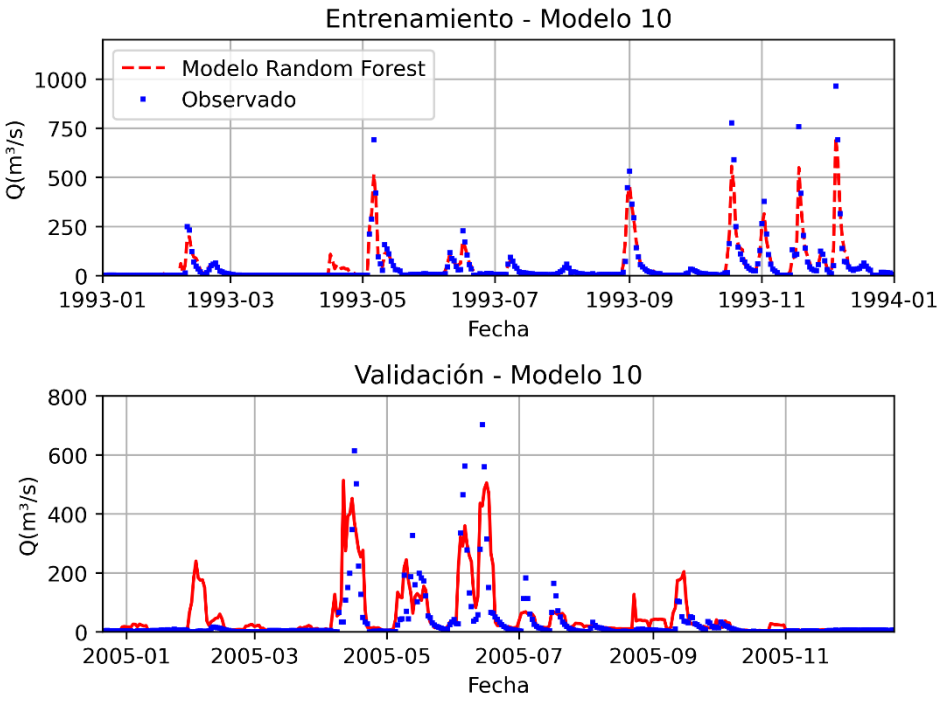

Figura 25.- Resultados de entrenamiento del Modelo 10.

Tabla 18.- Resultados de entrenamiento del Modelo 10.

Parámetro	Entrenamiento	Validación
NS	0.94	0.34
R^2	0.94	0.34
RMSE	24.8	73.6
KGE	0.73	0.45

7.3 Comparación de modelos

En la Tabla 19 se presentan comparativamente los resultados del modelo de base física para Paso Severino junto con los del Modelo 10. Si bien se obtuvo un mejor desempeño para el conjunto de validación con el Modelo 9, se elige el Modelo 10 porque es el segundo en desempeño, pero utiliza el total de las variables, por lo que permite extraer conclusiones sobre su importancia sin excluir ninguna.

Tabla 19.- Comparación de indicadores de ambos modelos.

	Entrenamiento		Validación	
Parámetro	**SWAT**	**Random Forest (Modelo 10)**	**SWAT**	**Random Forest (Modelo 10)**
NS	0.58	0.94	0.60	0.34
R^2	0.62	0.94	0.64	0.34
RMSE	64.8	24.8	57.4	73.6
KGE	0.44	0.73	0.57	0.45

Para complementar la comparación, se extrajeron los resultados del análisis de sensibilidad de ambos modelos. En el caso del modelo de base física, el mismo se realiza por el software SWAT-CUP a partir de los resultados del algoritmo SUFI-2. El resultado, que se presenta en la Figura 26 indica los parámetros que tienen mayor incidencia en la salida del modelo. El diagrama resulta de una prueba

de significancia estadística, por lo que cuanto menor sea el p-valor asociado, mayor será la sensibilidad al parámetro.

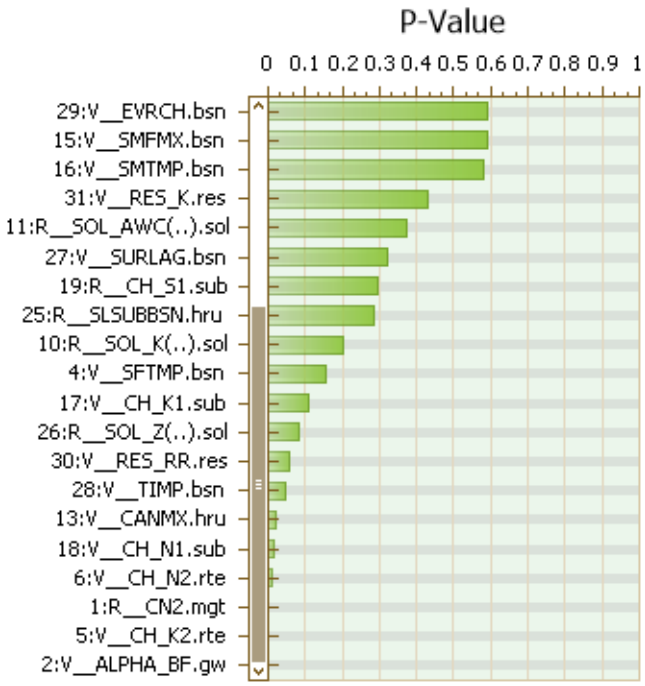

Figura 26.- Resultado de análisis de sensibilidad del modelo SWAT.

El algoritmo Random Forest permite cuantificar la importancia de las variables de entrada. En la Tabla 20 se presenta este resultado para las 10 variables más significativas del Modelo 10.

Tabla 20.- Importancia de variables resultante del Modelo 10.

Variable	Importancia
PP_LCru_acum7	0.015
PP_LCru_mean7	0.006
PP_LCru	0.036
PP_SGra	0.012
PP_SGra_acum7	0.003
PP_SGra_mean7	0.013
TMin_lag7	0.031
RHum	0.009
Wnd_max14	0.02
ETP_mean14	0.011

Cabe destacar que el análisis de sensibilidad del modelo SWAT, y el de importancia de variables del Modelo 10 no son directamente comparables, ya que en el primer caso se representa la sensibilidad a parámetros del modelo matemático, mientras que en el segundo se toman en cuenta las variables de entrada. No obstante, de ambos se pueden sacar conclusiones sobre la relevancia de los procesos físicos involucrados, que es en lo que se enfoca el trabajo.

8. Discusión

8.1 Modelación basada en procesos físicos

Al observar los gráficos comparativos de series temporales, se detectan algunos eventos de crecida que la serie de caudales observados en Paso Severino no representa correctamente. Se llega a esta conclusión porque dichos eventos sí se ven en la serie de Florida, y también en la salida del modelo, lo que quiere decir que responden a eventos de precipitación registrados por la red pluviométrica. Si bien esto ocurre en algunos eventos puntuales, podría verse afectada la calidad del modelo, ya que condiciona los parámetros de ajuste y por lo tanto también el proceso de calibración. La diferencia parece deberse a un problema de calidad de datos, no detectado durante la etapa de análisis, aunque también podría ser causada por el efecto de laminación que produce el embalse de Paso Severino. Sin embargo, esto no parece ser el caso ya que dicho efecto se refleja sistemáticamente en todos los eventos, en que los caudales máximos son mayores el Florida que en Paso Severino. No obstante, se requiere un análisis más en detalle de estos eventos en particular para determinar la causa del problema e intentar corregirlo.

Los gráficos correspondientes a la calibración muestran que el modelo es capaz de representar adecuadamente los caudales base, pero subestima los picos durante crecidas, tanto en Florida como en Paso Severino. Esto demuestra que el modelo aún puede ser mejorado. Para esto, se debe trabajar sobre la selección de parámetros para la calibración y el rango de valores considerado. A pesar de esto, en el período de validación los picos de caudal parecen representarse con mayor exactitud. Esto puede deberse a diferencias climáticas en ambos períodos. En particular el de validación resulta relativamente corto, y puede haberse seleccionado una ventana temporal caracterizada por crecidas de poca magnitud. Sería bueno, en este sentido, realizar otra validación con una ventana temporal diferente, para caracterizar lo mejor posible el desempeño del modelo.

Los valores de NS obtenidos, tanto para calibración como para validación resultan adecuados, si se tienen en cuenta los valores de referencia publicados en Moriasi et al. (2007), presentados en la Tabla 21.

Tabla 21.- Valores criterio de NS para modelo hidrológico SWAT. Extraído de Moriasi et al. (2007).

Calificación	**Rango de valores**
Satisfactorio	NS > 0.5
Adecuado	0.54 < NS < 0.65
Muy bueno	NS > 0.65

8.2 Modelación basada en datos

Lo primero que se observa, en todos los modelos generados a partir del algoritmo Random Forest, es la tendencia al sobreajuste del conjunto de datos de entrenamiento. Esto se ve reflejado en valores de NS cercanos a 1 durante el entrenamiento, pero significativamente más bajos al aplicar el modelo al conjunto de validación. Para mejorar este aspecto es necesario un ajuste de los parámetros del aprendizaje, que no fueron modificados durante este análisis de sensibilidad a las variables de entrada. Una vez seleccionadas las variables de trabajo, se debería limitar la cantidad de variables y observaciones que se utilizan para la construcción de cada árbol, y a su vez limitar su profundidad mediante algún algoritmo de poda. El procedimiento óptimo sería llevar a cabo una validación cruzada que permita determinar el conjunto de parámetros de aprendizaje más adecuado.

Comparando los resultados de los distintos modelos generados, resulta claro que la incorporación de las variables artificiales que se generaron es relevante para el aprendizaje del proceso precipitación-

escorrentía. Se interpreta que las mismas dan la información necesaria para la representación de los procesos de retención de agua en los distintos estratos del suelo. Se observa además que las variables _lag7 y _lag14, que representan el valor pasado de las variables, no son capaces por sí mismos de dar la información necesaria, sino que se requieren datos de agregación en las últimas semanas como acumulados (_acum7, _acum14) o promedios (_mean7, mean14). Los modelos que consideran estas variables son los que alcanzan los mejores resultados (Modelos 0, 5, 6, 7, 8, 9 y 10), siendo estos comparables entre sí. Los mejores parámetros se alcanzan para el Modelo 9, el cual no utiliza los acumulados en las semanas previas. Esto deja en evidencia que tal vez la inclusión de todos los parámetros esté aportando información redundante que favorece el sobreajuste. Por último, se destaca que la información de fecha desglosada en múltiples variables no parece aportar información adicional. Esto seguramente se deba a que las variables artificiales por sí mismas sean suficientes para el aprendizaje de la temporalidad por parte del algoritmo.

Si bien la comparación de series temporales muestra que los modelos de mejor desempeño logran capturar y representar los principales procesos físicos (flujo base, eventos de crecida), si se comparan los indicadores con los de otros modelos similares basados en Random Forest (Shortridge et al., 2016; Tongal & Booij, 2018; White, 2017) los mismos no pueden considerarse satisfactorios aún. Es esperable que estos mejoren una vez que se ajusten los parámetros del aprendizaje mediante validación cruzada.

8.3 Comparación de modelos

El nivel de desarrollo actual de ambos modelos muestra un mejor desempeño del basado en procesos físicos, ya que representa satisfactoriamente los principales procesos y muestra indicadores de desempeño adecuados. Los modelos basados en datos de mejor desempeño son capaces de representar la respuesta de la cuenca en escenarios de estiaje o crecida, pero no alcanzan indicadores satisfactorios.

El análisis de sensibilidad a los parámetros del modelo SWAT muestra que los más sensibles son “ALPHA_BF”, “CH_K2” y “CN2”. Tomando en cuenta las descripciones de la Tabla 3, se puede interpretar que los procesos de más relevancia son: el aporte subterráneo al flujo base, el transporte concentrado de flujo y el escurrimiento superficial durante tormentas, respectivamente, ya que esos son los procesos condicionados por estos parámetros.

La importancia de variables determinada por el algoritmo Random Forest muestra entre las principales a las variables artificiales “_acum7” y “_mean7” para la precipitación, junto con la precipitación original para el pluviómetro ubicado en La Cruz. Como es de esperar, la precipitación es la variable que tiene mayor relevancia, pues determina el volumen de agua dentro de la cuenca. A su vez, las variables artificiales “_acum7” y “_mean7” indican la importancia de los procesos asociados a la retención de agua en la cuenca, como pueden ser el aporte subterráneo al flujo base o el transporte del flujo concentrado, los cuales también resultaron relevantes en el modelo físicamente basado.

9. Conclusiones

Se implementaron dos modelos hidrológicos en la cuenca del río Santa Lucía Chico: un modelo basado en principios físicos, utilizando la herramienta SWAT; y un modelo basado en datos, a partir de aprendizaje automático mediante Random Forest. Si bien en ambos casos es posible y necesario realizar mejoras, se logró representar correctamente la temporalidad de los procesos y ambos modelos responden adecuadamente a las variables de input generando períodos de sequía y crecida en correspondencia con los datos observados. Dentro de las limitaciones de los modelos, se tiene que

el basado en procesos físicos presenta subestimación sistemática de los picos en eventos de crecida, mientras que el basado en datos no alcanza indicadores de ajuste satisfactorios.

La comparación de ambos desde el punto de vista de los procesos físicos mostró coincidencia en la relevancia que tienen los procesos de aporte subterráneo al flujo base, el transporte concentrado de flujo y el escurrimiento superficial durante tormentas. Esto es un indicio del potencial de los modelos basados en datos para interpretar la física detrás del problema.

Si bien se requiere continuar el trabajo sobre los modelos para mejorar sus indicadores de ajuste y su capacidad de representar el proceso precipitación-escorrentía, el nivel de desarrollo alcanzado permitió realizar una comparación de ambos desde el punto de vista de los subprocesos físicos involucrados. Por lo tanto se considera que se cumplieron los objetivos planteados para el trabajo.

10. Referencias

Abbaspour, K. C., Johnson, C. A., & Genuchten, M. T. van. (2004). Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure. *Vasose Zone Journal*, *3*(4), 1340–1352. https://doi.org/https://doi.org/10.2136/vzj2004.1340

Abbaspour, Karim C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., & Srinivasan, R. (2007). Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. *Journal of Hydrology*, *333*(2–4), 413–430. https://doi.org/10.1016/j.jhydrol.2006.09.014

Arnold, J. G., Srinivasan, R., Muttiah, R. S., & Wiliams, J. R. (1998). Large Area Hydrologic Modeling and Assessment Part I: Model development. *Journal of the American Water Resources Association*, *34*, 73–89. https://doi.org/https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Brieman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1201/9780429469275-8

Chow, V., Maidment, D., & Mays, L. (1994). *Hidrología aplicada*. McGraw-Hill, Inc.

Fu, M., Fan, T., Ding, Z., Salih, S. Q., Al-Ansari, N., & Yaseen, Z. M. (2020). Deep Learning Data-Intelligence Model Based on Adjusted Forecasting Window Scale: Application in Daily Streamflow Simulation. *IEEE Access*, *8*, 32632–32651. https://doi.org/10.1109/ACCESS.2020.2974406

Kan, G., He, X., Ding, L., Li, J., Hong, Y., Lei, T., Lei, T., Liang, K., Zuo, D., & Huang, P. (2017). Daily streamflow simulation based on the improved machine learning method. *Tecnologia y Ciencias Del Agua*, *8*(2), 51–60. https://doi.org/10.24850/j-tyca-2017-02-05

Kim, J., Han, H., Johnson, L. E., Lim, S., & Cifelli, R. (2019). Hybrid machine learning framework for hydrological assessment. *Journal of Hydrology*, *577*(July), 123913. https://doi.org/10.1016/j.jhydrol.2019.123913

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). NeuralHydrology – Interpreting LSTMs in Hydrology. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*, 347–362. https://doi.org/10.1007/978-3-030-28954-6_19

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). *Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks*. 6005–6022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Moriasi, D. N., Arnold, J. ., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, *50*(3), 886–900. https://doi.org/10.13031/2013.23153

Neitsch, S. ., Arnold, J. ., Kiniry, J. ., & Williams, J. . (2011). Soil & Water Assessment Tool Theoretical Documentation Version 2009. *Texas Water Resources Institute*, 1–647. https://doi.org/10.1016/j.scitotenv.2015.11.063

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016

Tongal, H., & Booij, M. J. (2018). Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology*, *564*(July), 266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004

Wang, Z., Song, H., Watkins, D. W., Ong, K. G., Xue, P., Yang, Q., & Shi, X. (2015). Cyber-physical systems for water sustainability: Challenges and opportunities. *IEEE Communications Magazine*, *53*(5), 216–222. https://doi.org/10.1109/MCOM.2015.7105668

White, E. (2017). Predicting Unimpaired Flow in Ungauged Basins: “Random Forests” Applied to California Streams. *ProQuest Dissertations and Theses*, 69. http://search.proquest.com.libraryproxy.griffith.edu.au/docview/2026286173?accountid=14543%0Ahttp://hy8fy9jj4b.search.serialssolutions.com/directLink?&atitle=Predicting+Unimpaired+Flow+in+Ungauged+Basins%3A+%22Random+Forests%22+Applied+to+California+Strea

Yaseen, Z. M., Allawi, M. F., Yousif, A. A., Jaafar, O., Hamzah, F. M., & El-Shafie, A. (2018). Non-tuned machine learning approach for hydrological time series forecasting. *Neural Computing and Applications*, *30*(5), 1479–1491. https://doi.org/10.1007/s00521-016-2763-0

Yaseen, Z. M., Naganna, S. R., Sa’adi, Z., Samui, P., Ghorbani, M. A., Salih, S. Q., & Shahid, S. (2020). Hourly River Flow Forecasting: Application of Emotional Neural Network Versus Multiple Machine Learning Paradigms. *Water Resources Management*, *34*(3), 1075–1091. https://doi.org/10.1007/s11269-020-02484-w