
# Tesis para “Maestría en explotación de datos y descubrimiento del conocimiento”

## *Estudio de impactos de sequía en rendimientos de cultivos agrícolas mediante métodos de aprendizaje automático*

Reporte de avance para Taller de Tesis I

Alumno: Ing. Santiago Luis Rovere

Director: Dr. Andrés Farall

Profesor de Taller de Tesis I: Dr. Ricardo Maronna

17 de julio de 2021

# 1 Sinopsis

El presente documento constituye el Reporte de avance N°1 correspondiente a la Tesis de Maestría a ser llevada a cabo por el autor. Este trabajo de tesis forma parte de un proyecto de investigación y desarrollo denominado "Diseño e implementación inicial de un Sistema de Información sobre Sequías para el Sur de América del Sur (SISSA)", el cual es financiado por el programa de Bienes Públicos Regionales del Banco Interamericano de Desarrollo (Cooperación Técnica RG-T3308, https://www.iadb.org/es/project/RG-T3308).

El documento "Plan de Trabajo" presentado previamente incluyó una breve descripción del Proyecto SISSA, la descripción del tema de abordar en la tesis y los objetivos buscados. También se presentó un del plan de trabajo a llevar a cabo para concretar el trabajo de tesis en el término de un año, así como también un cronograma de actividades tentativo. Por tal motivo, esta información no será replicada en este documento, salvo aquello que sea necesario para el desarrollo del presente reporte.

Dado que la naturaleza de este documento es presentar el grado de avance del trabajo de investigación, se comenzará por realizar un resumen del flujo de tareas presentadas en el plan de trabajo. Luego, para cada una de las tareas se detallarán las actividades desarrolladas. Para aquellas tareas que aún no hayan sido abordadas, se presentarán las ideas tentativas que se tienen en mente para su ejecución en el futuro.

Por lo expuesto previamente, se proseguirá con un resumen de las tareas vinculadas al plan de trabajo (Sección 2). El resto de las secciones (Sección 3 a Sección 8) se corresponderán con las principales tareas del plan de trabajo. Finalmente, en la Sección 9 se indicarán las actividades más inmediatas a ser llevadas a cabo.

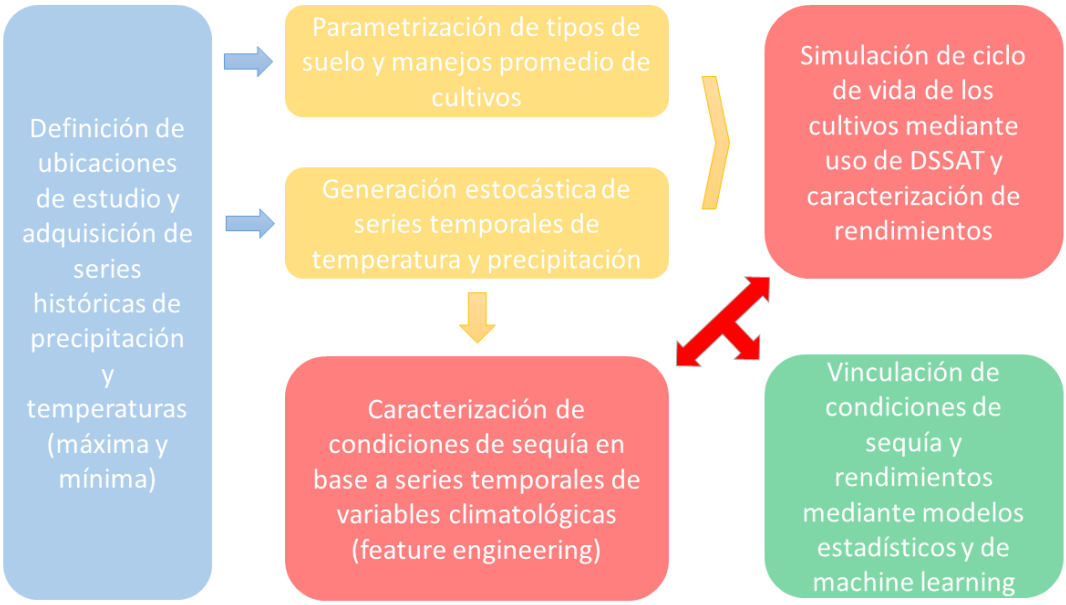
# 2 Plan de trabajo

Para llevar a cabo este trabajo de tesis han propuesto una serie de actividades basadas en al diagrama de la Fig. 1. Para poder abordar el estudio de impactos de sequías en rendimientos de cultivos, se deberán buscar regiones agrícolas importantes dentro del área abarcada por el CRC-SAS (https://www.crc-sas.org; Argentina, Bolivia, Brasil – debajo de 10°S, Chile, Paraguay y Uruguay). En estas regiones se determinarán ubicaciones puntuales para las cuales existan registros históricos largos (de al menos 30 años) de temperatura y precipitaciones.

Una vez definidas las ubicaciones puntuales que se utilizarán para el presente estudio, se deberá recabar información acerca de las actividades agrícolas más importantes, los manejos agronómicos típicos y los tipos de suelos predominantes para cada zona. Esta información permitirá caracterizar y parametrizar las actividades agrícolas y los cultivos cuyos ciclos de vida serán simulados haciendo uso de DSSAT a partir de numerosas series temporales sintéticas de precipitación y temperatura.

Como se mencionó en el párrafo previo, es necesario contar con numerosas series temporales de precipitación y temperatura para poder ejecutar las simulaciones de los ciclos de vida de los cultivos. Para ello se generarán series temporales estocásticas de precipitación y temperatura que tengan las mismas propiedades estadísticas que las series históricas originales. Esto se llevará a cabo haciendo uso del paquete de R *gamwgen* (https://github.com/CRC-SAS/weather-generator) que fue desarrollado, en parte, por integrantes del proyecto SISSA y de otros proyectos anteriores.

Fig. 1. Pipeline conceptual del plan de trabajo propuesto




Haciendo uso de las series temporales estocásticas generadas y las parametrizaciones de actividades agrícolas (cultivos, manejos, tipos de suelo, etc.), se ejecutarán las simulaciones de los ciclos de vida de los cultivos correspondientes. Esto significa que, para cada serie temporal asociada a una campaña agrícola, existirá un rendimiento resultante, producto de la simulación.

Cada una de las series temporales generadas deberá transformarse en un conjunto de atributos que permitan definir condiciones de sequía para cada momento del ciclo de vida del cultivo. Este proceso se realizará mediante el cálculo de eventos basados en índices de sequía actualmente utilizados por el SISSA (https://sissa.crc-sas.org/monitoreo/indices-de-sequia/). A través de este proceso de *feature engineering* se podrá construir un conjunto de datos tabular con atributos y resultados.

Una vez que se haya logrado construir un conjunto de datos tabular con atributos y resultados, será posible aplicar diversos modelos estadísticos y de aprendizaje automático que permitan vincular los atributos (los cuales definen condiciones de sequía) con los rendimientos asociados. Como todo proceso de *data mining*, deberá ser llevado a cabo de manera iterativa e interactiva hasta lograr los objetivos propuestos.

En las secciones subsiguientes se describirán cada uno de los componentes previamente mencionados. Cabe destacar que, dada la naturaleza colaborativa y multidisciplinaria del Proyecto SISSA, algunos de los componentes que se han listado requieren de la intervención de ciertos expertos de campo. Por ejemplo, para la parametrización de los tipos de suelos y de manejos de cultivos, será necesaria la intervención de ingenieros agrónomos que también participan en el proyecto.

También es importante destacar que las tareas no serán realizadas en un estricto orden secuencial, sino que, por cuestiones de interacción entre los integrantes del equipo, algunas de ellas se irán realizando de acuerdo con la dinámica de desarrollo del Proyecto SISSA. Para la ejecución de otras tareas también se hará uso de algunas herramientas y metodologías de cálculo desarrolladas en proyectos previos al SISSA (por ejemplo, para el cálculo de índices de sequía, como se describirá más adelante).

Finalmente, y antes de comenzar la exposición de cada uno de los componentes, debe mencionarse que el grado de avance alcanzado es muy heterogéneo. En algunos casos, aún no se han llevado a cabo actividades vinculadas a varios de los componentes. En otros casos, el grado de avance ha sido considerablemente mayor. Por este motivo, algunas de las secciones detalladas a continuación incluirán mucho menor contenido que otras.

# 3 Definición de zonas de estudio

Como se mencionó al inicio de este documento, el trabajo de tesis se encuentra enmarcado dentro del Proyecto SISSA. Este proyecto aborda, entre otras temáticas, el estudio de la sequía dentro del área abarcada por el Centro Regional del Clima para el Sur de América del Sur (CRC-SAS). Por este motivo, es de interés para el Proyecto SISSA, realizar un estudio de impactos de sequía para las zonas agrícolas más importantes de estos países.

De este modo, la primera etapa de esta tarea consta de una revisión de las actividades agrícolas para los 6 países definidos previamente con la finalidad de identificar las regiones y los cultivos agrícolas más importantes para cada país. Esta tarea está siendo liderada por un ingeniero agrónomo que forma parte del equipo del proyecto.

Como parte de la tarea de identificación de las zonas agrícolas importantes, también es necesario evaluar si existen datos suficientes para poder abordar el estudio de impactos de sequías en cada región. Los datos necesarios para llevar cabo el estudio son los siguientes: (a) series temporales diarias de precipitaciones, temperaturas máximas y mínimas y (b) información de actividades agrícolas, manejos típicos de cultivos y clasificación de suelos.

Los datos diarios de precipitaciones y temperaturas son necesarios para poder generar una cantidad importante de series temporales sintéticas de clima con las mismas propiedades estadísticas que las series históricas. Estas series sintéticas permitirán, a su vez, simular un número considerable de rendimientos para estudiar el vínculo entre las condiciones de sequía y los impactos en términos de rendimientos resultantes.

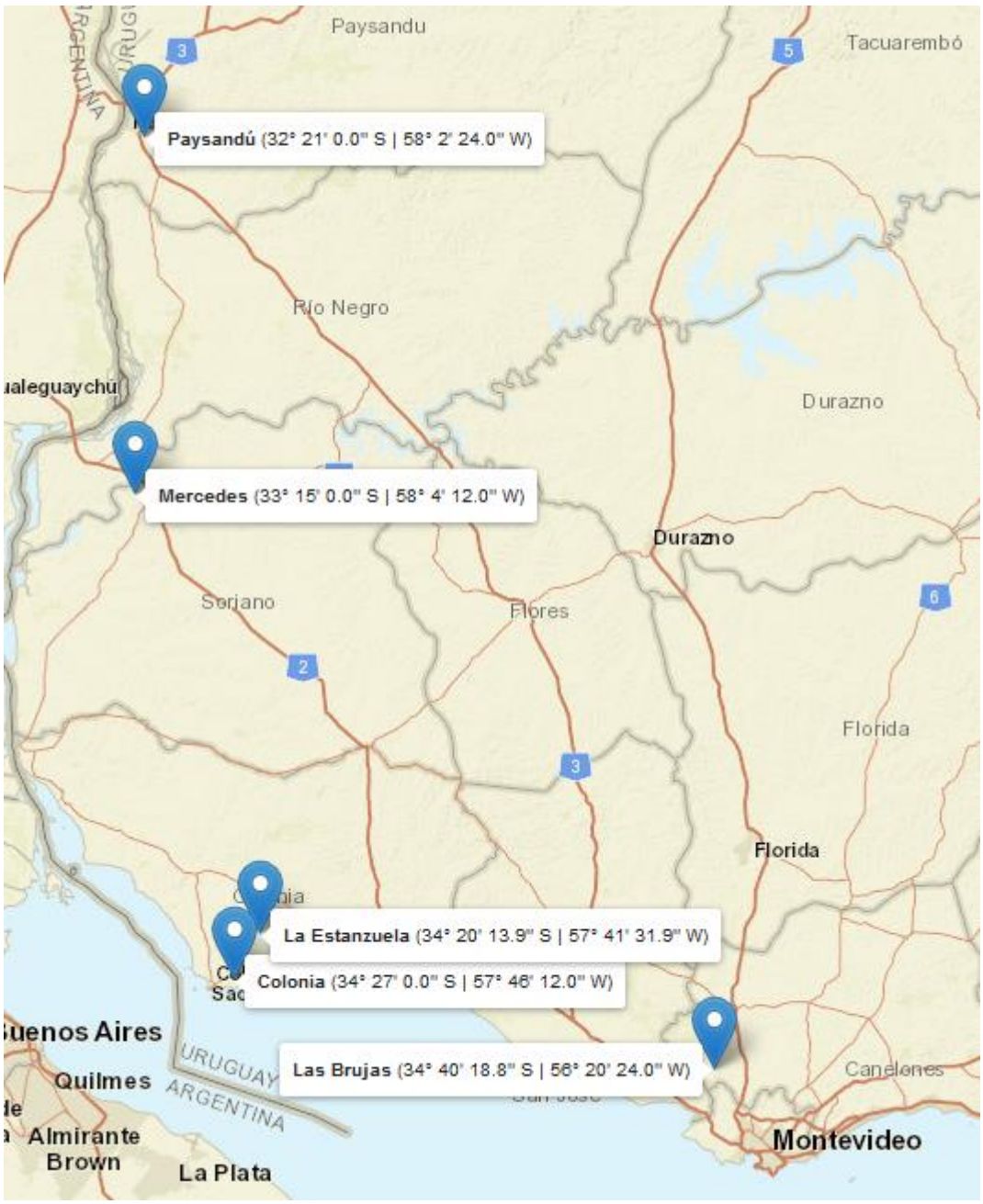
Si bien se dispone de una gran cantidad de información climática para la región del CRC-SAS, la mayoría de ella proviene de sensores automáticos, los cuales cuentan con unos pocos años de datos. Debe tenerse en cuenta que, para caracterizar el clima de una región determinada, es necesario contar con series climáticas largas (de al menos 30 años, según la Organización Climática Mundial) para poder incluir la mayor variabilidad posible.

Esta restricción nos obliga a utilizar datos de estaciones climáticas convencionales, las cuales tienen registros que exceden la cantidad de años necesaria (en general, desde 1961 a la fecha). Sin embargo, las estaciones meteorológicas convencionales carecen de una buena cobertura espacial (son bastante escasas en relación con las estaciones automáticas) y además muchas veces tienen datos faltantes. La selección de las ubicaciones puntuales de estaciones meteorológicas convencionales constituye el primer filtro necesario dentro de la presente tarea.

Por otro lado, para poder realizar las simulaciones de rendimientos con el software DSSAT, también es necesario definir ciertos parámetros asociados al cultivo a simular, su manejo agronómico y el tipo de suelo donde transcurre su ciclo de vida. Esta búsqueda debe ser realizada para aquellas ubicaciones puntuales que hayan sido filtradas a partir de la selección previa (según los registros climáticos).

Hasta el momento, y dado que es el país con menor extensión territorial de los 6 enumerados, Uruguay es el único para el cual se han definido las ubicaciones puntuales en las que se llevará a cabo el estudio

Fig. 2. Mapa de ubicaciones puntuales de Uruguay seleccionadas para el estudio de impactos de sequía. Estas ubicaciones corresponden a estaciones meteorológicas convencionales.




de impactos de sequía. Las ubicaciones seleccionadas corresponden a cinco estaciones meteorológicas convencionales y pueden visualizarse en el mapa de la Fig. 2.

# 4 Generación de series sintéticas de clima

Tal como se ha descripto en la Sección 2, una de las tareas necesarias para poder realizar simulaciones de ciclos de vida de los cultivos, implica la generación de series sintéticas de clima que presenten las mismas propiedades estadísticas que las series históricas observadas. De este modo, es posible

realizar una gran cantidad de simulaciones (con el propósito de aplicar el *método de Montecarlo*) considerando series climáticas que sean plausibles para cada una de las zonas de estudio.

Para llevar a cabo esta tarea se ha utilizado el paquete de R *gamwgen* (https://github.com/CRC-SAS/weather-generator), el cual contiene la implementación de un generador estocástico diario y multisitio de series meteorológicas sintéticas. Este paquete fue originalmente desarrollado como parte de la tesis doctoral del Dr. Andrew Verdin en el marco del proyecto *CNH: From Farm Management to Governance of Landscapes: Climate, Water, and Land-Use Decisions in the Argentine Pampas* (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1211613) financiado por la National Science Foundation de Estados Unidos. Este generador fue posteriormente continuado, optimizado, mejorado y extendido por varios integrantes y colaboradores del Proyecto SISSA (entre los que se incluye este autor).

El generador desarrollado es muy flexible y capaz de generar secuencias de valores diarios de precipitación y temperaturas máxima y mínima. A partir de estas últimas se pueden derivar otras variables como la radiación solar y la evapotranspiración. Esta flexibilidad se sustenta en el uso modelos generalizados aditivos (GAM) que permiten modelar con mucha precisión el comportamiento de las variables meteorológicas y capturar las propiedades estadísticas de los datos observados.

La modelación de las variables meteorológicas se divide en dos: por un lado, la ocurrencia y monto de precipitación y por otro, las temperaturas máxima y mínima. La ocurrencia de precipitación se modela a través de un modelo *probit* mientras que al monto se lo hace a través de una distribución aleatoria *gamma*. Para la temperatura se utiliza un modelo autorregresivo condicionado por la ocurrencia de lluvia. A su vez, estos modelos pueden ser espacialmente correlacionados con campos aleatorios Gaussianos que contemplan la variabilidad espacial y temporal regional.

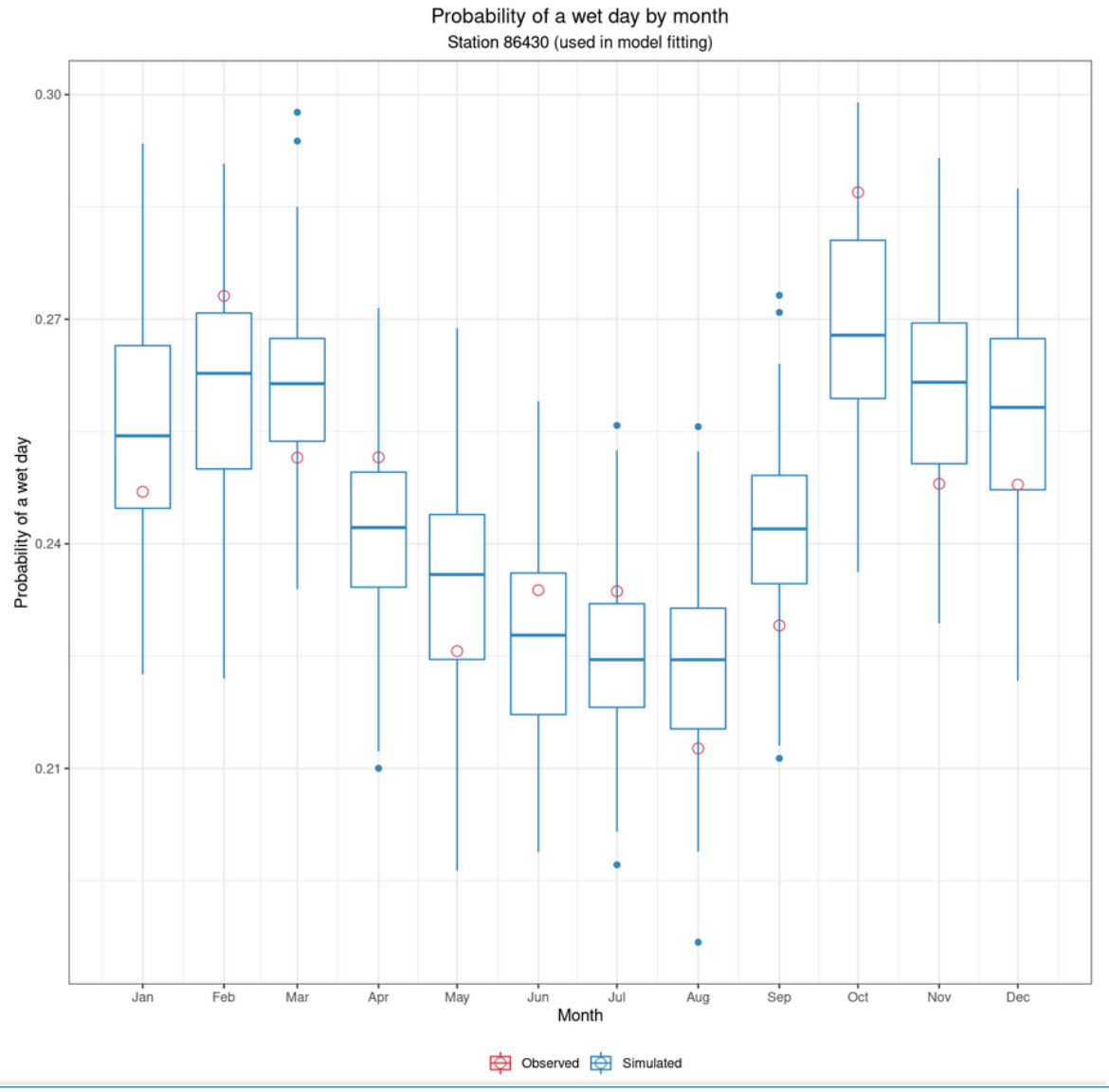
Este generador incluye una serie de diagnósticos estadísticos y gráficos desarrollados para verificar la bondad de ajuste estadística de los GAM y validar que las series sintéticas sean consistentes con los registros históricos. Los diagnósticos son exhaustivos e incluyen todas las propiedades de las series que podrían afectar el desempeño de las mismas durante el análisis probabilista.

Debe considerarse que una de las condiciones necesarias para seleccionar una estación meteorológica como candidata a participar de este estudio, es que las series climatológicas correspondientes sean largas y completas. La manera de validar esta condición es mediante el análisis de los diagnósticos implementados como parte del paquete *gamwgen*.

Al momento de escritura del presente reporte se han generado series sintéticas de clima para las 5 ubicaciones seleccionadas de Uruguay. Para cada una de estas ubicaciones se han utilizado series históricas de referencia con un largo de 60 años (desde el 1 de enero de 1961 hasta el 31 de diciembre de 2020). Para cada una de las series históricas se han generado 100 series sintéticas de clima. Como resultado, cada estación meteorológica cuenta con 6000 años de clima sintético los cuales pueden ser utilizados para simular una numerosa cantidad de ciclos de vida de diversos cultivos.

Para estas series sintéticas se han ejecutado los diagnósticos estadísticos mencionados anteriormente. Estos diagnósticos han sido cuidadosamente analizados para verificar que las series sintéticas generadas fueran estadísticamente consistentes con las series históricas. A continuación, se mostrarán solamente algunos de los diagnósticos analizados a modo de ejemplo (dado que está fuera del alcance de este documento el análisis exhaustivo de todos los resultados). Sin embargo, es posible visualizar la totalidad de los diagnósticos para las 5 estaciones meteorológicas de Uruguay en https://1drv.ms/u/s!As8wkljo8CRlgpBbVOleWcNBUr6hrA?e=qIdRUb.

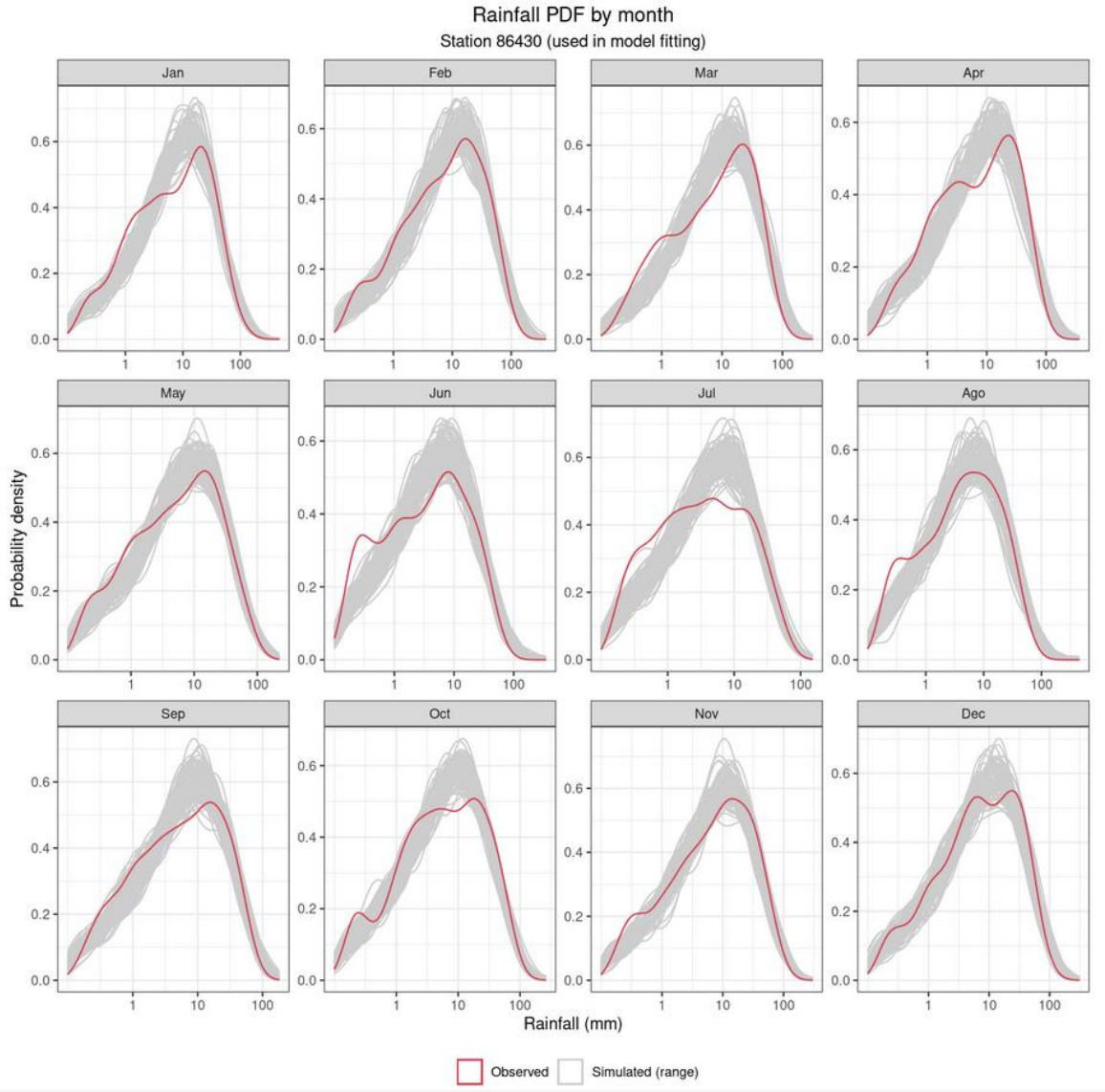
Fig. 3. Comparación entre frecuencia relativa observada de días húmedos (con precipitaciones mayores a 0,1mm) y la frecuencia relativa simulada para las 100 series sintéticas. Debido que la frecuencia observada de días húmedos tiene una dependencia estacional (es menor en meses de otoño e invierno y mayor en meses de primavera y verano), la comparación se ha llevado a cabo para cada mes del año.




En la Fig. 3 se muestra una comparación entre la frecuencia relativa observada de días húmedos (con precipitaciones mayores a 0,1mm) y la frecuencia relativa simulada para las 100 series sintéticas. Debido que la frecuencia observada de días húmedos tiene una dependencia estacional (es menor en meses de otoño e invierno y mayor en meses de primavera y verano), la comparación se ha llevado a cabo para cada mes del año. En general, se observa que los círculos rojos (valores observados) se encuentran dentro o en las proximidades de las cajas de los *boxplots* azules (valores simulados). También se observa que el patrón de variación estacional real es consistente con el patrón estacional simulado.

En la Fig. 4 se presentan doce paneles (uno por cada mes del año) con funciones de densidad de probabilidad (PDF por sus siglas en inglés) asociadas a la precipitación diaria de cada mes. Las líneas

Fig. 4. Comparación entre distribuciones de montos de precipitaciones diarias observadas (líneas rojas) y simuladas (líneas grises) para cada mes del año.

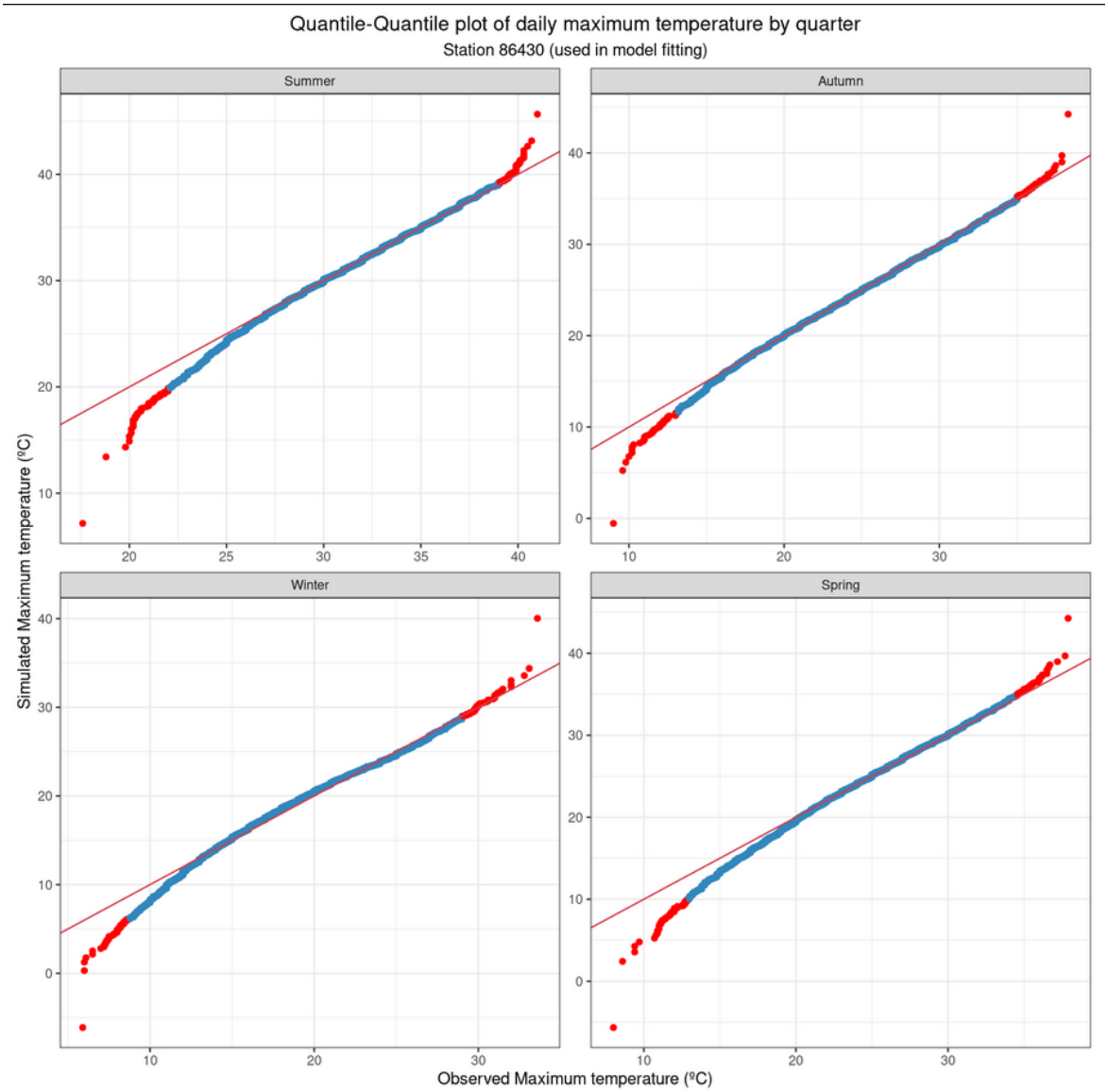



rojas corresponden a las funciones derivadas de datos observados, mientras que las líneas grises están vinculadas a datos simulados. También se observa que las distribuciones de probabilidad de montos diarios de precipitación son consistentes entre los datos históricos y los simulados.

Los próximos dos diagnósticos están vinculados a temperaturas. En la Fig. 5 se presentan cuatro QQ-plots que permiten comparar los percentiles de temperatura máxima diaria para cada estación climática entre los datos históricos observados y los simulados. Los puntos de color rojo están asociados a valores de temperatura máxima correspondientes a los percentiles 1 y 99. Se observa que, en general, las temperaturas más bajas son subestimadas por el generador y las más altas son sobrestimadas. Sin embargo, este comportamiento se observa para percentiles extremos (aunque no en todos los casos).

El último diagnóstico que se presenta en esta sección, corresponde a una combinación entre dos variables. La amplitud térmica diaria está definida como la diferencia entre la temperatura máxima y

Fig. 5. QQ-Plots para temperatura máxima diaria observada y simulada. Cada panel presenta temperaturas correspondientes a cada unas de las estaciones del año. Los puntos de color rojo están asociados a datos asociados a los percentiles 1 y 99.

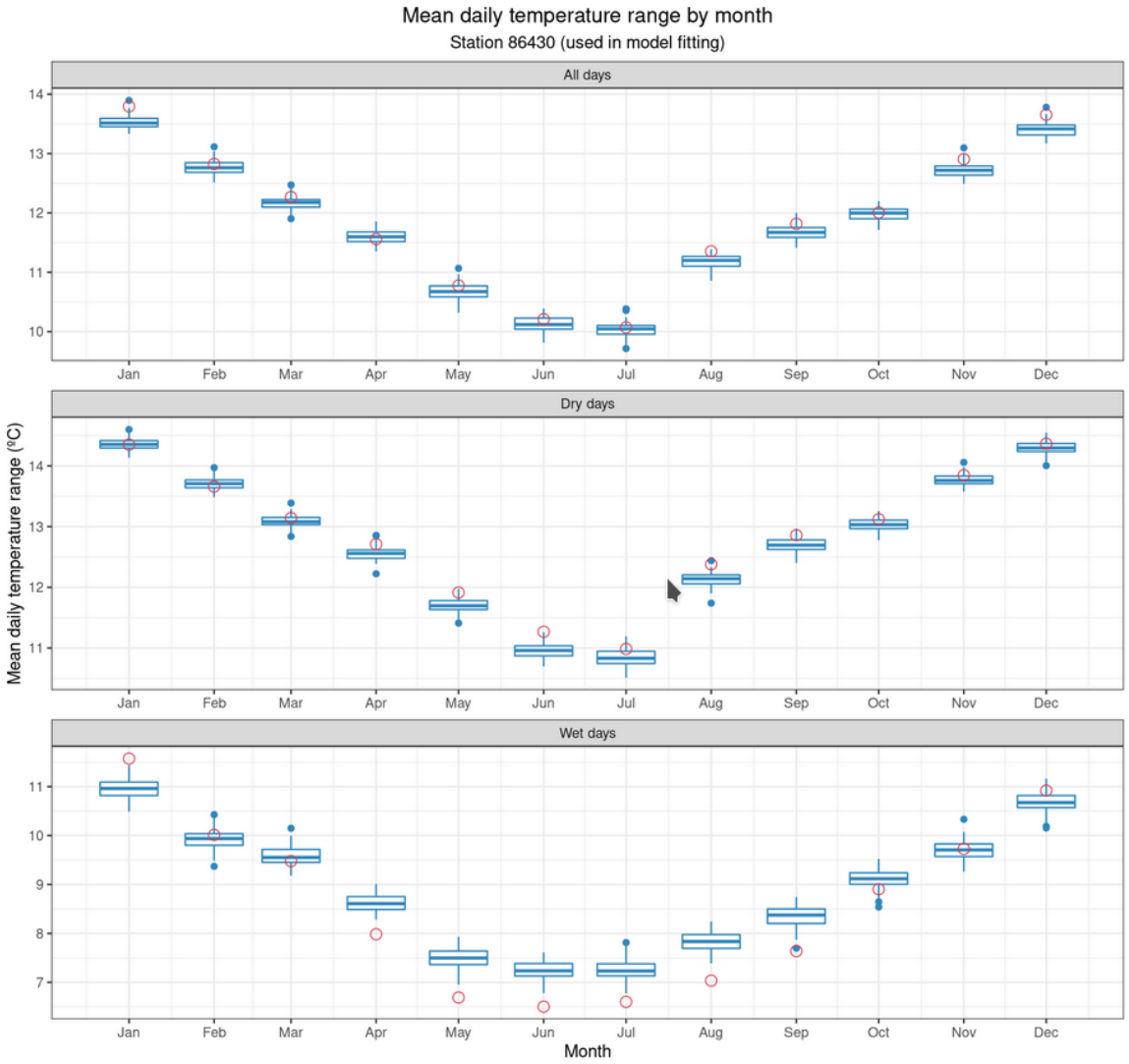



mínima de un día determinado. En la Fig. 6 se observan 3 paneles, cada uno de los cuales permite comparar la amplitud térmica promedio para cada mes del año entre la serie histórica y las series simuladas.

Cada panel muestra el mismo diagnóstico pero basado en conjuntos de datos diferentes. Esto permite realizar el análisis para todos los días (All days), los días secos (Dry days) y los días húmedos (Wet days). Esta discriminación es importante porque la amplitud térmica en días húmedos siempre suele ser menor que en días secos (debido a que la precipitación tiene el efecto de estabilizar la temperatura).

Se observa que, en general, existe consistencia entre el patrón observado y el simulado. Sin embargo, también puede notarse un pequeño sesgo positivo en cuanto a la amplitud térmica observada para los días lluviosos simulados durante la época otoñal-invernal. Es decir, la amplitud térmica simulada

Fig. 6. Comparación entre amplitud térmica promedio para cada mes del año. En el panel superior se muestra la comparación considerando todos los datos observados y generados. En el panel del medio (Dry days), la comparación fue realizada considerando solamente los días secos (sin lluvia) tanto para la series históricas como para las series simuladas. En el panel inferior (Wet days), la comparación fue realizada considerando solamente los días húmedos (con lluvia).




es moderadamente superior a la observada para esos meses del año. Se ha encontrado que este sesgo es sistemático, y constituye uno de los puntos a mejorar en futuras implementaciones del generador.

A pesar de algunas leves o moderadas diferencias presenten en algunos diagnósticos (debe tenerse en cuenta que hay aproximadamente 35 a 40 diagnósticos por estación), los datos simulados son razonablemente consistentes con los observados y pueden ser utilizados para la generación de rendimientos con DSSAT. Estos análisis también deberán ser llevados a cabo para las series climáticas que se generen para las demás estaciones meteorológicas que se consideren para el presente estudio de impactos de sequía.

# 5 Parametrización de tipos de suelo y manejos de cultivos

Para poder simular rendimientos de cultivos usando el software DSSAT, es necesario contar con datos de entrada que caractericen las siguientes condiciones:

- Condiciones climáticas: estos datos corresponden a las series temporales de clima sintético que fue explicado previemente en la Sección 4.
- Condiciones ambientales: esta información corresponde a las características de los suelos existentes en las ubicaciones (y sus inmediaciones) donde se simularán los ciclos de vida del cultivo.
- Condiciones de los sistemas de producción: estas condiciones caracterizan los cultivos cuyo ciclo de vida se simulará y los manejos agronómicos asociados a dichos cultivos.

En la sección previa se detalló el mecanismo para generar las series temporales de clima sintético que se usarán como datos de entrada para las simulaciones. Estas series sintéticas son de naturaleza estocástica y permiten modelar con cierta exhaustividad las condiciones climáticas que puedan afectar a las zonas de estudio. Por el contrario, las condiciones ambientales y de sistemas de producción son determinísticas y son representativas de cada zona en particular.

Las condiciones ambientales caracterizan los tipos de suelo que se encuentran en las regiones seleccionadas para las simulaciones. Dentro del campo de la agronomía, se suele caracterizar los suelos de distintas regiones a través de unidades cartográficas. Las unidades cartográficas son polígonos que reúnen los suelos más representativos de una zona. Por ejemplo, una unidad cartográfica determinada puede constar de un 20% de suelo A, un 30% de suelo B y un 50% de suelo C. Este modelo permite representar las características del suelo de forma macroscópica.

A su vez, cada tipo de suelo está caracterizado por las propiedades de sus *horizontes*. Los horizontes son estratos de distinta profundidad, los cuales tiene ciertas características comunes. Por ejemplo, un horizonte puede estar definido para profundidades de 0 a 20 centímetros y otro horizonte por profundidades de 20 a 40 centímetros. Cuanta más información de horizontes haya disponible para un tipo de suelo, mayor será su grado de representatividad.

Para poder ejecutar las simulaciones de los ciclos de vida de los cultivos en DSSAT, es necesario caracterizar los tipos de suelo de cada ubicación según las propiedades de sus horizontes. Disponer de información completa y precisa para los tipos de suelo es crucial para que las simulaciones derivadas devuelvan resultados apropiados. Algunas de las propiedades que son necesarias para caracterizar los horizontes de cada tipo de suelo, son:

- Porcentaje de limo, arcilla y arena (a partir de esta información se pueden derivar otras propiedades importantes como *capacidad de almacenamiento de agua* y *ambiente para desarrollo radical*).
- Porcentaje de materia orgánica
- Otras propiedades químicas (pH, capacidad de intercambio catiónico, base de cambio, etc.)

También se mencionó previamente que es necesario caracterizar los sistemas de producción a simular. Los sistemas de producción comprenden aquellas características asociadas a los cultivos cuyo ciclo de vida se va a modelar y los manejos agronómicos asociados a dichos cultivos. Las propiedades que son necesarias caracterizar son las siguientes:

- Cultivos: es necesario proveer información acerca de los aspectos del desarrollo (características de la fenología de cada cultivo y que permiten modelar las distintas etapas de

su ciclo de vida) y del rendimiento (por ejemplo, el número y peso de los granos, la biomasa, la cantidad de hojas, etc.).

- Manejos: caracterizan el sistema de producción aplicado al cultivo (fecha de siembra, densidad de siembra, espaciamiento entre surcos, uso de fertilizante, etc.)

Hasta el momento, se cuenta con información completa para Uruguay y se está trabajando en recopilar los mismos datos para las demás regiones que se encuentran en el proceso de selección descripto en la Sección 3.

# 6 Simulación de rendimientos de cultivos

Como se observa en la Fig. 1, para poder simular los rendimientos de los cultivos, es necesario realizar una parametrización de los tipos de suelos y manejos utilizados en las ubicaciones definidas (y a definir) para este estudio. Para ello es necesaria la coordinación con los ingenieros agrónomos que forman parte del Proyecto SISSA, dado que tal actividad requiere del conocimiento de expertos en esta área.

También es necesario contar con series temporales de precipitación, temperatura (mínima y máxima) y radiación solar diaria para cada una de dichas ubicaciones (por requerimiento del software que simula los rendimientos). Sin embargo, debe recordarse que el generador de clima sintético solamente es capaz de generar series temporales de temperatura y precipitación, pero no de radiación solar.

La radiación solar puede calcularse haciendo uso de varios modelos, de diversa complejidad y precisión. Pero debe tenerse en cuenta, que los modelos más precisos requieren de información sobre nubosidad, heliofanía u otras variables meteorológicas cuya información no está disponible a partir de las series climáticas generadas. Por tal motivo, se ha seleccionado el modelo de radiación de Bristow-Campbell (Bristow K. L., Campbell G. S., 1984), el único que permite estimar radiación solar a partir de valores resultantes del proceso de generación estocástica (temperaturas máxima y mínima).

Este modelo además incluye ciertas constantes paramétricas que pueden ser modificadas mediante un proceso de calibración a partir de valores de radiación medidos *in situ*. La calibración de estas constantes deberá ser realizada, en general, para todas las zonas agrícolas donde se vayan a ejecutar los procesos de generación estocástica de series climáticas.

Esta calibración solamente ha sido llevada a cabo para las ubicaciones definidas para Uruguay, permitiendo, de este modo, el cálculo de las series de radiación diaria asociadas a las series temporales de precipitación y temperaturas generadas estocásticamente. Este mismo proceso deberá realizarse para las zonas agrícolas de Argentina, Brasil, Chile y Paraguay.

Una vez que se puedan calcular las series temporales de radiación diaria y se hayan parametrizado los tipos de suelos y manejos típicos de cada una de las zonas agrícolas, se podrá comenzar la ejecución masiva de procesos de simulación (uno por cada año calendario simulado para cada ubicación). Para ello, deberá considerarse la implementación de un proceso que permita paralelizar la ejecución haciendo uso de múltiples nodos, con el fin de disminuir considerablemente el tiempo de procesamiento. Hasta el momento no se ha realizado ninguna simulación de rendimiento, pero el autor se encuentra activamente trabajando en esta tarea.

# 7 Caracterización de series sintéticas de clima

Para llevar adelante un proceso de aprendizaje automático que permita vincular condiciones de sequía con rendimientos de cultivos, se deberán efectuar transformaciones sobre las series de clima simulado de modo que las mismas permitan caracterizar eventos o condiciones de sequía. Este proceso requiere que se puedan cuantificar las condiciones de sequía mediante ciertos indicadores numéricos. Por lo tanto, en primera instancia debe indicarse qué se entiende por sequía.

No existe una única definición de sequía, debido a que este fenómeno se identifica por sus efectos o impactos sobre diferentes tipos de sistemas (agricultura, recursos hídricos, ecosistemas, economía, etc.). Los principales tipos de sequías son:

- Meteorológica: escasez de precipitación. Este tipo se sequías es el causante de otro tipo de sequías;
- Agrícola: escasez de agua para satisfacer las necesidades de un cultivo (es el tipo de sequía que se abordará en el presente estudio);
- Hidrológica: deficiencia de la disponibilidad de agua de superficie y/o subterránea. Se desarrolla más lentamente, debido a que hay un retraso entre la falta de lluvia y la reducción de agua en arroyos, ríos, lagos, embalses, etc.; y
- Socioeconómica: escasez hídrica con consecuencias sociales y económicas desfavorables. Es una consecuencia de los otros tipos de sequía y es claramente económica.

Los *indicadores de sequía* son variables o parámetros utilizados para describir las condiciones de las sequías. Cabe citar, por ejemplo, la precipitación, la temperatura, los caudales fluviales, los niveles de las aguas subterráneas y de los embalses, la humedad del suelo y el manto de nieve.

Los *índices de sequía* son medidas cuantitativas que se utilizan para caracterizar los niveles de sequía mediante la asimilación de uno o más indicadores de sequía. Suelen ser representaciones numéricas informatizadas de la gravedad de las sequías, determinadas mediante datos climáticos o hidrometeorológicos, entre los que se incluyen los indicadores enumerados. Tienen por objeto analizar el estado cualitativo de las sequías en el entorno en un período de tiempo determinado. Desde el punto de vista técnico, los índices también son indicadores.

Debido a la magnitud de sus impactos en la región, una de las principales líneas de trabajo del Centro Regional del Clima para el sur de América del Sur (CRC-SAS) ha sido el desarrollo de un sistema de vigilancia de sequías, para lo cual se ha implementado el cálculo de algunos índices de sequía, de los cuales pueden destacarse el SPI y SPEI por su uso e importancia.

## 7.1 Índice de Precipitación Estandarizado (SPI)

El Índice de Precipitación Estandarizado (IPE, o SPI por sus siglas en inglés) cuantifica las condiciones de déficit o exceso de precipitación en un lugar y para una escala determinada de tiempo. El SPI fue desarrollado por McKee et al. (1993) con la finalidad de mejorar la detección del inicio y el monitoreo de la evolución de las sequías meteorológicas (definidas únicamente en función de la precipitación). La principal ventaja de este índice es que su cálculo requiere una única variable climática para el cálculo: la precipitación. El SPI ha sido utilizado ampliamente a nivel mundial y es recomendado por la Organización Meteorológica Mundial.

El primer paso para el cálculo del SPI es el cálculo de los totales acumulados de precipitación correspondientes al mes/año y escala temporal deseada (ver discusión de escalas temporales más abajo). Luego, se ajusta una distribución teórica a los totales de cada mes en el período de referencia. Para ellos se ha utilizado la distribución *gamma*, debido a que ajusta adecuadamente las distribuciones empíricas de totales de precipitación para la mayoría de los meses y estaciones

consideradas, y además porque requiere solamente dos parámetros para su especificación: *alfa* (α, parámetro de forma) y *beta* (β, parámetro de escala).

Los parámetros estimados se usan luego para calcular el percentil correspondiente a los distintos valores de precipitación acumulada cuyo índice se desea calcular. Finalmente, los valores de SPI resultantes se obtienen a partir de los cuantiles correspondientes a estos percentiles para una distribución normal estandarizada (con media = 0 y desvío estándar = 1).

## 7.2 Índice de Precipitación – Evapotranspiración Estandarizado (SPEI)

El Índice de Precipitación – Evapotranspiración Estandarizado (IPEE o SPEI por sus siglas en inglés), es un índice cuyo cálculo es similar al del SPI, pero incorpora el efecto de la evapotranspiración (es decir, la demanda atmosférica de agua) que influye en las condiciones de sequía. El SPEI fue desarrollado por Vicente-Serrano et al. (2010), pero ya ha sido utilizado para analizar distintas características de la sequía, como ser su variabilidad, impactos y mecanismos atmosféricos que la producen (Beguería et al., 2013; Hernandez and Uddameri, 2013; Vicente-Serrano et al., 2015; Xu et al. 2015).

El SPEI utiliza como valor de entrada al balance hídrico (o sea, la diferencia entre precipitación y evapotranspiración potencial o PET por sus siglas en inglés). El cálculo de la evapotranspiración potencial es complicado, debido a que involucra muchos parámetros (temperatura, humedad del aire, viento y radiación, entre otros). Para el Proyecto SISSA se ha utilizado la ecuación de Hargreaves-Samani (1985), que es eficiente en el cálculo de la evapotranspiración potencial utilizando sólo medias mensuales de temperatura máxima y mínima, y radiación solar.

Dado que el balance entre precipitación y evapotranspiración puede tomar valores negativos, el ajuste del mismo se realiza con la distribución teórica *log-logística*, que acepta valores nulos y negativos. El método para ajustar estos parámetros a la distribución es el de máxima verosimilitud (Beguería et al., 2013).

## 7.3 Cálculo de índices para distintas escalas de tiempo

Los efectos de las sequías se manifiestan en diferentes escalas temporales, ya que las respuestas de diferentes sistemas hidrológicos y biológicos a las anomalías de precipitación varían mucho (Ji and Peters, 2003). Es decir, puede haber grandes diferencias en la duración de los déficits hídricos necesarios para causar impactos negativos en diferentes sistemas.

Como la sequía es un fenómeno multiescalar, es necesario el uso de indicadores que puedan capturar adecuadamente las escalas temporales relevantes para detectar impactos negativos sobre los diferentes sistemas de interés. Por ejemplo, la sequía agrícola – definida como la escasez de agua para satisfacer las necesidades de un cultivo – puede ser bien representada por las escalas de 2 y 3 meses, mientras que déficits en los caudales de río o arroyos se reflejan mejor por medio de las escalas de 3 a 6 meses. Asimismo, se han encontrado asociaciones entre la variación del nivel de la napa freática y los valores de los índices con escalas de 6 a 24 meses.

Los índices de sequía SPI y SPEI son calculados para múltiples escalas temporales: 1, 2, 3, 6, 9, 12, 18, 24, 36 y 48 meses. Es decir, los índices se basan en series de precipitaciones acumuladas para cada escala. Por ejemplo, para calcular un índice con una escala de 6 meses para junio de 2003, se considera la suma de los valores mensuales de precipitación desde enero hasta junio de 2003; este valor se pone en el contexto de los totales de lluvia para enero-junio en el registro histórico.

De la misma forma, se puede realizar el cálculo para otras escalas temporales. Para el cálculo del SPEI, que requiere temperaturas máximas y mínimas para estimar la evapotranspiración, se usa el promedio de estas temperaturas para cada escala temporal. Siguiendo con el ejemplo anterior, el cálculo de SPEI para una escala de 6 meses para junio de 2003 utilizará los promedios mensuales de temperaturas máximas y mínimas desde enero hasta junio de 2003.

Para el cálculo de los diferentes índices de sequía, generalmente es necesario estimar parámetros o cuantiles de la distribución empírica o teórica de precipitación. Esta estimación se puede repetir cada vez que se agregan datos nuevos a las series, es decir, al actualizar la base de datos del CRC-SAS. La desventaja de este enfoque es que los valores previos de un índice cambian a medida que las series se extienden, ya que cada vez el ajuste o estimación se hace con series diferentes (más largas).

Una alternativa para evitar que los valores anteriores cambien constantemente es utilizar un *período de referencia* fijo para estimar parámetros o cuantiles necesarios. El uso de un período de referencia implica que, al agregar más registros a las series climáticas, los valores anteriores no cambiarán ya que los índices se calculan a partir de parámetros o cuantiles estimados para el período de referencia que no se modifica.

Es deseable que el período de referencia contenga la mayor parte de la variabilidad de las series climáticas. La ventaja de contar con registros largos es que permiten capturar oscilaciones de baja frecuencia (épocas secas y húmedas asociadas con la variabilidad climática multidecadal). Los registros más cortos, en cambio, podrían contener sólo parte de esa variabilidad, y los valores de un índice podrían estar sesgados por el uso de un período de referencia seco o húmedo. Para todos los índices de sequía calculados (SPI y SPEI), se ha utilizado el período 1971-2010 (40 años) como período de referencia.

Este período de referencia y las funciones de densidad estimadas en base a datos históricos pueden ser utilizados como base de cálculo para los índices derivados de series sintéticas de clima. Esto es posible debido a que las series sintéticas de clima son generadas de modo de tener las mismas propiedades estadísticas que las series históricas originales.

## 7.4 Implementación de cálculo frecuente de índices

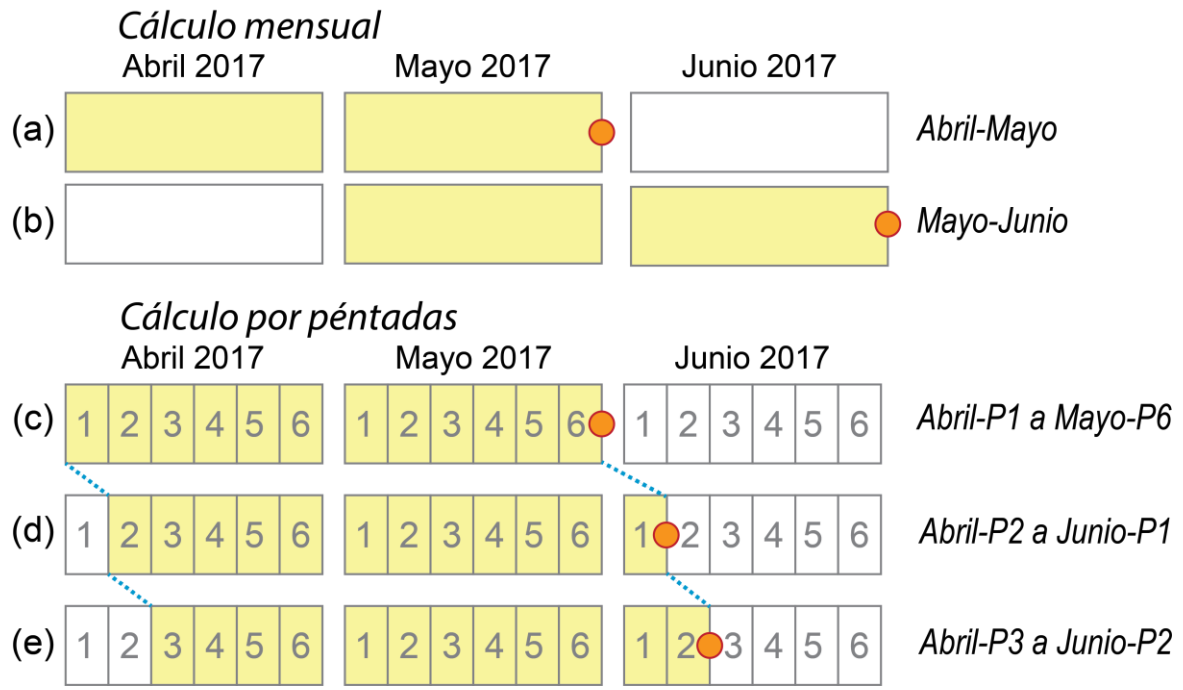
Dada la importancia que tienen los índices presentados para la vigilancia de la sequía, el CRC-SAS ha implementado un mecanismo para actualizar los valores de dichos índices con frecuencia mayor a una vez por mes (que es la frecuencia estándar que suele manejarse para el cálculo de estos índices). Este mecanismo implica el cálculo de valores de SPI-*n* o SPEI-*n* (donde *n* denota la escala temporal) a cierto intervalo de tiempo regular dentro de un mismo mes.

Para definir este intervalo de tiempo regular se hace uso del concepto de *péntada*. Una péntada en climatología denota un intervalo de tiempo de 5 días. La actualización de los valores de un índice por péntada, implica que en un mes de 30 días se actualizará 6 veces el valor de dicho índice. Las péntadas de un mes se consideran, de forma arbitraria, como los períodos que abarcan los días 1-5, 6-10, 11-15, 16-20, 21-25 y del 26 al final del mes. De este modo, las primeras 5 péntadas de un mes tienen todas 5 días, a excepción de la última que puede tener 3, 4, 5 o 6 días (dependiendo de la cantidad de días del mes). Esta última excepción permite que todos los meses tengan 6 péntadas completamente incluidas dentro del mismo.

En la Fig. 7 se ilustra el mecanismo de cálculo del índice SPI-2 con frecuencia de actualización mensual (panel superior) y pentadal (panel inferior). Debe destacarse que la escala temporal es siempre invariante (en el caso del ejemplo, 2 meses). Esta escala temporal constituye una ventana móvil de ancho fijo que se va corriendo en el tiempo según la frecuencia de actualización.

Con este mecanismo es posible derivar series temporales de índices de sequía a partir de las series temporales originales de temperaturas y precipitación. Esta transformación constituye el primer paso hacia la caracterización de las condiciones de sequía asociadas a las series climáticas que dan origen a los rendimientos de los cultivos. Sin embargo, estos índices por sí solos nos son suficientes para tal fin,

Fig. 7. Ilustración del cálculo de índices de sequías para meses calendario (filas a y b) y para péntadas filas c-e). Los datos utilizados para el cálculo de un índice de sequía con escala de dos meses se indican con amarillo en la figura. Los círculos naranja en cada fila indican el momento en que el valor del índice está disponible. En el cálculo por mes calendario, los índices se calculan una vez por mes, después de transcurrido por completo el último mes en el período a considerar (ej., el valor del índice para abril-mayo 2017 se calcula a fines de mayo 2017; el siguiente valor se calculará a fines de junio 2017.




debido a que existen otras características de las condiciones de sequía que deben ser tenidas en cuenta: la duración y su internsidad o magnitud.

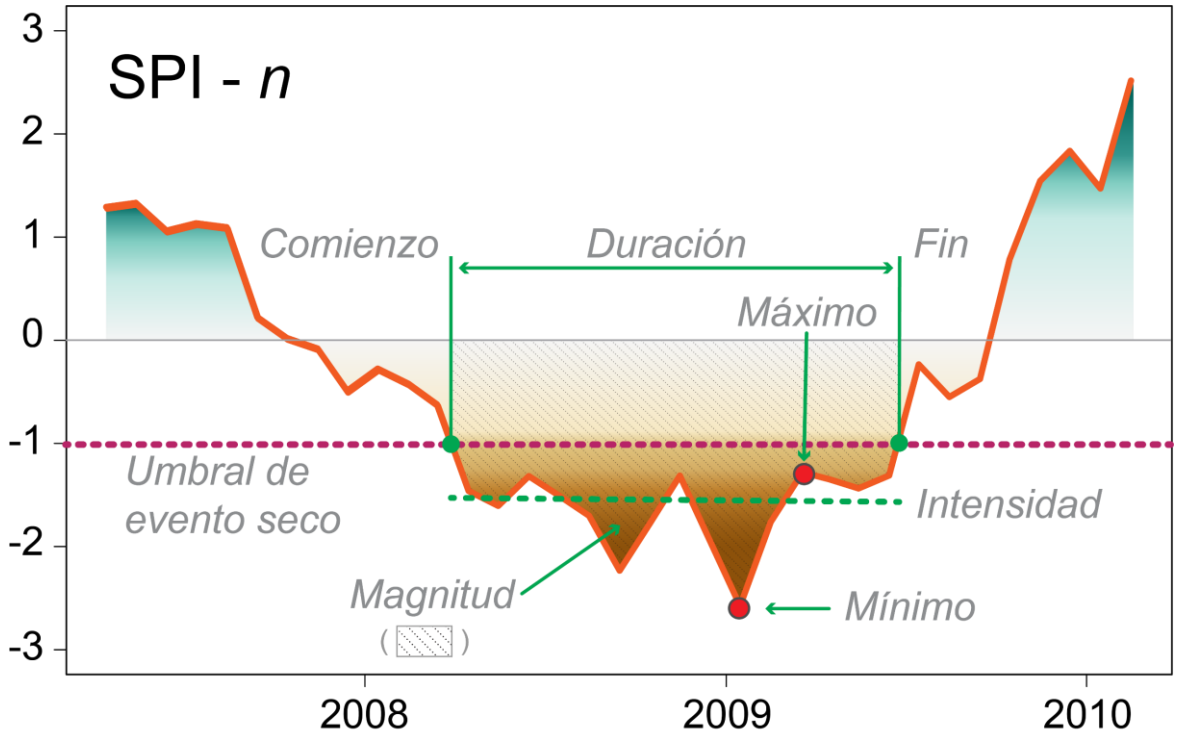
## 7.5 Determinación de eventos de sequía

Como se detalló en el Plan de Trabajo, resulta de interés estudiar el impacto de la sequía en relación con las características de los eventos y el momento de ocurrencia. Es sabido que la duración de una sequía tiene efectos diferentes (no es lo mismo una sequía corta que una larga), así como también su intensidad (sequía moderada vs. sequía excepcional) y su momento de ocurrencia (etapa de floración vs. etapa previa a la cosecha). Por este motivo, las características listadas previamente deben ser reflejadas de algún modo.

El momento de ocurrencia de la condición de sequía actualmente ya está representada por la escala del índice y la péntada de fin. Siguiendo el ejemplo de la Fig. 7, el valor de SPI-2 para P15 denota con precisión que el valor del índice está vinculado a las condiciones climáticas presentes entre las péntadas 4 a 15 del año inclusive (es decir, del 16 de enero al 15 de marzo). Sin embargo, el valor del índice por sí mismo no incluye información de la persistencia de una sequía ni de su magnitud durante el evento.

Es por ello que resulta necesario definir lo que se entiende por evento de sequía y las características de los mismos. El CRC-SAS define como evento de sequía al intervalo de tiempo durante el cual un valor de índice determinado (para una escala temporal específica) se encuentra por debajo de cierto umbral. Por ejemplo, dado que en general se considera que valores de SPI de -0,5 están en el umbral

Fig. 8. Definición de evento de sequía basado en un índice (en este caso, SPI). Un evento de sequía comienza cuando el valor del índice toma un valor por debajo de cierto umbral (dado por un parámetro). La duración del evento estará dada por el intervalo de tiempo durante el cual el índice mantenga su valor por debajo de ese umbral. Además de la duración, un evento de sequía está caracterizado por su intensidad (promedio de valores del índice) o magnitud (suma de los valores del índice).




de condiciones de sequía, se podría definir como un tipo de evento seco a aquellos intervalos de tiempo donde el SPI-3 está por debajo de -0,5. Debe reiterarse que la definición de evento seco es parametrizable respecto del umbral.

En la Fig. 8 se ilustra una posible serie temporal de SPI-*n*. Se observa que para el ejemplo de la figura se ha definido como umbral el valor -1. Luego, el evento seco asociado a este valor de umbral toma lugar durante el intervalo de tiempo en que el SPI-*n* está por debajo de -1. La extensión de dicho intervalo de tiempo se denomina *duración*. Esta propiedad permite caracterizar la extensión de un evento en el tiempo.

Para medir la intensidad de un evento se utilizan de forma indistinta las propiedades *intensidad* o *magnitud*. La intensidad de un evento denota el promedio de los valores de índice durante la vigencia de dicho evento. Por otro lado, la magnitud de un evento denota la suma de los valores del índice durante dicho evento. Como se observa, la magnitud es una variable que está asociada tanto a la intensidad como a la duración (es la multiplicación de ambas).

A partir de la identificación de eventos basados en índices de sequía será posible generar series temporales de "condiciones de sequía" que denoten los momentos de ocurrencia, su duración e intensidad. Estas series temporales deberán construirse en formato tabular ancho, de modo que cada uno de los valores represente un atributo que caracterice a una serie temporal (que será una fila de la tabla). A cada fila o serie temporal se le asociará el rendimiento resultante para un cultivo en una ubicación determinada. Este rendimiento podrá luego ser normalizado o categorizado según sea necesario para mejorar los resultados del proceso de aprendizaje.

# 8 Vinculación de condiciones de sequía con rendimientos

Una vez que se cuente con un conjunto de datos organizado de forma tabular (donde los atributos correspondan a condiciones de sequía en distintos momentos del cultivo y la variable dependiente a rendimientos resultantes) será posible llevar a cabo procesos de aprendizaje automático para encontrar patrones vinculantes entre condiciones de sequía y rendimientos de cultivos.

Si bien la ejecución de esta tarea aún está lejos de comenzarse, se ha iniciado la investigación acerca del estado del arte de esta área de estudio. En particular, se ha comenzado por (van Klompenburg, et al., 2020). Este *paper* consta de una revisión sistemática de 50 trabajos de investigación en esta temática. A partir de este trabajo (y de los otros trabajos a los que se hace referencia), se buscarán las alternativas óptimas para llevar a cabo esta tarea. Por el momento, solamente se puede adelantar que la gran mayoría de los trabajos de *machine learning* sobre rendimientos de cultivos, se han llevado a cabo mediante regresiones y clasificaciones basadas en redes neuronales profundas o el algoritmo *XGBoost*.

# 9 Próximas actividades

Al momento de la escritura del presente reporte el autor se encuentra trabajando en la implementación del código fuente en R que permita simular de forma programática los rendimientos asociados a los cultivos y manejos seleccionados para Uruguay. Una vez finalizada esta actividad, se continuará con la construcción de un conjunto de datos tabular como el detallado en la Sección 7. Posteriormente, se comenzarán las actividades asociadas a la tarea de la Sección 8 (vinculación de condiciones de sequía con rendimientos de cultivos). Este proceso deberá ser replicado para las ubicaciones seleccionadas en otros países.

# Referencias


Beguería, S., S. M. Vicente-Serrano, F. Reig, and B. Latorre (2013). *Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring.* International Journal of Climatology, 34, 3001-3023. http://dx.doi.org/10.1002/joc.3887.

Bristow K. L., Campbell G. S. (1984). *On the relationship between incoming solar radiation and daily maximum and minimum temperature*. Agricultural and Forest Meteorology, Volume 31, Issue 2, 159-166. https://doi.org/10.1016/0168-1923(84)90017-0.

Hargreaves, G. L., and Z. A. Samani (1985). *Reference crop evapotranspiration from temperature.* Applied Engineering in Agriculture, 1, 96-99. http://dx.doi.org/10.13031/2013.26773.

Hernandez, E. A., and V. Uddameri (2013). *Standardized precipitation evaporation index (SPEI)-based drought assessment in semi-arid south Texas*. Environmental Earth Sciences, 70(5), 1-11, doi:10.1007/s12665-013-2897-7. https://doi.org/10.1007/s12665-013-2897-7.

McKee, T. B., N. J. Doesken, and J. Kleist (1993). *The relationship of drought frequency and duration to time scales*. Eighth Conference on Applied Climatology, edited, pp. 179-184, American Meteorological Society, Anaheim, California.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno (2010). *A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index*. Journal of Climate, 23(7), 1696-1718. https://doi.org/10.1175/2009JCLI2909.1.

Vicente-Serrano, S. M., O. Chura, J. I. López-Moreno, C. Azorin-Molina, A. Sanchez-Lorenzo, E. Aguilar, E. Moran-Tejeda, F. Trujillo, R. Martínez, and J. J. Nieto (2015). *Spatio-temporal variability of droughts in Bolivia: 1955–2012*. International Journal of Climatology, 35(10), 3024-3040. https://doi.org/10.1002/joc.4190.

van Klompenburg T., Kassahun A., Catal C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, Volume 177. https://doi.org/10.1016/j.compag.2020.105709.

Xu, K., D. Yang, H. Yang, Z. Li, Y. Qin, and Y. Shen (2015). *Spatio-temporal variation of drought in China during 1961–2012: A climatic perspective*. Journal of Hydrology, 526, 253-264. https://doi.org/10.1016/j.jhydrol.2014.09.047.