Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería Agrícola
Programa de Doctorado en Recursos Hídricos y Energía para la Agricultura

Marco probabilístico para la estimación de evapotranspiración en huertos frutales: integración de modelos físicos y estructuras de error espaciotemporales (A probabilistic framework for evapotranspiration estimation in fruit orchards: integrating physical models and spatiotemporal error structures)

Tesis para optar al grado de Doctor en Recursos Hídricos y Energía para la Agricultura

LORENZO ESTEBAN CIGARRA GUÍÑEZ
CHILLÁN-CHILE
2026

Profesor Guía: Octavio Lagos Roa
Dpto. de Recursos Hídricos
Facultad de Ingeniería Agrícola
Universidad de Concepción

Esta tesis ha sido realizada en el Departamento de Recursos Hídricos de la Facultad de Ingeniería Agrícola, Universidad de Concepción.

Profesor Guía

Dr. Octavio Lagos Roa
Facultad de Ingeniería Agrícola
Universidad de Concepción

Comisión Evaluadora:

Dr. Mario Lillo Saavedra
Facultad de Ingeniería Agrícola
Universidad de Concepción

Dr. Sebastián Krogh Navarro
Facultad de Ingeniería Agrícola
Universidad de Concepción

Dr. Claudio Balbontín Nesvara
Centro de Investigación Intihuasi
Instituto de Investigaciones Agropecuarias

Director de Programa (S)

Dr. José Luis Arumí Ribera
Facultad de Ingeniería Agrícola
Universidad de Concepción

A mi familia,

por su apoyo incondicional y amor constante.

AGRADECIMIENTOS

Primero que todo, deseo expresar mi más sincero agradecimiento al Dr. Octavio Lagos, profesor guía de esta tesis, y al Dr. Daniele Zaccaria, cuyo apoyo conjunto hizo posible el desarrollo de este trabajo.

También extiendo mi gratitud al director de programa y miembro de mi comisión, Dr. Sebastián Krogh, por sus enseñanzas y muy valiosos aportes, y a los miembros de mi comisión de tesis, Dr. Mario Lillo y Dr. Claudio Balbontín, por sus observaciones, sugerencias y compromiso con la mejora continua de esta investigación. Agradezco eternamente a Clarita por todo su apoyo y orientación excepcional durante todos estos años.

Deseo expresar un agradecimiento muy especial al Dr. Pasquale Steduto, por su liderazgo, orientación científica y apoyo constante durante todo el proceso de esta investigación.

Asimismo, agradezco a Kristen Shapiro por su colaboración, guía y dedicación durante los trabajos de campo y la recopilación de información en terreno, los cuales fueron esenciales para el desarrollo práctico de esta tesis.

Extiendo también mi agradecimiento al Departamento de Land, Air and Water Resources de la University of California, Davis, por su disposición, recursos y apoyo institucional para la realización de esta tesis.

Del mismo modo, hago extensivo mi reconocimiento al Dr. Eduardo Holzapfel y

al Dr. Diego Rivera por sus enseñanzas, confianza y respaldo a lo largo de este proceso.

Este trabajo fue posible gracias al apoyo de la Agencia Nacional de Investigación y Desarrollo (ANID), a través de la Subdirección de Capital Humano/Doctorado Nacional 2022–21221319, que respaldó mis estudios de posgrado; así como del proyecto “Chileflux: una red científica para el monitoreo de flujos de agua, energía y CO_2 como herramienta para el control, adaptación y mitigación del cambio climático y la sequía en Chile” ANID/FSEQ210019, el Consorcio Tecnológico del Agua CoTH_2O CORFO/20CTECGH145896 y ANID/CTI250001, y el Centro de Recursos Hídricos para la Agricultura y la Minería (CRHIAM) ANID/FONDAP/1523A0001.

Finalmente, agradezco profundamente a la Facultad de Ingeniería Agrícola de la Universidad de Concepción por su constante apoyo académico y humano, y a todas las personas que me acompañaron con su buena disposición, energía y aliento durante este camino, contribuyendo a hacer de esta experiencia un proceso profundamente enriquecedor y gratificante.

ÍNDICE GENERAL

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

La estimación de la evapotranspiración (ET) es fundamental para la gestión del agua en sistemas agrícolas bajo presión hídrica creciente. Las plataformas satelitales disponibles operativamente presentan sesgos sistemáticos en huertos frutales y reportan exclusivamente estimaciones puntuales sin cuantificación de incertidumbre, lo que limita la capacidad de usuarios y gestores para evaluar la confiabilidad de los productos disponibles. Para abordar esta brecha, se desarrolló un marco metodológico probabilístico en dos fases orientadas a construir un ensamble con diversidad estructural. El modelo SEB-PV (desarrollado exclusivamente para superficies parcialmente vegetadas) fue identificado como candidato complementario a los seis algoritmos de la plataforma OpenET, pero requería superar barreras operacionales. La primera fase evaluó su viabilidad utilizando productos globales de humedad del suelo y datos de estaciones agrometeorológicas, comparando siete algoritmos de escalamiento temporal. La segunda fase integró el modelo operacionalizado con OpenET en un marco de Promedio de Modelos Bayesiano, contrastando arquitecturas de calibración bajo independencia condicional versus estructuras de error espaciotemporales explícitas, evaluadas mediante validación cruzada temporal con encadenamiento hacia adelante.

Los resultados de la primera fase indicaron que SEB-PV puede operar con precisión aceptable utilizando datos globalmente disponibles, aunque no se identificó un algoritmo de escalamiento temporal universal, siendo su desempeño dependiente del contexto climático y estructural del huerto. En la segunda fase, la especificación explícita de la estructura de error espaciotemporal resultó el factor dominante de la exactitud predictiva

fuera de muestra, superando consistentemente a los marcos de independencia condicional en todas las arquitecturas evaluadas. La calibración de intercepto dependiente del estado resultó la arquitectura con mayor exactitud predictiva, con una mejora relativa del 46% en error normalizado respecto al promedio determinístico simple, superando el umbral planteado en la hipótesis, y mostró mayor estabilidad frente a condiciones meteorológicas no observadas durante la calibración. Esta investigación establece que la especificación de la estructura de error es el determinante principal de la exactitud predictiva en ensambles satelitales de ET, orientando dónde concentrar esfuerzos metodológicos. El marco probabilístico desarrollado aborda ambas limitaciones identificadas: reduce el sesgo sistemático mediante calibración adaptativa y produce distribuciones predictivas posteriores que cuantifican la incertidumbre asociada a cada estimación, proveyendo a usuarios y gestores una herramienta para discernir cuándo confiar y cuándo cuestionar los productos satelitales. El código Stan está disponible públicamente para facilitar su adopción y adaptación a otros sistemas.

ABSTRACT

Evapotranspiration (ET) estimation is fundamental for water management in agricultural systems under increasing hydrological pressure. Operationally available satellite-based ET platforms exhibit systematic biases in fruit orchards and report exclusively point estimates without uncertainty quantification, limiting the ability of users and water managers to evaluate the reliability of available products. To address this gap, a probabilistic methodological framework was developed in two phases oriented toward building an ensemble with structural diversity. The SEB-PV model (developed exclusively for partially vegetated surfaces) was identified as a complementary candidate to the six algorithms of the OpenET platform, but required overcoming operational barriers. The first phase evaluated its viability using global gridded soil moisture products and agrometeorological station data, comparing seven temporal upscaling algorithms. The second phase integrated the operationalized model with OpenET within a Bayesian Model Averaging framework, contrasting calibration architectures under conditional independence versus explicit spatiotemporal error structures, evaluated through forward-chaining temporal cross-validation.

Results from the first phase indicated that SEB-PV can operate with acceptable accuracy using globally available data, although no universal temporal upscaling algorithm was identified, with performance being dependent on the climatic and structural context of the orchard. In the second phase, explicit specification of the spatiotemporal error structure emerged as the dominant driver of out-of-sample predictive accuracy, consistently outperforming conditional independence frameworks across all evaluated architectures.

State-dependent intercept calibration yielded the highest out-of-sample predictive accuracy, with a 46% relative improvement in normalized error over the deterministic simple average, surpassing the threshold stated in the hypothesis, and showed greater stability under weather conditions not observed during calibration. This research establishes that error structure specification is the primary determinant of predictive accuracy in satellite-based ET ensembles, guiding where methodological efforts should be concentrated. The probabilistic framework developed addresses both identified limitations: it reduces systematic bias through adaptive calibration and produces posterior predictive distributions that quantify the uncertainty associated with each estimate, providing users and water managers a tool to discern when to trust and when to question satellite products. The Stan code is publicly available to facilitate adoption and adaptation to other systems.

CAPÍTULO 1

INTRODUCCIÓN

La estimación precisa de la evapotranspiración (ET) es fundamental para la gestión eficiente del agua en sistemas agrícolas, especialmente en regiones con recursos hídricos limitados y alta demanda atmosférica como el Valle Central de California y la zona central de Chile. La ET representa el proceso integrado donde el agua líquida se transforma en vapor, involucrando tanto la evaporación desde la superficie del suelo como la transpiración de las plantas (Allen et al., 1998). En California, la producción de cítricos representa el 92% de la producción nacional de cítricos frescos (Weber et al., 2023), mientras que en Chile, los huertos frutales constituyen un componente esencial de la economía exportadora (Quezada et al., 2025). Sin embargo, ambas regiones enfrentan desafíos crecientes relacionados con sequías recurrentes, disminución de las reservas de nieve y patrones de precipitación alterados (Quezada et al., 2025; Pathak et al., 2018; Sandoval-Solis, 2020), lo que exige herramientas de monitoreo hídrico de alta precisión y confiabilidad.

Los modelos de balance de energía superficial (SEB) proporcionan estimaciones de ET mediante relaciones físicas entre variables de entrada y salida (Wang and Dickinson, 2012). La adopción operacional de estos modelos ha seguido trayectorias metodológicas distintas. Los enfoques contextuales, ejemplificados por SEBAL y METRIC (Allen et al., 2007), emplean píxeles de anclaje que representan condiciones extremas de humedad dentro de la imagen satelital para escalar ET mediante diferencias de temperatura superficial. Esta estrategia enfrenta limitaciones cuando la imagen o subimagen procesada no contiene

superficies que satisfagan estos supuestos extremos, como ocurre en áreas de cultivo irrigado sin zonas desnudas secas adyacentes, o tras eventos de precipitación que uniformizan el estado hídrico de la escena. Los enfoques de píxel único, en contraste, evalúan el balance energético independientemente en cada ubicación sin requerir calibración intra-imagen, aunque requieren parametrización explícita de resistencias superficiales en lugar de calibrarlas implícitamente mediante píxeles de referencia. Una distinción adicional surge entre modelos de fuente única, que tratan la superficie como una capa uniforme, y enfoques de dos fuentes que distinguen explícitamente entre componentes de suelo y vegetación (Shuttleworth and Wallace, 1985; Anderson et al., 2007). Esta diferenciación resulta particularmente relevante en sistemas con cobertura parcial vegetada, donde durante estrés hídrico la transpiración se reduce por cierre estomático mientras que la evaporación del suelo puede mantenerse si hay riego reciente (Souto et al., 2019). Los modelos de dos fuentes capturan este desacoplamiento funcional, mientras que modelos de fuente única que promedian ambos componentes pierden precisión cuando sus respuestas divergen bajo déficit hídrico.

A pesar de estos avances metodológicos, sesgos sistemáticos han sido documentados cuando estos modelos se aplican a huertos frutales. La plataforma OpenET (Melton et al., 2022), que integra seis algoritmos satelitales mediante promedios simples para generar estimaciones diarias de ET a 30 m de resolución, representa un sistema operacional avanzado. Sin embargo, un estudio reciente en California documenta sobreestimación del uso anual de agua en huertos de cítricos hasta en 30% (Dhungel et al., 2024), evidenciando que la agregación determinística de modelos estructuralmente similares no elimina sesgos compartidos. Esta limitación refleja tanto los desafíos al representar heterogeneidad espacial en superficies parcialmente vegetadas como la ausencia de cuantificación rigurosa de incertidumbre en enfoques que asumen independencia entre estimaciones de modelos y entre observaciones temporales.

El modelo *Surface Energy Balance for Partially Vegetated surfaces* (SEB-PV) (Lagos

et al., 2012) representa un enfoque multi-capa avanzado diseñado específicamente para distinguir la canopia, la superficie del suelo y las capas del subsuelo. Su ventaja clave radica en la capacidad de separar la transpiración de la evaporación del suelo mediante una estructura de cuatro capas con resistencias explícitas, relevante para huertos con cobertura heterogénea. Esta especialización, sin embargo, ha venido acompañada de limitaciones para su adopción operacional. Además de los inputs meteorológicos estándar requeridos por otros modelos SEB, el modelo requiere parámetros adicionales como altura de canopia, ancho de hoja y propiedades del suelo que pueden obtenerse de mediciones puntuales de campo o bases de datos globales, pero las mediciones continuas de humedad del suelo in situ para calibración de resistencias superficiales no están disponibles en redes agrometeorológicas convencionales. Adicionalmente, carece de un algoritmo estandarizado para escalar estimaciones instantáneas a valores diarios, una limitación compartida con otros modelos SEB que estiman ET únicamente al momento de paso del satélite (Liu, 2021). Evaluar la viabilidad de sustituir las mediciones de humedad del suelo in situ por productos grillados globalmente disponibles, y establecer estrategias de escalamiento temporal, constituyen prerrequisitos para que modelos especializados como SEB-PV contribuyan a sistemas operacionales de gestión hídrica.

La integración de múltiples modelos mediante técnicas de ensamble frecuentemente supera el desempeño de modelos individuales en estimación de ET (Sun et al., 2019; Bai et al., 2021). Sin embargo, los enfoques existentes operan predominantemente bajo marcos determinísticos, promedios simples o ponderados por desempeño histórico, que asumen independencia entre predicciones de modelos y entre observaciones sucesivas (Hastie et al., 2009). Este supuesto es problemático en aplicaciones agrícolas donde modelos comparten forzantes meteorológicos comunes y procesos físicos subyacentes similares, condiciones bajo las cuales sus residuos tienden a exhibir correlación (F. Dormann et al., 2007). Adicionalmente, la memoria de humedad del suelo y patrones atmosféricos regionales generan correlaciones temporales y espaciales en residuos

que permanecen después del condicionamiento en predictores (Rahmati et al., 2024; Legendre, 1993). Cuando estas dependencias no se modelan explícitamente, los ensambles resultantes son subdispersos, donde los intervalos de predicción, si se reportan, subestiman la incertidumbre y el riesgo asociado a decisiones de manejo (Raftery et al., 2005).

El Promedio de Modelos Bayesiano (BMA) proporciona un marco formal para combinar predicciones de múltiples modelos mientras se cuantifica la incertidumbre mediante distribuciones predictivas posteriores (Raftery et al., 2005). Implementaciones previas en estimación de ET, sin embargo, han mantenido supuestos de independencia condicional (Wu et al., 2025), produciendo intervalos de credibilidad que reflejan solo incertidumbre paramétrica sin capturar dependencias residuales. Marcos probabilísticos con estructuras de error espaciotemporales explícitas, mediante procesos autorregresivos y matrices de correlación espacial, están bien establecidos en estadística espacial (F. Dormann et al., 2007; Banerjee et al., 2014), pero su aplicación sistemática a ensambles de ET satelital es limitada. Esta brecha metodológica es particularmente crítica para gestión operacional del agua, donde la cuantificación apropiada de incertidumbre es fundamental para la toma de decisiones bajo criterios de riesgo.

Integrar un modelo físico especializado como SEB-PV, operacionalizado mediante datos globalmente disponibles, dentro de un marco BMA que modela explícitamente dependencias espaciotemporales ofrece una vía para superar estas limitaciones estructurales. Sin embargo, una pregunta sin respuesta clara es: ¿las mejoras predictivas en ensambles de ET provienen principalmente de la complejidad de la arquitectura de calibración o de la especificación explícita de la estructura de error espaciotemporal? Y bajo una especificación de error espaciotemporal, ¿qué arquitectura de calibración ofrece mayor exactitud predictiva fuera de muestra? Estas preguntas tienen implicaciones operacionales: la complejidad injustificada puede aumentar la incertidumbre paramétrica sin mejorar las predicciones (Kim et al., 2023). Sistemas que ignoran dependencias

residuales tienden a sobreestimar los beneficios de calibraciones jerárquicas (Gelman and Hill, 2006), mientras que la calibración de parámetros sin modelar correlaciones espaciotemporales produce intervalos de predicción engañosamente precisos (Raftery et al., 2005). Resolver esta cuestión requiere un diseño factorial que permita aislar los efectos de cada componente, evaluando arquitecturas de calibración global, jerárquica y dependiente del estado bajo especificaciones de error alternativas mediante validación cruzada rigurosa que respete el ordenamiento temporal de los datos.

Esta tesis doctoral aborda estos desafíos mediante un enfoque secuencial de dos fases, estructurado en los siguientes capítulos principales, correspondientes a los manuscritos científicos desarrollados:

Capítulo 2: Aborda la operacionalización del modelo SEB-PV. Se evalúa su desempeño utilizando exclusivamente conjuntos de datos globalmente disponibles (productos de humedad de suelo grillados CFSv2 y SMAP L4) y se comparan siete métodos de escalamiento temporal en huertos de avellanos y pistachos en Chile y California. Este capítulo establece las bases para incluir un modelo físico especializado en un ensamble operativo sin depender de instrumentación in situ.

Capítulo 3: Desarrolla e implementa un marco de ensamble probabilístico avanzado. Integra el modelo SEB-PV (operacionalizado en el Capítulo 2) con seis modelos de la plataforma OpenET utilizando seis años de datos de torres de flujo en cítricos de California. Este estudio evalúa las fuentes de mejora predictiva, contrastando supuestos de independencia condicional versus estructuras de error espaciotemporales explícitas, y evaluando arquitecturas de calibración desde globales hasta dependientes del estado.

Finalmente, se presentan las conclusiones generales que sintetizan cómo la especificación explícita de dependencias espaciotemporales en marcos probabilísticos mejora la cuantificación de incertidumbre para la gestión del agua en agricultura.

Hipótesis y objetivos

Hipótesis (H): La integración del modelo SEB-PV operacionalizado en un marco de ensamble probabilístico con especificación explícita de dependencias espaciotemporales reducirá el error de estimación en más del 40% respecto a ensambles determinísticos (umbral establecido como valor de referencia dado que mejoras relativas comparables no han sido reportadas previamente en la literatura), siendo la estructura de error el factor dominante frente al refinamiento de la arquitectura de calibración.

Objetivo General (OG): Evaluar la jerarquía de importancia entre la especificación de la estructura de error espaciotemporal y la sofisticación de la arquitectura de calibración en el desempeño predictivo fuera de muestra de un ensamble probabilístico con diversidad estructural, que integra el modelo SEB-PV operacionalizado con modelos satelitales de OpenET para la estimación de evapotranspiración en huertos frutales.

Objetivos Específicos (OE):

OE1. Evaluar la viabilidad técnica de operar el modelo SEB-PV utilizando productos de humedad del suelo grillados y datos de estaciones agrometeorológicas como alternativas a la instrumentación in situ, y determinar los algoritmos de escalamiento temporal con mayor exactitud predictiva bajo distintas condiciones climáticas y estructurales del huerto.

OE2. Cuantificar la contribución relativa de la especificación de la estructura de error espaciotemporal frente a la complejidad de la arquitectura de calibración en el desempeño predictivo fuera de muestra de un ensamble de modelos satelitales de ET con diversidad estructural, e identificar la arquitectura con mayor exactitud predictiva bajo especificación explícita de dependencias espaciotemporales.

CAPÍTULO 2

Enhancing Evapotranspiration Estimates in Orchards with the Surface Energy Balance for Partially Vegetated surfaces (SEB-PV) Model through Combined Use of Gridded Soil Moisture and Temporal Upscaling Methods

Science of The Total Environment, Volume 1002, Article 180569 (2025)

doi: https://doi.org/10.1016/j.scitotenv.2025.180569

Submitted: 27 June 2025

Status: Published (1 November 2025)

Abstract

The Surface Energy Balance for Partially Vegetated surfaces (SEB-PV) model provides accurate evapotranspiration (ET) estimates for orchard crops. However, it faces two operational limitations: requiring specific input data unavailable from conventional agro-meteorological stations and lacking an evidence-based algorithm for upscaling instantaneous ET to daily values.

This study addresses these limitations by evaluating SEB-PV performance under three conditions: (1) using measured soil moisture with in-situ meteorological equipment; (2) using gridded soil moisture products (Climate Forecast System and Soil Moisture Active

Passive) with in-situ meteorological equipment; (3) using gridded soil moisture products with agro-meteorological stations' data. Seven temporal upscaling methods were compared for ET estimation in commercially-produced, micro-irrigated hazelnut (Chile) and pistachio (California) orchards. A Model Decision Making Indicator (MDMI), combining Kling-Gupta efficiency and normalized root mean square error (NRMSE), is proposed to enhance parameter optimization sensitivity.

SEB-PV performance using gridded soil moisture products demonstrated comparable accuracy to configurations using measured soil moisture after parameter adjustment (MDMI values $>$70 for hazelnuts, NRMSE $\sim$ 21%; $>$59 for pistachios, NRMSE $\sim$ 29%). Transitioning from in-situ meteorological measurements to agro-meteorological stations minimally impacted hazelnut orchards but required careful consideration for pistachios. Methods that upscale instantaneous ET to daily values on the basis of net radiation performed optimally for hazelnut orchards grown in Mediterranean climatic conditions (NRMSE $\sim$ 15%), while meteorological inputs-based methods were preferable for semi-arid pistachio orchards (NRMSE $\sim$ 30%). These findings show that SEB-PV can maintain acceptable accuracy using globally available datasets, improving operational applicability through guidance for input selection and temporal upscaling tailored to orchard characteristics.

2.1 Introduction

Orchard crop production in Chile and California faces intensifying water resource challenges from recurrent droughts, diminishing snowpack, and altered precipitation patterns, requiring precise evapotranspiration (ET) estimation for irrigation water management (Quezada et al., 2025; Pathak et al., 2018; Sandoval-Solis, 2020). Physically-based surface energy balance (SEB) models provide robust frameworks for accurate ET estimation due to well-established causal relationships between input variables and outputs (Wang and Dickinson,

2012). While many SEB models have achieved operational status, no single approach demonstrates optimal performance across different landscapes, with trade-offs between accuracy and operational feasibility particularly evident in orchard systems(Steduto et al., 2023).

Remote sensing-based SEB models are classified into contextual approaches (e.g., SEBAL, METRIC) that utilize anchor pixels representing extreme temperature and moisture conditions within the image scene to scale ET across all pixels (Allen et al., 2007), and pixel-based approaches that assess energy balance within individual pixels without requiring scene-wide calibration. Models are further distinguished as single-source approaches (SEBS, SEBAL, METRIC, SSEBop) treating surfaces as uniform layers, versus two-source approaches (TSEB, ALEXI, DisALEXI) that differentiate vegetation and soil energy transfer (Shuttleworth and Wallace, 1985; Anderson et al., 2007).

The Surface Energy Balance for Partially Vegetated Surfaces (SEB-PV) model (Lagos et al., 2012) represents an advanced multi-layer approach employing a four-layer structure that distinguishes canopy, ground surface, upper soil, and lower sub-soil layers. This structure enables separate estimation of net radiation (Rn), sensible heat flux (H), latent heat flux (λE), and soil heat flux (G) for each layer, including effects of crop residues on soil evaporation. The model's key advantage lies in its ability to separate transpiration from soil evaporation, which is particularly valuable for orchard systems with heterogeneous canopy coverage and exposed soil surfaces between tree rows.

Multi-layer models like SEB-PV generally enhance accuracy in complex landscapes (Anderson et al., 2007; Lagos et al., 2009), and recent developments have enhanced SEB-PV performance in orchard crops through wetted diameter parameterization for micro-irrigation systems (Souto et al., 2019, 2022). However, validation in orchard crops with complex spatial configurations and seasonal growth patterns remains challenging (Grisafi et al., 2021).

Despite its proven accuracy, SEB-PV faces two significant operational constraints that

limit widespread implementation. First, the model requires specialized input parameters including detailed canopy characteristics, soil physical properties, and site-specific soil moisture measurements that necessitate extensive field campaigns and specialized equipment unavailable from conventional agro-meteorological station networks. The model calculates soil surface resistance to water vapor diffusion using an exponential function that includes a $beta$ parameter requiring calibration with measured soil moisture data, while canopy resistance is modeled through a multi-factor equation where $c1$ serves as a coefficient relating canopy resistance to leaf area index, vapor pressure deficit, and solar radiation.

Second, SEB-PV lacks an evidence-based algorithm for upscaling instantaneous ET estimates to daily values limiting its utility in irrigation water management. When applied to satellite imagery, SEB models detect only the instantaneous rate of evapotranspiration at satellite overpass time, which is not directly applicable for practical water management (Liu, 2021). This temporal upscaling challenge represents a key limitation in remote sensing-based ET estimation (Liu et al., 2020; Chen and Liu, 2020), particularly in orchard crop systems where temporal variations are influenced by row orientation, canopy architecture, crop-specific stomatal conductance responses to vapor pressure deficit (VPD), air temperature, relative humidity, and irrigation methods (ASCE-EWRI, 2005).

Recent advancements have demonstrated the potential of gridded soil moisture products to improve ET estimation accuracy when integrated into remote sensing algorithms. The SMAP Level-4 (9-km) and CFSv2 (0.25°) products provide global, regularly updated soil moisture estimates (Reichle et al., 2017; Saha et al., 2014). Studies integrating SMAP soil moisture into ET algorithms have shown significant improvements, with error reductions of 23% for soil evaporation and 17% for transpiration components, particularly in water-limited regions (Purdy et al., 2018). However, the accuracy of these products can degrade significantly under orchard canopies compared to open agricultural fields (Reichle et al., 2017), raising questions about their applicability in orchard crop systems.

For temporal upscaling, different SEB models employ distinct approaches based on multiplying instantaneous ET by scaling ratios to obtain daily estimates. These algorithms typically assume constant ratios between ET and available energy ($Rn - G$) or reference ET ratios throughout daylight hours (Brutsaert and Sugita, 1992; Colaizzi et al., 2006). Validation studies across diverse landscapes show context-dependent performance; with some highlighting insolation-based methods (Cammalleri et al., 2014) and others favoring available energy ratio approaches (Chávez et al., 2008), with variations influenced by sky conditions, time of day, and ecosystem type (Jiang et al., 2021). This variability necessitates systematic evaluation of temporal upscaling approaches specifically adapted to orchard crops (Odi-Lara et al., 2016; Mobe et al., 2020).

This study addresses the operational limitations of SEB-PV by evaluating whether the model can achieve acceptable accuracy using exclusively globally available datasets, thus expanding its geographic applicability to regions lacking specialized field infrastructure. The research evaluates: 1) the viability of gridded soil moisture products (CFSv2 and SMAP L4) as alternatives to in-situ soil moisture measurements; 2) the performance of seven temporal upscaling methods for converting instantaneous to daily ET estimates. The research study was conducted on hazelnut and pistachio orchards grown with microirrigation for commercial production in Chile and California, respectively, selected for their contrasting canopy structures and climatic conditions, covering growing seasons from 2015 to 2022. Addressing this gap is a critical step toward expanding the model's applicability, which would add a valuable multi-layer, physically-based approach to the suite of operational remote-sensing surface energy balance models, thus contributing to improved management of agricultural water supplies in fruit and nut production areas.

2.2 Materials and methods

The SEB-PV model performance was assessed under three configurations characterized by different data inputs using seven alternative algorithms for temporal upscaling ET from instantaneous to daily values for commercial hazelnut orchards in Chile and pistachio orchards in California. The methodological framework used in this study was illustrated in Figure 2.1.

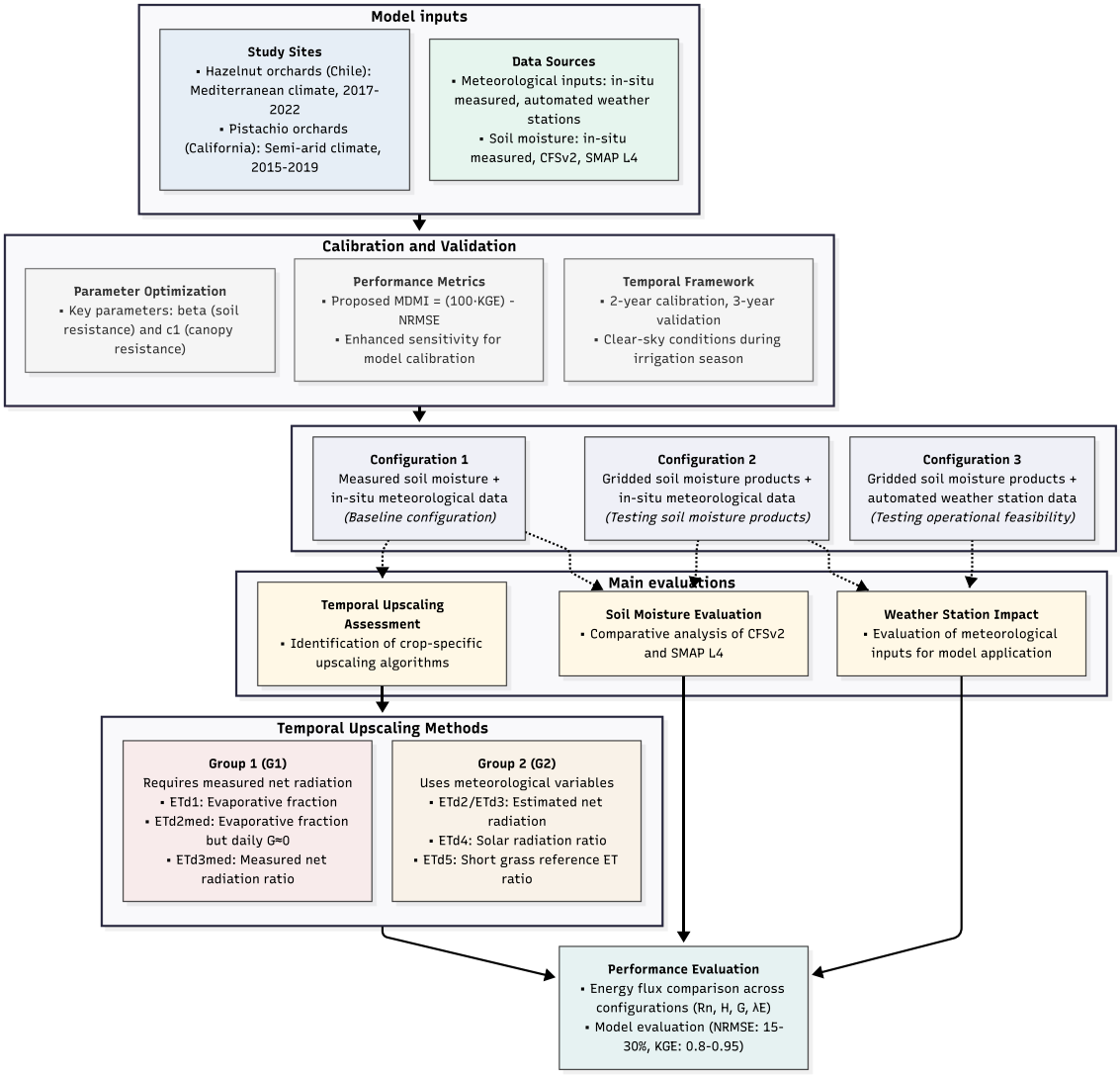

Figure 2.1: Methodological framework for evaluating the SEB-PV model with alternative data sources and temporal upscaling methods.

The methodology was structured in four main stages, as illustrated in Figure 2.1. First, data collection and preparation were conducted, including half-hourly energy balance components measurements from flux towers, in-situ meteorological data within the orchard canopy environment, hourly data from networks of automated weather stations (Chile Agricultural Research Institute and California Irrigation Management Information System), on-site soil moisture measurements, and gridded soil moisture products (CFSv2 at 0.25° resolution, SMAP L4 at 9-km resolution). The integration and processing of all spatial and temporal datasets, including Landsat 8/9 imagery, were performed using the Google Earth Engine platform (Gorelick et al., 2017). Second, an experiment was designed to test three distinct model configurations: (1) a baseline configuration using measured soil moisture and in-situ meteorological data, (2) a second configuration testing gridded soil moisture products with in-situ meteorological data, and (3) a third configuration testing operational feasibility with both gridded soil moisture and data from agro-meteorological networks. Third, for each configuration, the model's key $beta$ (soil) and $c1$ (canopy) parameters were optimized using a novel Model Decision Making Indicator (MDMI) that combines Kling-Gupta efficiency (KGE) and normalized root-mean-square error (NRMSE). Finally, a comprehensive performance evaluation was conducted, focused on two main analyses: an assessment of the model's accuracy across the three input data configurations, and a comparative analysis of seven temporal upscaling approaches for estimating daily ET. These upscaling methods were categorized into two groups based on their input requirements: those requiring measured net radiation (Group 1) and those using standard meteorological variables (Group 2). Each component of this methodology is described in detail in the following subsections.

2.2.1 Study sites

Four sites at 36°S (Chile) and 36°N (California) were encompassed in the study: two in Chile (Larqui and Pullami) and two in California, U.S. (Flores and Nichols), as shown in Figures 2.2 and 2.3. The Chile sites had warm-summer Mediterranean climates

(Csb), characterized by mild temperatures and substantial winter rainfall ($>$1,000 mm annually). In contrast, the California sites had cold semi-arid climates (BSk), characterized by hot summers, cool winters, and limited annual precipitation ($<$250 mm), with rainfall concentrated between November and March. Orchard photographs illustrating canopy structure differences between hazelnut and pistachio sites were presented in Figure 2.4.

Chile Sites

Larqui Site The Larqui site (36°42' 30.53" S, 72°21' 38.25" W, 72 m above sea level) was located in the Bulnes commune, Ñuble region of Chile. For the 2017-2022 period, the following climate characteristics were recorded: a mean annual temperature of 15.2°C (with average annual daily minimums of 8.6°C and maximums of 21.8°C), and mean annual precipitation of 594.1 mm. Mean reference evapotranspiration (ET_0) during the growing season for this period was 822.4 mm. This site was characterized by a Lewis cv. hazelnut orchard that had been established in 2013, with a mature canopy height averaging 4.0 m during the study period and tree spacing of 5.0 m × 3.5 m. The orchard block covered 4.3 hectares and is characterized by clay loam soil. Irrigation was provided through a double-line drip system with four 3.8 L h^{-1} emitters spaced at 0.9 m intervals, with irrigation interval of 2-3 days throughout the growing season.

Pullami Site The Pullami site (36°35' 20.07" S, 71°47' 55.51" W, 254 m above sea level) was located in the Coihueco commune, Ñuble region of Chile. For the 2017-2022 period, the following climate characteristics were recorded: a mean annual temperature of 13.4°C (with average annual daily minimums of 5.7°C and maximums of 21.0°C), and mean annual precipitation of 802.7 mm. Mean ET_0 during the growing season for this period was 762.4 mm. This site was characterized by a hazelnut orchard of "Tonda di Giffoni" cv. that had been established in 2011, with a mature canopy height averaging 4.0 m during the study period and tree spacing of 5.0 m × 2.5 m. The orchard block covered 14.4 hectares

and is characterized by deep clay soil. Irrigation was provided through a micro-sprinkler system (one micro-sprinkler per tree) with a nominal flow rate of 30 L h^{-1} per tree, with six irrigations per week in November-December, 5 days per week in January-February, and 7 days per week in March.

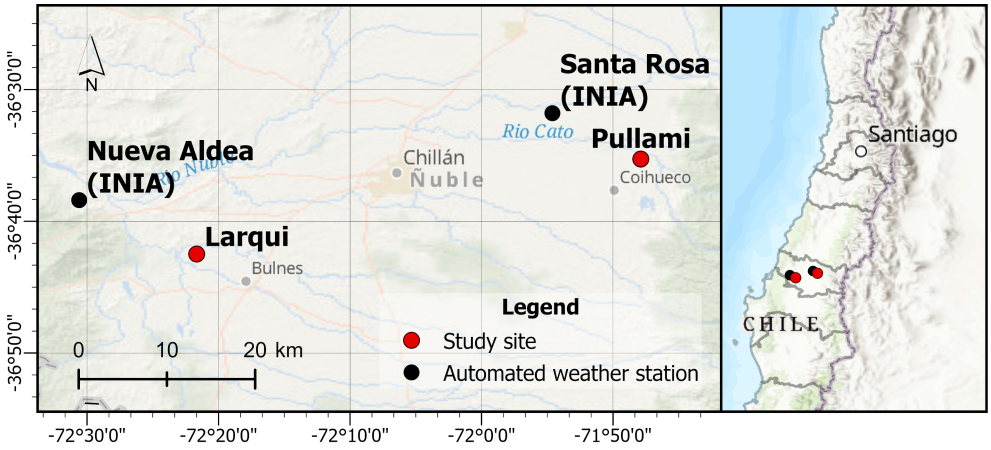

Figure 2.2: Geographic locations of the hazelnut orchard study sites (Larqui and Pullami) in the Ñuble region, Chile.

California Sites

Flores Site The Flores site (36°14' 34.46" N, 119°56' 16.44" W, 72 m above sea level) was located near Lemoore, in the south-western part of the San Joaquin Valley. For the 2015-2019 period, the following climate characteristics were recorded: a mean annual temperature of 18.1°C (with average annual daily minimums of 9.9°C and maximums of 26.2°C), and mean annual precipitation of 148.6 mm. Mean ET_0 during the growing season for the study period was 1,443.5 mm. This site was characterized by a pistachio orchard with trees of Kerman cv. grafted onto Pioneer Gold 1 rootstock that had been established during the late 1980s, with a mature canopy height averaging 2.9 m during the study period and tree spacing of 5.0 m × 5.0 m. The orchard covered 58.2 hectares and is mainly characterized by saline-sodic clay-loam soil. Irrigation was provided through a double-line drip system with eight 1.4 L h^{-1} emitters per tree spaced at 1.3 m intervals, and irrigation

frequency that varied from once every 10 days in April-May to multiple times daily during peak water demand period (July-August).

Nichols Site The Nichols site (36°15' 29.25" N, 119°30' 27.04" W, 77 m above sea level) was located near Hanford, in the south-central part of the San Joaquin Valley. The climate characteristics for the 2015-2019 period were consistent with those previously described for the Flores site. This site was characterized by a pistachio orchard with trees of Kerman cv. grafted onto Pioneer Gold 1 rootstock that had been established in 1985, with a mature canopy height averaging 4.7 m during the study period and tree spacing of 6.0 m × 5.0 m. The orchard covered 61.1 hectares and is characterized by fine sandy-loam soil. Irrigation was provided through a double-line drip system with ten 4 L h^{-1} self-compensating emitters per tree spaced at 1.0 m intervals, with irrigation frequency that varied from once every 10 days in April-May to multiple times daily during peak water demand period (July-August).

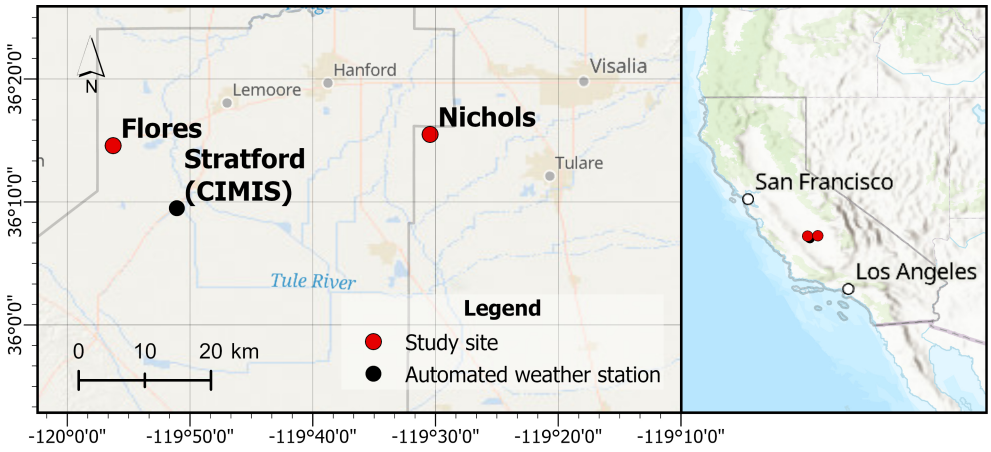

Figure 2.3: Geographic locations of the pistachio orchard study sites (Flores and Nichols) in California's San Joaquin Valley.

2.2.2 In situ measurements

Actual ET was determined using the residual of the surface energy balance method from micro-meteorological measurements using a combination of eddy covariance and surface

Figure 2.4: Representative photographs illustrating canopy structure differences between the hazelnut (Larqui, Pullami) and pistachio (Flores, Nichols) study orchards.

renewal equipment. Sensible heat flux (H) was measured at all sites using a 3-D sonic anemometer (RM Young 81000) sampling at 10 Hz. At the California sites, gaps in sonic anemometer data were filled using the Surface Renewal technique, also sampled at 10 Hz (Paw U et al., 1995). Data from the growing seasons of 2015-2019 for pistachios and 2017-2022 for hazelnuts were analyzed, allowing for a multi-year analysis that accounts for annual variations in climate and crop physiology. The location of each station was detailed in Table 2.1.

Table 2.1: Summary of ET station characteristics at the study sites.

Station name	Crop	Latitude (°)	Longitude (°)	Elevation (m)
Larqui	Hazelnut	-36.708	-72.361	72
Pullami	Hazelnut	-36.588	-71.798	254
Flores	Pistachio	36.243	-119.938	72
Nichols	Pistachio	36.258	-119.508	77

2.2.3 Instrumentation and Setup

At all sites, the following sensors and instrumentation were installed at each station:

Energy Balance Components

- **Net Radiation (Rn):** NRLite2 net radiometers (Kipp & Zonen Inc., Delft, Netherlands) mounted 1.5 m above the canopy.
- **Sensible Heat Flux (H):** RM 81000 three-dimensional sonic anemometers (R.M. Young Inc., Traverse City, MI, USA) mounted approximately 2.0 m above the canopy.
- **Soil Heat Flux (G):** Multiple soil sensor packages were deployed at each site. Each package contained:
  - Soil heat flux plates installed horizontally at 0.08 m depth: HFP01SC sensors (Hukseflux, Delft, Netherlands) at Chile sites and HFT3 sensors (REBS, Bellevue, Washington) at California sites
  - Soil temperature thermocouples (TCAV, Campbell Scientific Inc., Logan, UT, USA) installed at depths of 0.02-0.06 m
  - Soil moisture sensors installed at 0.03 m depth: CS616 sensors (Campbell Scientific Inc., Logan, UT, USA) at Chile sites and EC5 sensors (Decagon Devices, Pullman, WA, USA) at California sites

  At the Chile sites, two packages were installed (one in the tree row and one between rows), while at California sites, three packages were installed along a transect perpendicular to the tree row. The mean of all measurements was used to calculate soil heat flux.

Meteorological Measurements

- **Precipitation:** TR-525M rain gauges (Texas Electronics, Dallas, TX, USA)

- **Air Temperature and Relative Humidity:** Temperature and humidity probes mounted approximately 1.0 m above the canopy

These meteorological measurements were collected within the orchard canopy environment, capturing the site-specific microclimate conditions experienced by the crops.

Data Collection and Processing Sonic anemometer and Surface Renewal data were sampled at 10 Hz, and all measurements were averaged and recorded at 30-minute intervals using CR3000 data loggers (Campbell Scientific Inc., Logan, UT, USA). Latent heat flux (λE) was determined as the residual of the energy balance method:

$$\lambda E = Rn - H - G \tag{2.1}$$

where Rn is net radiation, H is sensible heat flux, and G is soil heat flux, all expressed in W m^{-2}.

2.2.4 Quality Control and Maintenance

Rigorous quality control procedures were implemented to address invalid or missing data due to sensor malfunctions or disturbances. In California sites, when sonic anemometer data were unavailable, sensible heat flux was estimated using the surface renewal technique to maintain data consistency. Monthly maintenance procedures were conducted, which included sensor cleaning, leveling, and verification of instruments' orientation. This approach to data collection and quality control ensured the availability of robust field datasets for validating and refining the SEB-PV model across different orchard types and meteorological conditions.

2.2.5 SEB-PV Model

The Surface Energy Balance for Partially Vegetated Surfaces (SEB-PV) model was implemented with specific adaptations for orchard crop systems using globally available datasets. The model employed a four-layer resistance network that explicitly partitioned evapotranspiration into canopy transpiration and soil evaporation from both shaded (beneath canopy) and sunlit (inter-row) areas, addressing the heterogeneous spatial distribution characteristic of orchard systems. The model's conceptual diagram for both sensible and latent heat fluxes was illustrated in Figure 2.5.

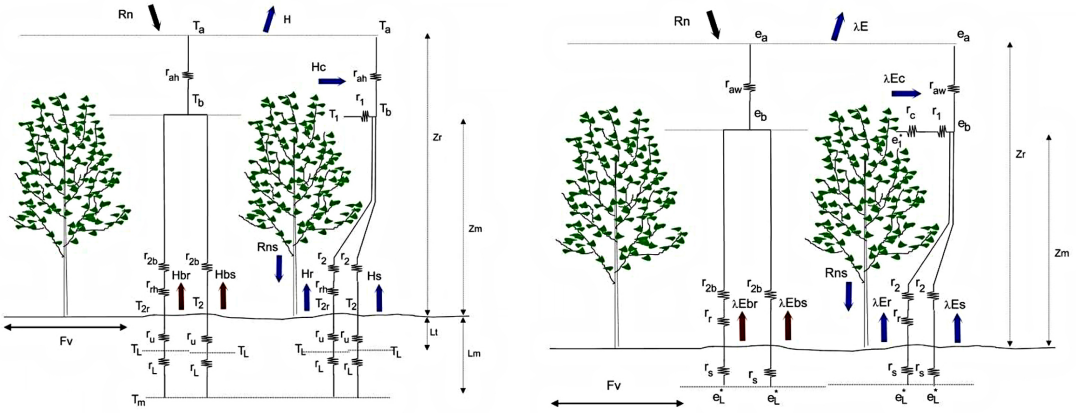

Figure 2.5: Conceptual diagram of the SEB-PV model's four-layer structure and its analogous resistances. The diagram illustrates pathways for sensible heat flux (left) and latent heat flux (right), including resistances for vegetated and bare soil areas. Source: Lagos et al. (2012).

The fundamental energy balance equation was solved for each surface component within the orchard system:

$$Rn = \lambda E + H + G \tag{2.2}$$

where net radiation (Rn) was partitioned into latent heat flux (λE), sensible heat flux (H), and soil heat flux (G), all expressed in W m^{-2}. The critical advancement of SEB-PV was achieved through its spatial partitioning of total evapotranspiration based on fractional

vegetation cover (F_v, dimensionless):

$$\lambda E = [\lambda E_c + \lambda E_s(1 - f_r) + \lambda E_r f_r]F_v + [\lambda E_{bs}(1 - f_r) + \lambda E_{br} f_r](1 - F_v) \quad (2.3)$$

where λE_c represented canopy transpiration (W m^{-2}), λE_s was soil evaporation beneath the canopy (W m^{-2}), λE_r was evaporation from residue-covered soil beneath the canopy (W m^{-2}), λE_{bs} was evaporation from inter-row bare soil areas (W m^{-2}), λE_{br} was evaporation from residue-covered inter-row soil (W m^{-2}), and f_r was the fraction of soil covered by residue. Given the absence of residue on the orchard floor at all study sites ($f_r = 0$), this equation simplified to:

$$\lambda E = [\lambda E_c + \lambda E_s]F_v + [\lambda E_{bs}](1 - F_v) \quad (2.4)$$

The heterogeneous nature of orchard systems required explicit treatment of radiation reaching different surface components, diverging fundamentally from homogeneous canopy assumptions employed in traditional models. Net radiation beneath the canopy was calculated using Beer's law attenuation:

$$Rn_s = Rn \cdot \exp(-k_{ext} \cdot LAI) \quad (2.5)$$

where k_{ext} was the extinction coefficient for net radiation (dimensionless) and LAI was the Leaf Area Index, a measure of the total leaf area per unit ground area (m^2 m^{-2}). In contrast, full incident radiation was received by inter-row areas ($Rn_{bs} = Rn$), creating distinct energy environments within the orchard system. Net radiation absorbed by the canopy (Rn_c, W m^{-2}) was calculated as:

$$Rn_c = Rn - Rn_s \quad (2.6)$$

This partitioning was essential for accurate soil evaporation estimates, as uniform

radiation assumptions would have systematically overestimated evaporation from shaded areas beneath tree canopies while underestimating evaporation from exposed inter-row soil surfaces.

For this implementation, the original SEB-PV conceptual model was enhanced through several refinements:

1. Enhanced net radiation (Rn) calculation was implemented based on Allen et al. (2007)

2. Landsat Collection 2 Level-2 Surface Temperature Science Product was integrated for land surface temperature determination

3. Updated solar constant value of 1366.1 W m^{-2} was adopted (ASTM, 2000) - where the solar constant value was defined as the total solar irradiance at normal incidence on a surface in free space at the earth's mean distance from the sun

4. Improved albedo estimation was performed using coefficients from Ke et al. (2016) derived from the SMARTS2 model

5. Advanced soil thermal conductivity (K) modeling was implemented following Lu et al. (2014)

Half-hourly weather data collected during Landsat 8/9 satellite overpasses (11:30-12:00 local time for all study orchards) were used to implement the model, including air temperature and relative humidity, wind speed, and volumetric soil water content measurements from both row and inter-row positions. Google Earth Engine cloud computing platform (Gorelick et al., 2017) was used to conduct all satellite data processing and spatial analyses, which allowed efficient processing of Landsat imagery and integration with ground-based measurements. Site-specific bio-physical parameters (crop height and leaf width) were inputted as time-averaged values for different periods for each study orchard.

The original model's sensitivity analysis by Lagos et al. (2012) identified soil surface resistance (r_s) - resistance to water vapor diffusion from the soil surface to the atmosphere - as a critical parameter, particularly in orchards with sparse vegetation covers. This resistance was calculated according to Eq. 2.7:

$$r_s = \frac{L_t \cdot \tau_s}{Dv \cdot \phi} \cdot \exp\left(-\frac{beta \cdot \theta}{\theta_s}\right) \tag{2.7}$$

where L_t was the thickness of the surface soil layer (m), τ_s was a soil tortuosity factor (dimensionless), Dv was the water vapor diffusion coefficient (m^2 s^{-1}), ϕ was soil porosity, θ was the average volumetric water content in the surface layer (m^3 m^{-3}), θ_s was the saturation water content (m^3 m^{-3}), and $beta$ was a fitting parameter (dimensionless). The resulting soil surface resistance (r_s) was expressed in s m^{-1}.

The model also showed high sensitivity to canopy resistance (r_c, s m^{-1}; Eq. 2.8) under full coverage, moderate sensitivity to residue resistance (resistance to water vapor transfer through crop residues covering the soil), and lower sensitivity to vapor pressure deficit. Canopy resistance regulated transpiration from vegetated surfaces and was modeled as:

$$r_c = \left[\frac{c1 \cdot LAI}{LAI_{max}} \cdot \frac{c2}{1 - \exp(-VPD)} \cdot \frac{R_s \cdot (R_{smax} + c3)}{R_{smax} \cdot (R_s + c3)}\right]^{-1} \tag{2.8}$$

where LAI_{max} was the maximum value of LAI, fixed at 6.0 (m^2 m^{-2}) as defined by Lagos et al. (2012); VPD was the vapor pressure deficit, representing the difference between the amount of moisture in the air and how much moisture the air could hold when saturated (kPa); R_s was the solar radiation (W m^{-2}); R_{smax} was the maximum value of solar radiation, fixed at 1,000 W m^{-2} following Lagos et al. (2012); and $c1$, $c2$, and $c3$ were regression coefficients (dimensionless) used to fit the model to field observations. Eq. 2.8 did not account for stress effects due to soil moisture and focused solely on how canopy resistance was influenced by LAI, vapor pressure deficit, and solar radiation.

Following the model's sensitivity analysis by Lagos et al. (2012), the calibration process

was focused on determining two key parameters: $beta$, which controlled soil surface resistance to water vapor diffusion, and $c1$, which regulated canopy resistance through LAI dependency. Instantaneous satellite-based surface energy balance components (R_n, G, H, and λE) with corresponding half-hourly measured energy fluxes were compared using calibration curves. Relative differences were minimized between modeled and measured fluxes while maximizing overall model performance, using initial parameter values of $beta$ = 6.5 and $c1$ = 5.0 as recommended by Lagos et al. (2012).

The complete mathematical derivation of the original SEB-PV framework was provided in Lagos (2008), including model implementation algorithms. The extension to include residue effects, which formed the basis for the orchard adaptations presented here, was described in Lagos et al. (2012).

2.2.6 Datasets

Automated weather stations

Nearby automated weather stations from agro-meteorological networks were used to obtain air temperature and relative humidity (2.0 m in Chile, 1.5 m in California), wind speed (2.0 m), and solar radiation, to provide the meteorological inputs required by the SEB-PV model. These stations provided standardized, hourly data measured over uniform surfaces, representing broad-area reference meteorological conditions rather than the microclimate within the orchard canopies. For weather stations in Chile (Figure 2.2), data were taken from the agro-meteorological network of the Chile Agricultural Research Institute. For California (Figure 2.3), data were sourced from the California Irrigation Management Information System, that is operated by the California Department of Water Resources. The nearest weather stations to each study site were identified and selected, and were presented in Table 2.2.

These data were not sufficient to run the model simulations, and were complemented with a volumetric soil water content dataset as indicated in section 2.2.6 of this manuscript.

Table 2.2: Features of the automated weather stations from agro-meteorological networks employed as alternative meteorological data sources used for the model simulations, including locations, distances from study sites, and elevations.

Station name	Location	Site name	Distance to the site (km)	Lat (°)	Lon (°)	Elevation (m)
Nueva Aldea	Ránquil, Chile	Larqui	15.4	-36.64	-72.51	82
Santa Rosa	Chillán, Chile	Pullami	11.9	-36.53	-71.91	194
Stratford	Stratford, CA	Flores	12.2	36.16	-119.85	59
Stratford	Stratford, CA	Nichols	32.9	36.16	-119.85	59

Mid-day (noon) was selected as the representative instant for the model, as the values represented an average of the preceding 60 minutes. In addition, the time of Landsat satellite overpass in summer occurred between 11 am and 12 pm local time for all study orchards (between 10:00 and 11:00 am standard time).

Soil moisture datasets

Measured soil moisture The volumetric soil water content (VWC) was measured by the sensors described in section 2.2.2 at a depth from 0 to 0.06 m, both in the row middles and between trees along the row, which were inputs for the SEB-PV model. To allow running the model globally without relying on site-specific field measurements, VWC was estimated using two alternative products. Although these global product estimations may exhibit poor accuracy, their use allowed for broader model applicability. A single pixel from each product was extracted and used as input for the SEB-PV model at each site, with their impact on model performance evaluated in following sections.

SMAP Level-4 SMAP Level-4 Surface and Root Zone Soil Moisture product version 7 (Reichle et al., 2022) provided global, 3-hourly, 9-km resolution VWC content estimates for the surface (0–5 cm), root zone (0–100 cm), and profile (0 cm to bedrock) layers. The algorithm assimilated L-band radiometer brightness temperature observations from SMAP into the NASA Catchment land surface model driven with gauge-based precipitation

observations. The utilized product was a transformed soil moisture raster from the Equal-Area Scalable Earth Grid version 2 projection to geographic coordinates, achieved using GDAL libraries with a resolution of 11,000 meters. This product showed an average accuracy of 0.026 m^3 m^{-3} bias-adjusted root mean squared error across the 9-km core site reference pixels; however, error values could increase to 0.057 m^3 m^{-3} or higher under orchard canopies or natural vegetation covers (Reichle et al., 2017). The estimated VWC at 1330 UTC for Chile and 1930 UTC for California was utilized, which represent the closest available times to the satellite overpass.

CFSv2 The CFSv2 product, version 2 of the National Centers for Environmental Prediction's (NCEP's) Climate Forecast System, was initialized four times daily (0000, 0600, 1200, and 1800 UTC) (Saha et al., 2014). It provided estimates of volumetric soil moisture content at various depths below the surface layer, specifically at 5 cm, 25 cm, 70 cm, and 150 cm. Unlike SMAP Level-4, comprehensive validation studies documenting the error metrics specifically for agricultural applications were limited for this product. For this study, the estimated VWC at 1200 UTC (Chile) and 1800 UTC (CA) was used.

Since the estimates from gridded products match the target depth of field measurements, the surface layer zone estimates at 5 cm for CFSv2 and (0-5 cm) for SMAP L4 were used. Time series of measured volumetric water content in row (VWC) and between rows (VWCbr) along with these products were presented in Figure 2.6. Further preliminary comparison of these products is provided in B.

Preliminary assessment of gridded products Prior to model evaluation, the correspondence between gridded and measured soil moisture was assessed to understand baseline product performance (Figure 2.14). Both products showed systematic underestimation during irrigation periods, with Nash-Sutcliffe efficiency (NSE) consistently below 0.2 across most site-product combinations. The sole exception was at Larqui when comparing inter-row measurements (NSE $>$0.87). Despite poor representativeness, these

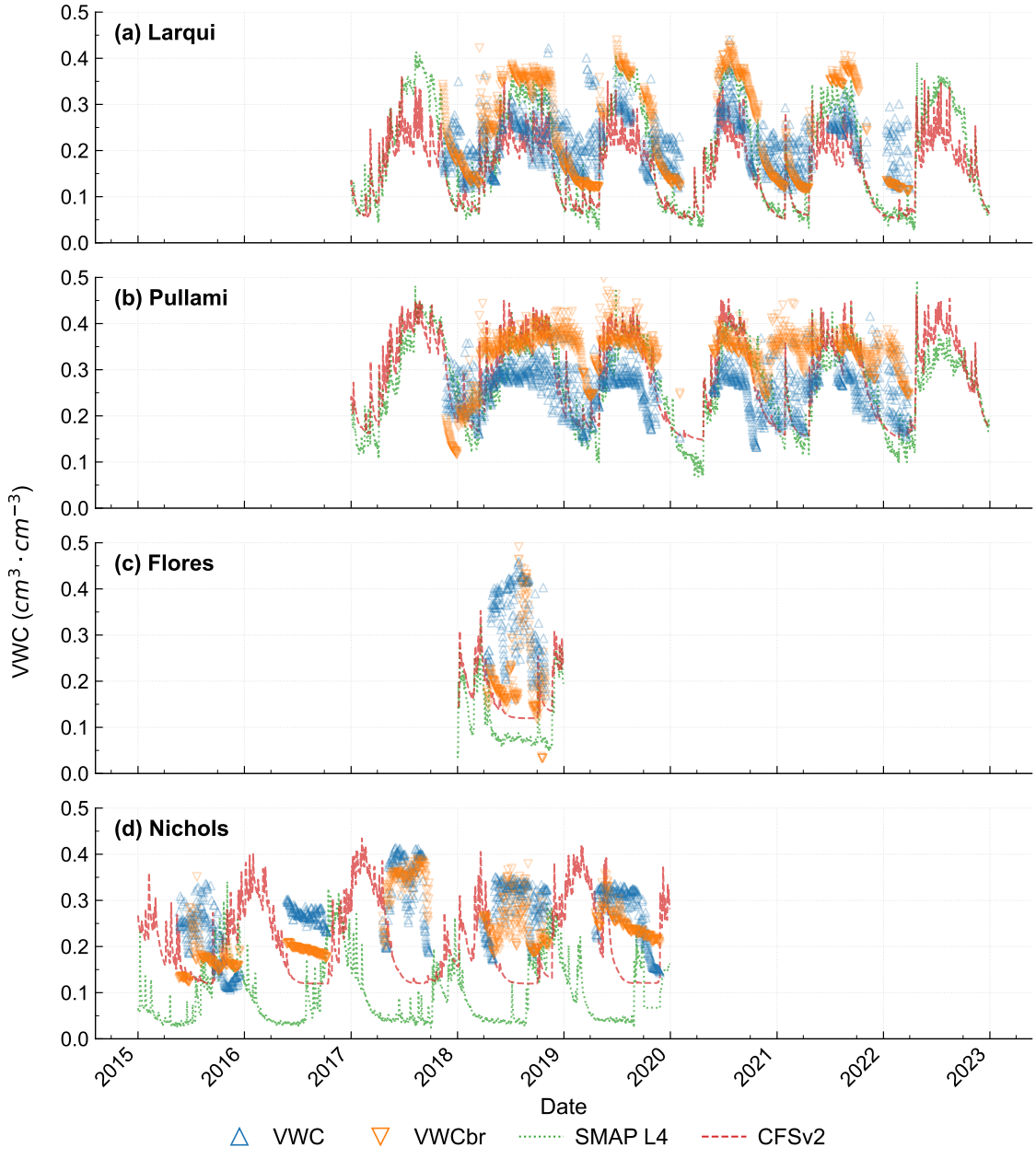

Figure 2.6: Comparison of volumetric water content (VWC) time series from in-situ measurements (row and inter-row) and gridded products (CFSv2 and SMAP L4) at all study sites.

products provided sufficient information for model calibration through parameter adjustment, as demonstrated in subsequent analyses.

Soil physics characteristics (SoilGrids)

The SoilGrids product developed by Poggio et al. (2021) generated maps of soil properties for the entire globe at medium spatial resolution (250 m cell size) using state-of-the-art machine learning methods, providing maps of pH, soil organic carbon content, bulk density, coarse fragments content, sand content, silt content, clay content, cation exchange capacity (CEC), total nitrogen as well as soil organic carbon density and soil organic carbon stock; at depths of 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm. In this study, only bulk density, as well as sand and clay contents were used.

Satellite imagery

Landsat Collection 2 Level-1 Product, available from the U.S. Geological Survey, included Landsat 1-5 MSS, Landsat 4-5 TM, Landsat 7 ETM+ and Landsat 8-9 OLI/TIRS. This second collection offered various improvements, including enhanced radiometric and geometric accuracy for all Landsat missions, with significant advancements in Landsat 8 (L8) TIRS and OLI sensors. A total of 59 unique Landsat 9 (L9) and 245 unique L8 images were used, amounting to 304 unique images, some of which spanning multiple study sites. Table 2.3 shows the number of unique images used per study site and path/row of Landsat 8-9. Images with more than 1% cloud cover within a 200 m buffer around the flux tower footprint were excluded (the spatial area contributing to the flux measurements, see section 2.2.8), and only images from dates with corresponding ground-based ET measurements were retained.

Table 2.3: Number of quality-controlled Landsat 8-9 images (cloud cover $<$1% within a 200 m flux tower footprint buffer) used per study site, with corresponding path/row information.

Field name	path/row	Total
Larqui	001/085, 233/085, 233/086	63
Pullami	233/085	15
Flores	042/035, 043/035	23
Nichols	042/035	50

2.2.7 Temporal upscaling models

To address the lack of a built-in algorithm in SEB-PV for determining daily ET values from instantaneous ET estimates, five equations evaluated by Chávez et al. (2008) were implemented. The temporal upscaling methods were categorized into two alternative approaches: (1) methods that use measured net radiation as input (Group 1), and (2) methods that rely on other meteorological inputs such as solar radiation, air temperature, and reference evapotranspiration (Group 2).

The first method was based on Eq. 2.9 (Shuttleworth et al., 1989) and Eq. 2.10 (Brutsaert and Sugita, 1992).

$$EF = \frac{\lambda E_i}{(Rn - G)_i} \tag{2.9}$$

$$\lambda E_i = (Rn - G - H)_i \tag{2.10}$$

where EF is the evaporative fraction (dimensionless), $(Rn - G)_i$ is the instantaneous available energy (AE) in W m^{-2}, λE_i is the instantaneous remotely sensed based latent heat flux rate in W m^{-2}, $(Rn - G - H)_i$ is the instantaneous remotely sensed based net radiation, soil heat flux, and sensible heat flux, respectively, in W m^{-2}. Based on AE, the first method is described by Eq. 2.11.

$$\mathrm{ET}_{d1} = [EF(Rn - G)_d] \times \left[\frac{cf}{\lambda_v \rho_w}\right] \tag{2.11}$$

considering ET_{d1} as the method ETd1 for daily ET rate, mm d^{-1}, $(Rn - G)_d$ is the mean measured 24 h AE, W m^{-2}, cf = time (unit) conversion factor equal to 86400 s d^{-1} for daily ET and 3600 s h^{-1} for hourly ET, λ_v is the latent heat of vaporization (Eq. 2.12), MJ kg^{-1} and ρ_w is the water density (Eq. 2.13). The latent heat of vaporization was described by Harrison (1965) as:

$$\lambda_v = \frac{(2500.84 - 2.3601 \cdot T_a)}{1000} \tag{2.12}$$

where T_a is the air temperature in ℃. The water density (ρ_w) was described as a third order quadratic equation by Kell (1975) as:

$$\begin{aligned}\rho_w = {} & \frac{999.83952 + 16.945176t - 7.9870401 \times 10^{-3}t^2}{1 + 16.897850 \times 10^{-3}t} \\ & - \frac{46.170461 \times 10^{-6}t^3 + 105.56302 \times 10^{-9}t^4 - 280.54253 \times 10^{-12}t^5}{1 + 16.897850 \times 10^{-3}t}\end{aligned} \tag{2.13}$$

with t denoting the water temperature in ℃. In this study, the remotely sensed based surface temperature was chosen for this calculation.

For sites with available net radiation measurements, method ETd2med (Eq. 2.14) followed the same fundamentals as ETd1, but excludes G from the AE throughout the day based on the widely used assumption that daily G tends to zero (Allen et al., 1998), and is written as:

$$ET_{d2med} = [EF(Rn_{\text{measured}})_d] \times \left[\frac{cf}{\lambda_v \rho_w}\right] \tag{2.14}$$

When net radiation measurements throughout the day are not available, a modification of the previous method was proposed, ETd2 was determined using the Eq. 2.15.

$$ET_{d2} = [EF(Rn_{\text{estimated}})_d] \times \left[\frac{cf}{\lambda_v \rho_w}\right] \tag{2.15}$$

where $(Rn_{estimated})_d$ is the estimated daily average net radiation. This estimation was calculated even when Rn measurements were available, relying on solar radiation and temperature measurements throughout the day following the model of Samani et al. (2007), and it can be described as:

$$(Rn_{\text{estimated}})_d = \frac{Rn \times (R_s)_d}{R_s} \times \left(\frac{T_d + 273.15}{T_a + 273.15}\right)^4 \tag{2.16}$$

in which $(R_s)_d$ represents the daily average solar radiation in W m^{-2}, R_s is the hourly or half-hourly averaged solar radiation in W m^{-2}, T_d is the daily average air temperature in ℃, and T_a is the hourly or half-hourly averaged air temperature in ℃.

Similar to ETd2med, method ETd3med (Eq. 2.17) ignores G in both daily and instantaneous AE, increasing the total AE:

$$ET_{d3med} = \left[\frac{\lambda E_i}{Rn_i}\right] \times (Rn_{\text{measured}})_d \times \left[\frac{cf}{\lambda_v \rho_w}\right] \tag{2.17}$$

Similarly, when net radiation measurements are not available, ETd3 estimates daily net radiation from Eq. 2.16 according to Eq. 2.18.

$$ET_{d3} = \left[\frac{\lambda E_i}{Rn_i}\right] \times (Rn_{\text{estimated}})_d \times \left[\frac{cf}{\lambda_v \rho_w}\right] \tag{2.18}$$

Method ETd4 (Eq. 2.19) was based on the assumption that ET is well correlated and proportional to solar radiation R_s (Jackson et al., 1983).

$$ET_{d4} = \left(\frac{R_{s_d}}{R_s}\right) \times ET_i \tag{2.19}$$

where R_s is the average measured solar radiation in W m^{-2}, $(R_s)_d$ is the daily accumulated solar radiation in MJ m^{-2} d^{-1}, and ET_i is the instantaneous remotely sensed

based evapotranspiration in mm h^{-1}. Hourly average R_s values from agro-meteorological stations were used. For half-hourly time steps, interpolation was performed to fill the gaps. Finally, the short grass reference ET ratio method (ETd5) proposed by Colaizzi et al. (2006) based on Allen et al. (2002) was evaluated, which assumed that the ratio of actual to reference ET remains constant throughout the day:

$$ET_{d5} = \left(\frac{ET_i}{(ET_0)_i}\right) \times (ET_0)_d \tag{2.20}$$

considering $(ET_0)_i$ as the hourly or half-hourly grass reference evapotranspiration (mm h^{-1}), and $(ET_0)_d$ is the daily grass reference evapotranspiration in mm d^{-1}. These parameters were calculated according to the ASCE-EWRI (2005).

2.2.8 Model performance assessment and parameter adjustment

The Flux Footprint Prediction model (Kljun et al., 2015) was used to determine the spatial representativeness of eddy covariance measurements. Footprints were calculated for each flux tower considering site-specific parameters, including sensors' heights ranging from 4.5 to 7.0 m above ground level and zero-plane displacement heights defined as two-thirds of the tree canopy height at full development. For model validation purposes, the 90% contribution isopleth of the averaged footprint was used to define the source area of the measured energy fluxes. These footprints were subsequently used to extract SEB-PV model outputs from the corresponding Landsat pixels for comparison with the measurements at the energy flux tower.

The SEB-PV model was configured with site-specific inputs across the study sites based on field observations at each orchard. Mean leaf widths were set at 12 cm for Larqui and Pullami, and at 4 cm for Flores and Nichols. Average crop heights were set at 4.0 m for Larqui and Pullami, while Flores exhibited a height of 2.9 m and Nichols 4.7 m. The canopy resistance coefficients $c2$ and $c3$ were uniformly set at 0.00333 and 300 (dimensionless), respectively. Additional details on other constants can be found in Lagos et al. (2012) and

Souto et al. (2019).

To account for the alternate bearing characteristic observed in fruit orchard production (Singh, 1948), where data were available, a two-year calibration period was established. In the cases of Larqui and Nichols, five years of data were collected and three years were used for validation at each site. For Pullami, only four years of data were available; however, due to the limited data from the 2019–2020 irrigation season, which had only two usable measurements, data from the 2018–2019 season were utilized for calibration. Finally, for the Flores site, only one season of data was available, which allowed for calibration but not for independent validation for instantaneous flux analysis. However, this single season was sufficient for daily ET analysis since temporal upscaling methods employ mathematical algorithms that do not require calibration and directly utilize the already-calibrated instantaneous parameters. Results reported for Flores represent calibration performance for instantaneous fluxes and algorithmic application of upscaling methods for daily ET.

One parameter at a time was varied during the sensitivity analysis. Given the lower sensitivity and importance of $c1$, $beta$ was varied first (0.5 to 29.5 in increments of 0.5) while $c1$ remained constant, followed by $c1$ variation over the same range. Each curve was optimized using the minimize function from the optimization module of the SciPy library, implemented in Python 3.8.3.

For this purpose, the normalized root mean squared error (NRMSE), which measures the average distance between the estimations and the measured values, and the Kling-Gupta efficiency (KGE), which is an indicator of goodness-of-fit, were calculated according to Eq. 2.21 and Eq. 2.22.

$$NRMSE(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(X_{\text{OBS}} - X_{\text{EST}})^2}{n}}/\bar{X}_{\text{OBS}} \tag{2.21}$$

where X_{OBS} corresponds to the in-situ measurements recorded in the field, $\bar{X}_{\text{OBS}}$ is the mean value of the measurements, $X_{(EST)}$ corresponds to the estimated values of the

SEB-PV model, and n represents the number of data points considered in the analysis.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2.22}$$

where the square root term represents the Euclidean distance from the ideal point, α is a measure of relative variability in the simulated and measured values, β represents bias and is the ratio between the simulated mean and the observed mean, and r is the correlation coefficient.

The RMSE, which has the same units as the variables and is the most commonly used indicator to report uncertainty in evapotranspiration (Tran et al., 2023), can be obtained by multiplying the NRMSE in its fractional form by the average of the field-measured data. Therefore, it was reported for comparison with other studies.

Relying on a single performance metric for model evaluation has recognized limitations (Liemohn et al., 2021). Optimizing a model to improve one performance aspect, such as predictive accuracy (e.g., NRMSE), can inadvertently degrade its ability to represent other crucial behaviors (Jackson et al., 2019). To address this trade-off, the Model Decision Making Indicator (MDMI) was developed as a composite metric that integrates prediction accuracy (NRMSE) with comprehensive model behavior assessment (KGE) (Eq. 2.23).

$$MDMI = f_1 \cdot (100 \cdot KGE) - f_2 \cdot NRMSE \tag{2.23}$$

where f_1 and f_2 are weighting factors that control the relative importance of each metric in the combined assessment. Given the importance of accurate ET estimation for irrigation management and KGE's comprehensive evaluation of correlation, bias, and variability (Gupta et al., 2009), equal weights ($f_1 = f_2 = 1$) were chosen to balance prediction accuracy with overall model behavior.

Parameter sensitivity was quantified using mean absolute deviation (MAD) from the optimal performance state as detailed in Eq. 2.24.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_{opt}| \tag{2.24}$$

where x_i represents each evaluation metric value across all $beta$ parameter values tested, x_{opt} is the evaluation metric value at the best-performing $beta$ setting, and n is the total number of $beta$ values evaluated (59 values ranging from 0.5 to 29.5). This metric quantifies the expected magnitude of deviation from the reference performance when the $beta$ parameter deviates from its calibrated setting, providing a direct measure of parameter sensitivity cost across different performance metrics. MAD was chosen over standard deviation as it provides a more robust measure less sensitive to outliers and offers direct interpretation of average expected deviation from the reference performance state.

2.3 Results

2.3.1 Model calibration and analysis of energy flux tower's footprint

The footprint analysis of energy flux towers revealed distinct source area patterns across study sites (Figure 2.7), with spatially-explicit footprints used to extract the SEB-PV model estimates from corresponding Landsat pixels for validation. The calibration process systematically varied $beta$ from 0.5 to 29.5 while keeping $c1$ constant at 5.0 (followed by optimization of $c1$ across the same range), yielding $beta$ values of 4.5, 3.0, and 3.5 for NRMSE, KGE, and MDMI, respectively.

The MDMI indicator demonstrated superior parameter discrimination compared to individual metrics. While $beta$ optimization showed limited variation for NRMSE (35.3% to 47.2%) and KGE (0.68 to 0.83), MDMI exhibited a substantially wider range (21.1 to 47.6), providing 2.2 $\times$ greater sensitivity for parameter optimization compared to individual metrics (Figure 2.8). This broader range improves discrimination between parameter sets when performance differences are minimal.

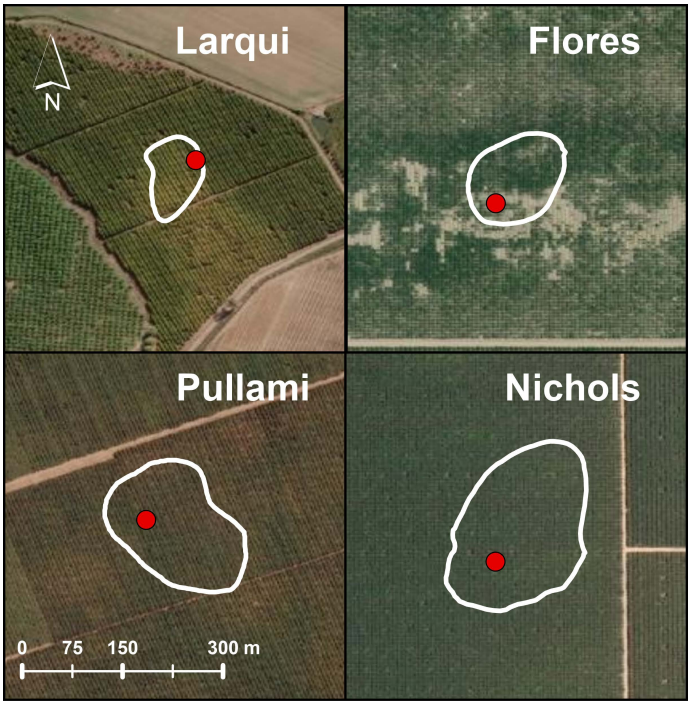

Figure 2.7: Footprint analysis of the energy flux towers at the study sites. The 90% source area contribution isopleths for measured fluxes are shown, indicating the areas used for SEB-PV model validation.

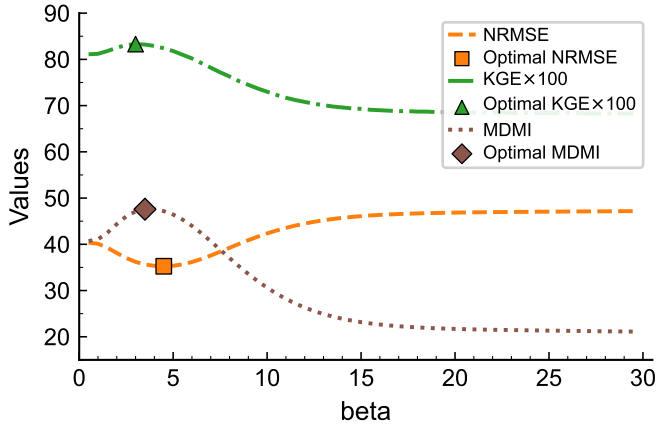

Figure 2.8: Sensitivity of SEB-PV model performance metrics (NRMSE, 100 $\times$ KGE, and MDMI) to the $beta$ parameter during calibration at the Flores site. Results demonstrate the enhanced sensitivity of the proposed MDMI metric for parameter optimization (dimensionless).

2.3.2 Model performance using gridded soil moisture inputs

Both CFSv2 and SMAP L4 achieved comparable accuracy to measured soil moisture after parameter adjustment (Table 2.4). Model performance after calibrating $beta$ and $c1$ parameters differed by $<$0.3 MDMI from measured soil moisture at Chile sites but showed larger differences ($\sim$ 6 MDMI) at California sites when using gridded inputs. Parameter $beta$ required significant adjustment depending on soil moisture source, with California sites requiring larger adjustments than Chilean sites, indicating that $beta$ functions as a correction factor when using gridded products.

Figure 2.9 shows that model performance using gridded products inputs resulted in slightly higher scatter in energy flux estimation while maintaining comparable overall performance. The model systematically underestimates sensible (H) and soil heat flux (G), leading to overestimation of latent heat flux (λE). The site-disaggregated results (Figure 2.13) showed regional differences : Chile sites showed consistent net radiation underestimation, while California sites exhibited greater scatter governed by poor soil heat flux correlation, which compromises instantaneous energy balance partitioning with implications for temporal upscaling algorithms discussed in section 2.4.3.

CFSv2 outperformed SMAP L4 with consistently higher MDMI values and demonstrated lower $beta$ parameter variability, making it the preferred choice for operational implementation given its longer historical record.

2.3.3 Model performance with automated weather stations

Model evaluation using automated weather station data instead of in-situ measurements revealed differential sensitivity across climatic conditions (Table 2.5). While both measurement approaches record the same meteorological variables, they represent fundamentally different environments: in-situ sensors capture the modified microclimate within orchard canopies (influenced by canopy architecture, transpiration, and soil-plant-atmosphere interactions), whereas automated weather stations measure standardized

Table 2.4: SEB-PV model optimal parameters ($c1$, $beta$) and instantaneous SEB performance metrics (NRMSE, KGE, MDMI) using meteorological inputs from in-situ measurements combined with different soil moisture datasets (measured, CFSv2, and SMAP L4).

Dataset	Subset	$c1$	$beta$	NRMSE (%)	KGE	MDMI
Measured	(a) Larqui calibration	6.0	29.0	13.6 ± 8.6	0.96 ± 0.05	82.3 ± 13.3
	validation			22.5 ± 4.9	0.93 ± 0.04	70.4 ± 8.5
	(b) Pullami calibration	7.0	26.5	16.4 ± 1.1	0.91 ± 0.01	74.3 ± 2.4
	validation			20.1 ± 1.4	0.96 ± 0.02	75.4 ± 3.6
	(c) Flores calibration	29.5	2.5	34.4 ± 9.6	0.85 ± 0.13	50.8 ± 22.6
	validation			-	-	-
	(d) Nichols calibration	6.0	8.5	22.6 ± 5.1	0.95 ± 0.04	72.5 ± 9.3
	validation			25.4 ± 3.1	0.91 ± 0.03	65.8 ± 5.0
CFSv2	(a) Larqui calibration	29.5	29.5	15.3 ± 10.4	0.96 ± 0.05	80.3 ± 15.5
	validation			22.8 ± 2.3	0.93 ± 0.04	70.5 ± 5.9
	(b) Pullami calibration	7.0	19.0	16.1 ± 0.8	0.91 ± 0.02	74.5 ± 2.0
	validation			19.9 ± 1.2	0.95 ± 0.02	75.4 ± 3.4
	(c) Flores calibration	24.5	5.5	27.9 ± 9.9	0.91 ± 0.11	62.8 ± 20.9
	validation			-	-	-
	(d) Nichols calibration	5.5	14.0	25.4 ± 3.8	0.94 ± 0.04	68.8 ± 7.8
	validation			27.8 ± 6.3	0.91 ± 0.04	63.4 ± 9.5
SMAP L4	(a) Larqui calibration	29.5	29.5	15.5 ± 10.2	0.96 ± 0.05	80.1 ± 14.9
	validation			23.5 ± 2.1	0.94 ± 0.04	70.1 ± 5.3
	(b) Pullami calibration	7.0	15.5	15.8 ± 0.9	0.91 ± 0.01	74.7 ± 1.9
	validation			19.5 ± 1.3	0.95 ± 0.02	75.4 ± 3.3
	(c) Flores calibration	29.5	10.5	30.8 ± 4.0	0.88 ± 0.04	57.1 ± 8.3
	validation			-	-	-
	(d) Nichols calibration	29.5	29.5	27.1 ± 4.5	0.93 ± 0.04	66.2 ± 9.0
	validation			30.5 ± 6.2	0.90 ± 0.04	59.2 ± 9.7

Values after $\pm$ represent parameter sensitivity estimates (mean absolute deviation from optimal parameter value) quantifying model sensitivity to $beta$ parameter calibration. (a), (b), (c) and (d) represent study sites.

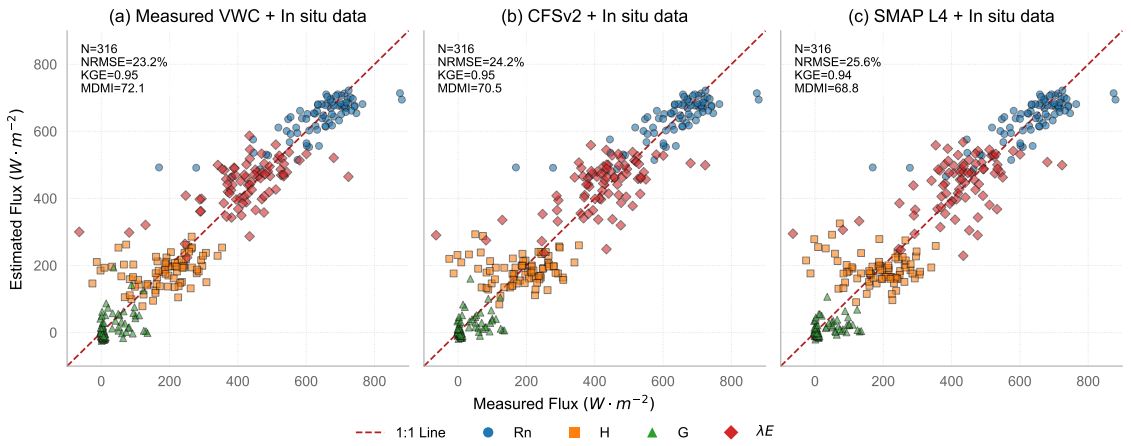

Figure 2.9: Comparison of measured vs. SEB-PV modeled instantaneous energy balance components (Rn, H, G, and λE) during validation. Model inputs include half-hourly meteorological data from in-situ measurements combined with different soil moisture sources: (a) measured, (b) CFSv2, and (c) SMAP L4.

reference conditions over uniform surfaces that characterize broad-area atmospheric conditions. Correspondence analysis between these contrasting measurement environments at satellite overpass times revealed regional differences, with Chile sites exhibiting stronger agreement (temperature R^2 = 0.93, relative humidity R^2 = 0.67) compared to California sites (temperature R^2 = 0.34, relative humidity R^2 = 0.50). This differential correspondence directly explains the observed model performance patterns: stable performance at Chile sites (MDMI constant at Larqui, +1.2 improvement at Pullami) versus significant degradation at California sites (MDMI decreased 15.4 at Flores, 6.1 at Nichols). The weaker correspondence in California reflects the greater divergence between atmospheric conditions and orchard microclimate under semi-arid conditions, where high vapor pressure deficit amplifies the biophysical coupling between crop transpiration and atmospheric demand, making orchard-specific meteorological conditions critical for accurate ET modeling.

The operationally relevant configuration combining CFSv2 soil moisture with weather station data maintains acceptable accuracy while significantly expanding model applicability

Table 2.5: SEB-PV model optimal parameters ($c1$, $beta$) and instantaneous SEB performance metrics (NRMSE, KGE, MDMI) using meteorological inputs from nearby weather stations combined with gridded soil moisture datasets (CFSv2 and SMAP L4).

Dataset	Subset	$c1$	$beta$	NRMSE (%)	KGE	MDMI
CFSv2	(a) Larqui calibration	15.5	29.5	15.0 ± 8.1	0.96 ± 0.05	80.6 ± 15.0
	validation			22.7 ± 4.3	0.93 ± 0.04	70.5 ± 6.5
	(b) Pullami calibration	5.0	19.0	13.9 ± 1.0	0.91 ± 0.02	77.3 ± 2.6
	validation			19.2 ± 2.8	0.96 ± 0.03	76.4 ± 5.1
	(c) Flores calibration	28.0	8.5	36.0 ± 5.2	0.82 ± 0.06	45.8 ± 11.4
	validation			-	-	-
	(d) Nichols calibration	5.5	14.5	26.9 ± 6.4	0.92 ± 0.05	65.1 ± 10.2
	validation			30.7 ± 8.0	0.88 ± 0.05	57.5 ± 11.5
SMAP L4	(a) Larqui calibration	17.0	29.5	15.1 ± 7.5	0.96 ± 0.05	80.6 ± 14.2
	validation			23.8 ± 3.1	0.94 ± 0.04	70.0 ± 5.4
	(b) Pullami calibration	5.0	20.0	13.9 ± 1.0	0.91 ± 0.02	77.4 ± 2.4
	validation			18.7 ± 3.0	0.96 ± 0.03	76.8 ± 5.1
	(c) Flores calibration	28.5	18.5	36.9 ± 4.3	0.80 ± 0.03	43.3 ± 7.6
	validation			-	-	-
	(d) Nichols calibration	29.5	29.0	26.6 ± 3.5	0.92 ± 0.01	65.5 ± 4.1
	validation			33.0 ± 3.9	0.86 ± 0.01	53.0 ± 4.1

Values after $\pm$ represent parameter sensitivity estimates (mean absolute deviation from optimal parameter value) quantifying model sensitivity to $beta$ parameter calibration. (a), (b), (c) and (d) represent study sites.

to regions lacking field measurements (Figure 2.10(b)), though requiring $beta$ parameter calibration.

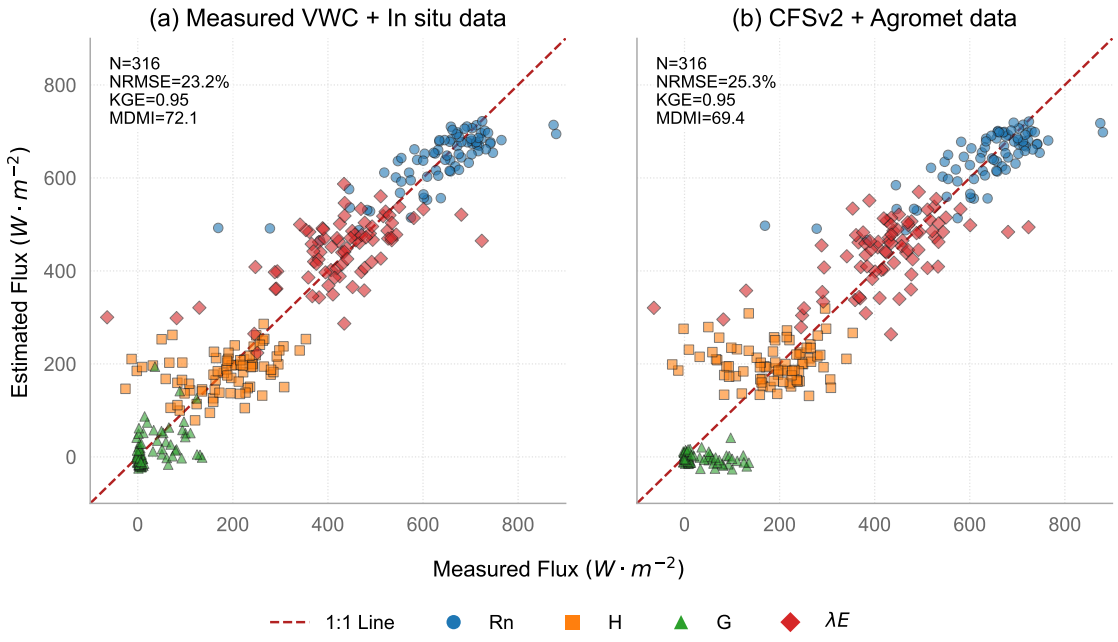

Figure 2.10: Comparison of SEB-PV modeled instantaneous energy balance components (Rn, H, G, and λE) against measurements during validation, using: (a) fully in-situ inputs (measured soil moisture and half-hourly in-situ meteorological data), and (b) globally available inputs (CFSv2 soil moisture and hourly data from nearby weather stations).

2.3.4 Temporal evapotranspiration upscaling methods

For temporal upscaling analysis, the optimal $beta$ parameter from instantaneous calibration was applied to all upscaling methods without recalibration. Temporal upscaling accuracy varied significantly across sites due to systematic differences in soil heat flux model performance and contrasting atmospheric demand conditions (Table 2.6). For the Chile orchards in Mediterranean conditions, Group 1 algorithms (requiring net radiation) substantially outperformed Group 2 approaches (that rely on meteorological variables): average MDMI $\sim$ 53 vs. $<$28, indicating diurnal ET patterns driven predominantly

by available energy. ETd2med achieved optimal hazelnut performance with RMSE 0.8 $\pm$0.2 mm d^{-1}.

Conversely, at California pistachio sites, G2 algorithms performed better under substantially higher atmospheric demand (ET_0 = 1,443.5 mm during growing season) compared to Chilean sites (792.4 mm), which increases the likelihood of any source of water stress (both management-induced and unintentional stress) and enhances vapor pressure deficit control over transpiration. Additionally, systematic soil heat flux underestimation (measured: 146.1 vs. estimated: 39.3 W m^{-2} at Flores; 65.9 vs. 34.0 W m^{-2} at Nichols) forces compensatory overestimation of sensible and latent heat fluxes, compromising the $\lambda E/Rn$ relationship fundamental to G1 temporal upscaling algorithms. The combination of higher water stress likelihood and soil heat flux model limitations makes G2 algorithms more suitable because they rely on meteorological variables that better capture physiological responses under these conditions rather than energy balance partitioning. ETd4 achieved superior performance with RMSE of 1.3 $\pm$ 0.5 mm d^{-1}.

Figure 2.11 demonstrates that no single algorithm group dominates across both orchard types. Both groups exhibit systematic biases: G1 methods generally underestimate daily ET while G2 methods tend toward overestimation. Although temporal upscaling performance depends on underlying instantaneous flux estimation accuracy, the poor performance of G2 methods over study sites, specifically hazelnut orchards despite good instantaneous flux estimates, reveals fundamental limitations of these models for specific orchard crops.

2.4 Discussion

2.4.1 Model performance with alternative soil moisture products

Despite poor representativeness between gridded and measured soil moisture (Nash-Sutcliffe efficiency $<$0.2 for most site-product combinations), parameter adjustment enabled comparable model performance. The systematic underestimation during irrigation periods

Table 2.6: Performance of seven temporal upscaling methods for estimating daily evapotranspiration (ET) from instantaneous SEB-PV model outputs. Methods are grouped by input data requirements (G1: requiring measured net radiation; G2: using meteorological data from weather stations).

Model	**Group**	**RMSE (mm d^{-1})**	**NRMSE (%)**	**KGE**	**MDMI**
(a) Larqui					
ETd1	G1	0.9 ± 0.2	21.7 ± 5.7	0.67 ± 0.06	45.1 ± 11.5
ETd2med	G1	0.9 ± 0.2	21.9 ± 5.5	0.68 ± 0.06	45.6 ± 11.1
ETd3med	G1	0.9 ± 0.3	22.5 ± 6.4	0.67 ± 0.07	44.5 ± 13.4
ETd2	G2	1.3 ± 0.2	32.4 ± 4.6	0.56 ± 0.05	23.9 ± 9.6
ETd3	G2	1.4 ± 0.2	33.4 ± 5.5	0.56 ± 0.07	22.8 ± 12.2
ETd4	G2	1.5 ± 0.2	36.7 ± 5.8	0.55 ± 0.07	18.0 ± 12.2
ETd5	G2	1.8 ± 0.3	42.9 ± 7.7	0.47 ± 0.08	4.4 ± 15.0
(b) Pullami					
ETd1	G1	0.7 ± 0.0	12.2 ± 0.5	0.73 ± 0.01	60.6 ± 0.7
ETd2med	G1	0.7 ± 0.0	12.5 ± 0.4	0.74 ± 0.01	61.5 ± 0.7
ETd3med	G1	0.7 ± 0.1	13.8 ± 1.2	0.72 ± 0.01	58.3 ± 2.0
ETd2	G2	1.3 ± 0.0	23.3 ± 0.3	0.46 ± 0.02	23.1 ± 2.2
ETd3	G2	1.3 ± 0.0	23.7 ± 0.6	0.44 ± 0.01	20.5 ± 1.8
ETd4	G2	1.4 ± 0.0	26.1 ± 1.0	0.43 ± 0.02	16.6 ± 3.0
ETd5	G2	1.7 ± 0.1	31.3 ± 2.4	0.43 ± 0.03	11.7 ± 5.1
(c) Flores					
ETd1	G1	1.9 ± 1.0	47.9 ± 24.7	0.40 ± 0.32	-8.4 ± 56.7
ETd2med	G1	1.8 ± 0.9	45.6 ± 21.8	0.44 ± 0.35	-1.8 ± 56.4
ETd3med	G1	1.9 ± 1.0	48.1 ± 24.7	0.40 ± 0.39	-7.8 ± 63.8
ETd2	G2	1.3 ± 0.5	34.3 ± 11.5	0.55 ± 0.08	20.2 ± 18.3
ETd3	G2	1.4 ± 0.4	35.7 ± 10.9	0.51 ± 0.07	15.1 ± 16.2
ETd4	G2	1.3 ± 0.7	33.0 ± 16.8	0.53 ± 0.10	20.2 ± 26.1
ETd5	G2	0.9 ± 1.9	22.0 ± 49.1	0.74 ± 0.45	52.5 ± 94.0
(d) Nichols					
ETd1	G1	1.6 ± 0.3	29.7 ± 6.1	0.56 ± 0.05	26.5 ± 10.9
ETd2med	G1	1.6 ± 0.3	29.7 ± 5.8	0.58 ± 0.05	28.3 ± 10.9
ETd3med	G1	1.6 ± 0.4	31.2 ± 6.8	0.59 ± 0.07	27.5 ± 13.6
ETd2	G2	1.3 ± 0.2	25.1 ± 2.6	0.57 ± 0.03	31.9 ± 5.7
ETd3	G2	1.2 ± 0.2	23.3 ± 3.9	0.62 ± 0.04	39.1 ± 8.2
ETd4	G2	1.2 ± 0.2	23.4 ± 4.3	0.65 ± 0.05	41.6 ± 9.1
ETd5	G2	1.8 ± 0.3	33.5 ± 6.1	0.66 ± 0.07	32.3 ± 12.9

Values after $\pm$ represent parameter sensitivity estimates (mean absolute deviation from optimal parameter value) indicating expected performance variability under suboptimal parameter selection. (a), (b), (c) and (d) represent study sites.

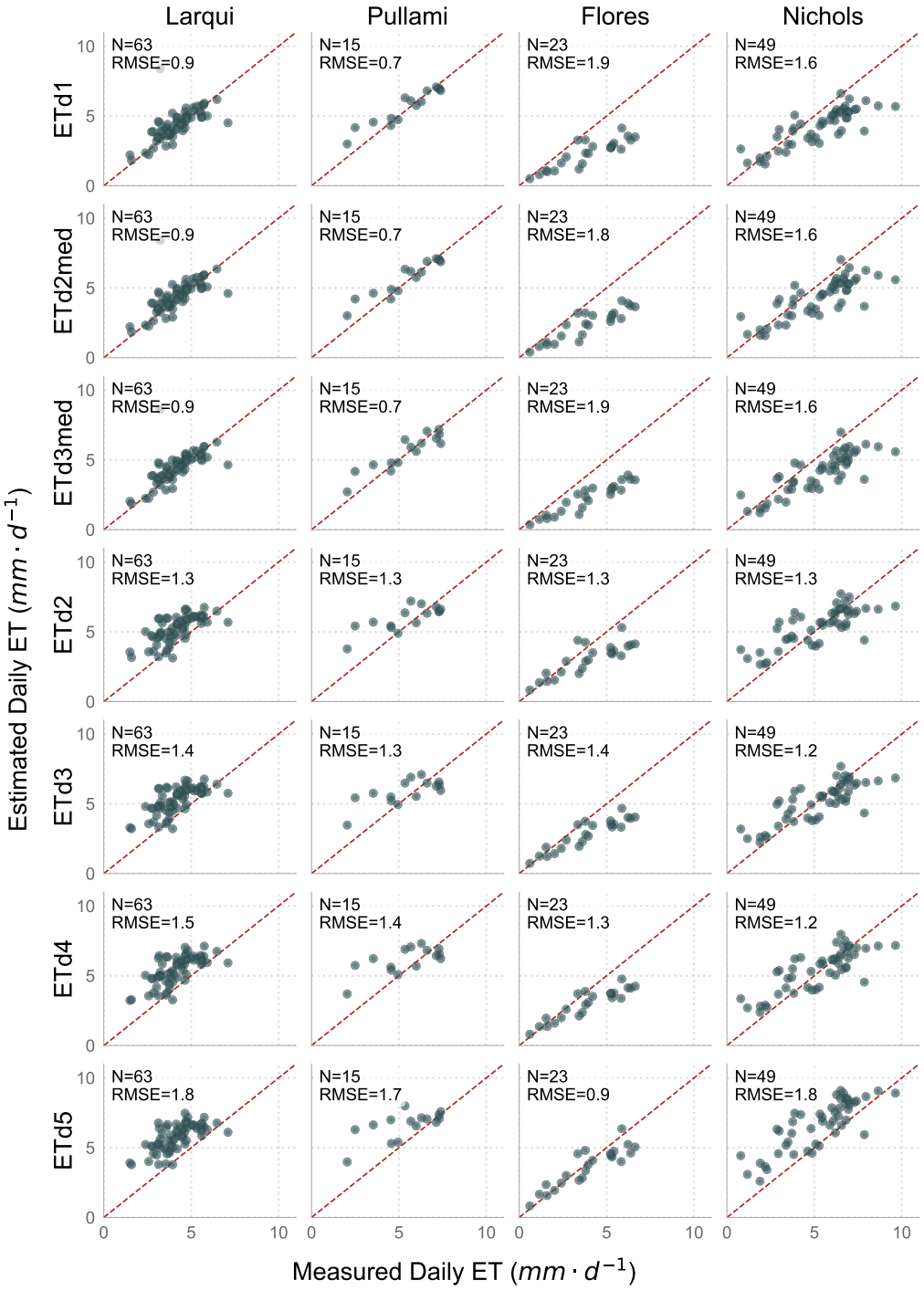

Figure 2.11: Site-by-site comparison of measured vs. estimated daily evapotranspiration (ET) using seven temporal upscaling methods. The dashed red line represents the 1:1 relationship.

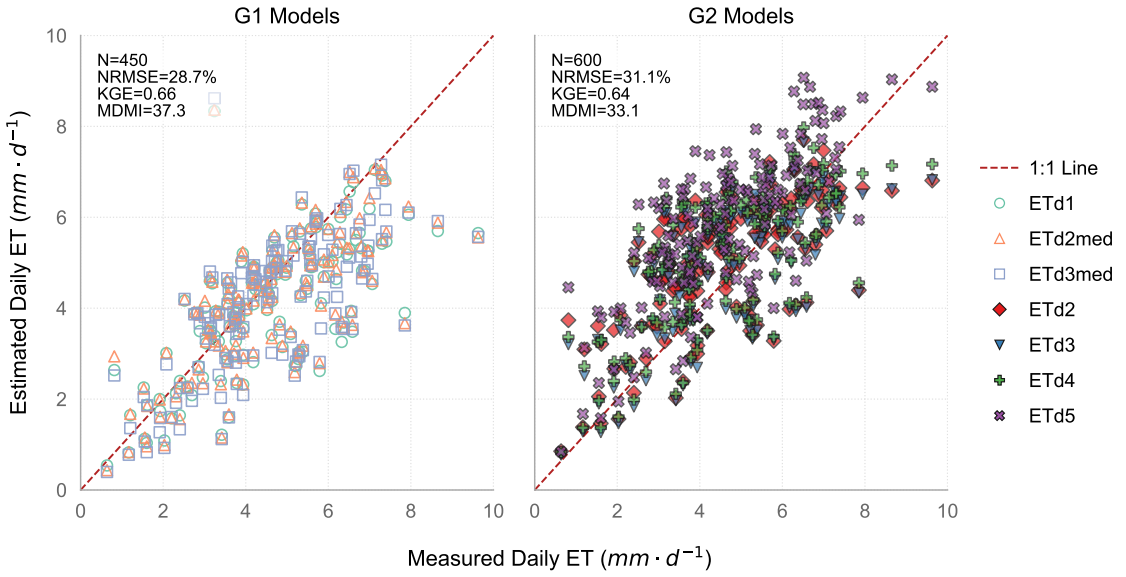

Figure 2.12: Aggregate performance of temporal upscaling method alternatives for daily ET estimation across study sites. Comparison of G1 algorithms (requiring measured net radiation; left panel) vs. G2 algorithms (using data from weather stations; right panel).

reflects limitations of coarse spatial resolution products that rely predominantly on gauge-based precipitation observations (section 2.2.6).

The Flores site demonstrated improved performance with gridded products compared to measured soil moisture, illustrating the trade-off between point-source measurements and spatially integrated estimates. Although average model performance remains comparable when using gridded soil moisture inputs, the systematic underestimation of soil heat flux (G) at California sites compromises instantaneous energy balance partitioning accuracy, as evident in the 1:1 comparison plots (Figure 2.13) where latent heat flux (λE) exhibits compensatory overestimation. This degradation in soil heat flux estimation performance worsens with gridded soil moisture products, highlighting the need for enhanced soil thermal dynamics representation in SEB-PV.

Large net radiation estimation offsets at Chilean sites (118% overestimation at Larqui on March 27-28, 2022; 20% underestimation at Pullami on November 28, 2021) resulted

from cloud variability during measurement periods rather than model structural limitations, since satellite imagery captures instantaneous conditions (11:34-11:40 local time) while flux measurements represent 30-minute averages.

2.4.2 Operational implications

The demonstrated performance using globally available datasets (CFSv2 + weather stations) with RMSE of 1.2 mm d^{-1} represents a 37% improvement over METRIC's reported 1.9 mm d^{-1} for Nichols (Jin et al., 2018), enabling water management with parameter sensitivity estimates of $\pm$ 0.2 mm d^{-1}, though total model uncertainty may be higher when considering all error sources. This expands SEB-PV applicability to regions where direct measurements are unavailable.

2.4.3 Temporal upscaling methods

The crop-specific performance differences reflect fundamental energy balance dynamics: hazelnuts respond to energy-driven processes captured by G1 algorithms, while pistachios in semi-arid conditions exhibit stronger VPD control better represented by G2 methods. Systematic soil heat flux underestimation, affecting instantaneous SEB partition estimates (H and λE), propagates errors more severely through G1 than G2 algorithms, explaining the regional performance reversal.

These findings complement previous studies that have sought universal upscaling approaches. While Cammalleri et al. (2014) identified ETd4 as robust across diverse land covers, Chávez et al. (2008) favored ETd1 for annual crops, and Colaizzi et al. (2006) demonstrated superior performance of grass reference ET methods (ETd5) in field crops, this study reveals that optimal methods vary with orchard crop and environmental conditions. The poor performance of the widely successful ETd4 and reference ET-based ETd5 algorithms in hazelnut orchards underscores the critical importance of crop-specific upscaling algorithm selection. This finding aligns with Jiang et al. (2021), who reported

correction factors for G1 algorithms ranging from 1.12 to 1.34 across different ecosystems at satellite overpass times, highlighting inherent variability in temporal scaling relationships.

The anomalous performance at Flores—where ETd5 achieved an MDMI of 52.5 while G1 methods averaged -6.0 despite showing the poorest instantaneous performance among all sites (Figure 2.13)—warrants careful interpretation. This result exhibits high parameter sensitivity, evidenced by the large mean absolute deviation in RMSE ($\pm$ 1.9 mm d^{-1}), with accuracy declining significantly when the $beta$ parameter is not carefully calibrated. Furthermore, the poor transferability of the ETd5 method to other orchards (Larqui, Pullami, Nichols RMSE $\sim$ 1.8 mm d^{-1}) demonstrates that this widely used reference ET-based algorithm performs poorly in these orchard environments despite well-calibrated instantaneous surface energy balance, highlighting that algorithm performance varies considerably even within the same crop type under different local conditions. This case reinforces the fundamental principle that improving instantaneous energy balance estimation is prerequisite to obtaining reliable daily ET estimates, while also demonstrating that orchard crops require specialized temporal upscaling approaches rather than universal methods. Parameter sensitivity analysis (Table 2.6) provides practical guidance for operational implementation, with algorithms showing lower beta parameter sensitivity being more resilient to suboptimal calibration conditions.

2.4.4 Practical implications and limitations

The demonstrated performance with globally available datasets enables water managers to:

- apply SEB-PV model using CFSv2 soil moisture and weather station data where measured soil moisture and in-situ meteorological measurements are unavailable
- use G1 algorithms for hazelnut orchard when net radiation measurements are available (RMSE $\sim$ 0.8 mm d^{-1})

- use G2 algorithms (ETd4) for pistachio orchards in semi-arid conditions (RMSE $\sim$1.3 mm d^{-1})
- achieve daily ET estimates with parameter sensitivity estimates of $\pm$ 0.2 mm d^{-1} in hazelnut orchard and up to $\pm$ 0.5 mm d^{-1} in pistachio orchard

Several limitations of this study should be highlighted. First, using static seasonal footprints for validation may not capture day-to-day variations in flux source areas. Second, systematic soil heat flux underestimation at California sites represents a fundamental model limitation that requires attention in future SEB-PV development. Third, coarse-resolution gridded products cannot capture field-scale irrigation events, though no downscaling was undertaken to preserve the globally applicable approach. Fourth, parameter sensitivity estimates ($\pm$) represent mean absolute deviation from optimal parameter values, quantifying model sensitivity to parameter calibration but not encompassing input data errors, model structural uncertainty, or measurement uncertainties. Fifth, the temporal upscaling analysis was conducted using measured soil moisture only to isolate temporal scaling uncertainties from soil moisture input uncertainties. This approach limits direct assessment of how instantaneous flux errors using gridded soil moisture products propagate to daily ET estimates.

Future research should focus on improving SEB-PV's soil thermal dynamics representation and developing robust temporal upscaling algorithms specifically for orchard crops. Additionally, sensitivity analysis of MDMI weighting factors could optimize performance for different cropping systems and weather conditions.

2.5 Conclusions

This study evaluated SEB-PV model performance when using gridded soil moisture data and seven temporal upscaling approaches in hazelnut and pistachio orchards, addressing (1) the viability of gridded soil moisture products as alternatives to in-situ measurements,

and (2) evaluating the performance of seven temporal upscaling methods for converting instantaneous to daily ET estimates.

SEB-PV model performance using both CFSv2 and SMAP L4 soil moisture products achieved comparable accuracy to the baseline configuration using in-situ soil moisture measurements after parameter adjustment. Validation MDMI values exceeded 70 in hazelnut orchards (NRMSE $\sim$ 21%) and 57 for pistachios (NRMSE $\sim$ 29%). CFSv2 is recommended over SMAP L4 due to superior performance and longer historical availability.

Method selection depends on orchard crop and weather conditions. In hazelnut orchards in Mediterranean conditions, available energy-based methods (G1) substantially outperformed meteorological-based methods (G2), with ETd2med achieving best performance (MDMI $>$45.6; RMSE 0.8 mm d^{-1}). In pistachio orchards in semi-arid conditions, where both algorithm groups showed limited accuracy, G2 algorithms were less problematic, with ETd4 being the best available option (MDMI $>$20.2; RMSE 1.3 mm d^{-1}).

The model maintained acceptable performance when using automated weather station data combined with gridded soil moisture products, reducing dependency on specialized field sensors while maintaining functionality for water management applications. These findings demonstrate that gridded soil moisture products can provide acceptable accuracy for orchard ET estimation when combined with appropriate parameter calibration, expanding SEB-PV's geographic applicability despite inherent limitations of coarse spatial resolution. The approach offers a practical solution for irrigation water management in orchard crop systems facing water scarcity, particularly in regions with limited field monitoring capabilities.

Data availability

The SEB-PV model outputs used in this study, including instantaneous model inputs and various temporal upscaling methods, are publicly available at Zenodo repository

https://doi.org/10.5281/zenodo.15476997.

Meteorological data used as model inputs are accessible through their respective institutional repositories:

For Chilean sites (Larqui and Pullami): Chilean Agricultural Research Institute agrometeorological network (https://agrometeorologia.cl) For California sites (Flores and Nichols): California Irrigation Management Information System (https://cimis.water.ca.gov)

The soil moisture gridded products used in this study are available from their original sources:

SMAP Level-4 data: NASA National Snow and Ice Data Center (https://doi.org/10.5067/EVKPQZ4AFC4D) CFSv2 data: NOAA Climate Forecast System (https://cfs.ncep.noaa.gov)

Soil physical properties data were obtained from SoilGrids250m 2.0, available at: https://soilgrids.org

Additional data or information related to this study are available from the corresponding author upon reasonable request.

Acknowledgements

We express our gratitude to Richard Snyder, Frank Anderson, Carlos Cea and Samuel G. Metcalf for their valuable assistance with flux tower data collection and field measurements.

A Appendix. Simulated vs. Observed Instantaneous Fluxes

The comprehensive visualization in Figure 2.13 provides a side-by-side comparison of energy flux estimations across all study sites and input data configurations. This figure serves as a visual complement to the quantitative analyses presented in Tables 2.4 and 2.5, illustrating how energy balance components (Rn, H, G, and λE) are estimated

under different model configurations. The consistent clustering of data points along the 1:1 line across configurations demonstrates the robustness of the SEB-PV model when using alternative data sources. Notably, even the fully operational configuration using weather station data with gridded soil moisture products (rightmost columns) maintains comparable accuracy to baseline estimates using measured inputs (leftmost column), further supporting the model's potential for widespread application in regions lacking specialized field measurements.

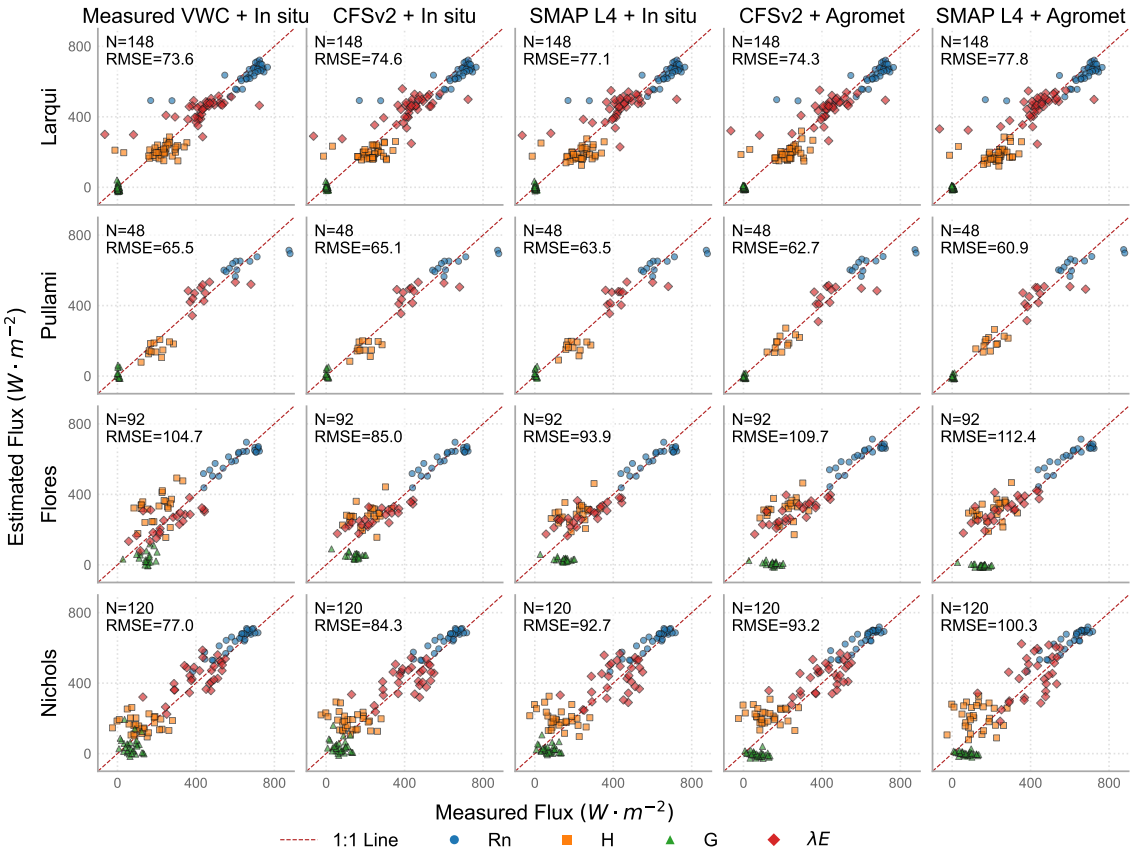

Figure 2.13: Matrix comparison of SEB-PV modeled vs. measured instantaneous energy balance components (Rn, H, G, λE) across all study sites and model input configurations. Results demonstrate consistent model performance from fully measured inputs (leftmost column) to globally available data sources (rightmost columns).

B Appendix. Estimated vs. Observed Soil Volumetric Water Content

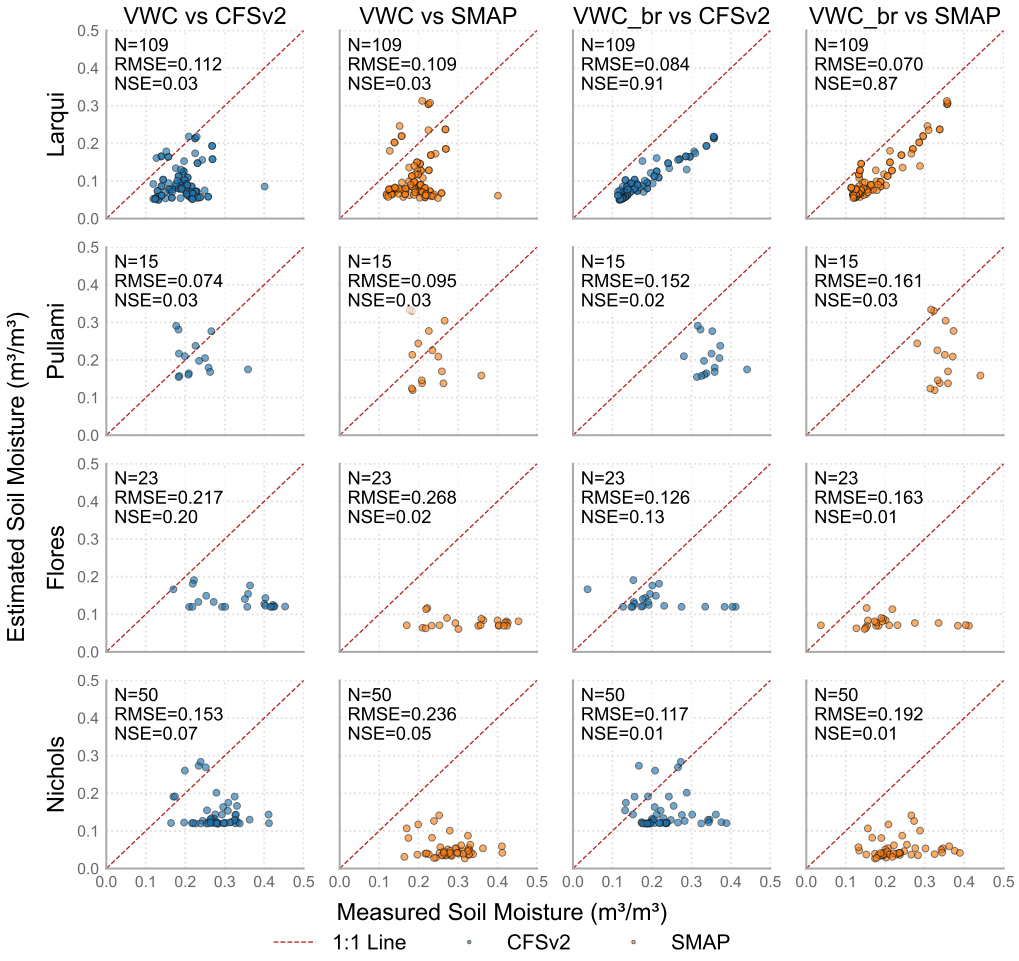

Figure 2.14: Scatter plots comparing in-situ measured volumetric water content (VWC: in-row, VWCbr: between-row) with estimates from gridded products (CFSv2 and SMAP L4) for all study sites. RMSE and NSE values are provided for each site-product combination.

CAPÍTULO 3

A Probabilistic Ensemble Framework for Evapotranspiration: Spatiotemporal Error Structure Dominates Calibration Complexity

Irrigation Science (Special Issue: Remote Sensing and Irrigation Monitoring)
Status: In review

Abstract

Satellite-derived evapotranspiration (ET) estimates systematically overestimate water use in California citrus orchards by as much as 30%, limiting operational reliability for an industry producing 92% of U.S. fresh citrus. While multi-model ensembles consistently outperform individual models, current operational platforms rely on deterministic averaging under independence assumptions, reporting accuracy metrics rather than prediction intervals and ignoring spatiotemporal error dependencies, which leads to systematically overconfident uncertainty estimates.

We developed a Bayesian Model Averaging framework to answer a fundamental question: do predictive improvements arise from sophisticated calibration architectures or from explicitly modeling spatiotemporal error structures? Combining six OpenET models with the specialized Surface Energy Balance for Partially Vegetated surfaces algorithm, we implemented parallel model suites under conditional independence versus multivariate

spatiotemporal assumptions, testing global, hierarchical, and state-dependent calibrations. Using six years (2019–2024) of flux tower data from three commercial sites and rigorous temporal cross-validation, we systematically evaluated both dimensions.

Error structure dominates calibration complexity. Multivariate frameworks yielded gains of 37–128 expected log predictive density (ELPD) units, outperforming conditional independence counterparts by 5–7 standard errors of the ELPD difference across all architectures. Spatiotemporal structure collectively accounts for 95-98% of residual variance versus 1-2% observational, with the specific partition between spatial and temporal components varying by model architecture. State-dependent intercept calibration proved optimal, maintaining accuracy under unseen weather conditions where static approaches failed.

These findings establish that operational ET ensembles must prioritize spatiotemporal error specification over calibration refinement to deliver reliable prediction intervals for risk-informed water management under intensifying climate variability.

3.1 Introduction

State-of-the-art remote sensing platforms systematically overestimate annual evapotranspiration (ET) in California's high-value citrus orchards by as much as 30% (Dhungel et al., 2024). This represents a worsening compared to foundational methods—previous findings using the FAO-56 crop coefficient approach have been shown to overestimate citrus consumptive water use by 20% when compared to eddy covariance measurements (Er-Raki et al., 2009). The increasing magnitude of these systematic biases affects the reliability of remote sensing ET estimates for operational water management in California's citrus industry, which accounts for 92% of U.S. fresh citrus production (Weber et al., 2023).

Multi-model ensemble methods often improve ET estimation relative to individual

models (Sun et al., 2019; Bhattarai et al., 2019; Bai et al., 2021), but most operational systems still rely on deterministic averaging schemes that produce single-valued estimates without prediction intervals or explicit treatment of error dependence (Tran et al., 2023). These schemes assume that model errors are independent and that data are abundant (Hastie et al., 2009)—assumptions that fail in agricultural settings where data are sparse, models share common forcing, and residuals exhibit spatial and temporal correlation (Legendre, 1993; F. Dormann et al., 2007). This produces under-dispersive ensembles that underestimate uncertainty (Raftery et al., 2005), and when uncertainty estimates are provided, they are systematically overconfident. Operational water management requires reliable prediction intervals to assess whether ET estimates are adequate for irrigation scheduling, yet current platforms provide neither intervals nor the probabilistic infrastructure to generate them.

Current operational ET platforms illustrate these constraints. The Jet Propulsion Laboratory EvapoTranspiration (JET) ensemble combines several algorithms to derive ET from spaceborne sensors, but its application is limited by the irregular coverage of the International Space Station and the absence of consistent interpolation between overpasses (Fisher et al., 2020). The OpenET platform (Melton et al., 2022) provides daily ET at 30 m resolution for the conterminous United States by integrating six models (DisALEXI, eeMETRIC, geeSEBAL, PT-JPL, SIMS, SSEBop) through simple averaging of Landsat-based estimates. These products have expanded access to field-scale ET information, yet systematic bias persists in citrus systems (Dhungel et al., 2024). In parallel, specialized models such as the Surface Energy Balance for Partially Vegetated surfaces (SEB-PV) (Lagos et al., 2012) have been developed for orchards with partial canopy cover and have shown improved agreement with in situ measurements when combined with appropriate temporal upscaling schemes (Cigarra-Guíñez et al., 2025). However, such models are not yet embedded in probabilistic ensemble frameworks that handle uncertainty and spatiotemporal dependence in a coherent way.

Moving beyond deterministic ensembles requires addressing two critical elements that govern predictive performance. The first is the **calibration architecture**: whether ensemble parameters are global, vary across sites in a hierarchical manner, or depend on environmental conditions. Global calibration assumes that a single set of parameters is valid for all locations and periods, which is convenient but may mask systematic differences among orchards (Stojanović et al., 2022). Hierarchical structures allow site-level deviations while sharing information across locations through partial pooling. State-dependent calibration allows parameters to change with meteorological variables, which may be necessary when model errors vary with energy availability, atmospheric demand, or phenological stage.

The second element is the **error structure**. Standard regression and ensemble formulations assume that residuals are independent and identically distributed after conditioning on predictors. In ET applications, this assumption is rarely tenable, because residuals are influenced by common forcing, regional atmospheric patterns, and soil moisture memory. These processes induce spatial correlation among sites and temporal autocorrelation within sites (Legendre, 1993; F. Dormann et al., 2007). If these dependencies are not represented, model comparison and calibration will be biased, and predictive intervals will be misleading.

The central challenge is identifying which component yields meaningful predictive gains: calibration refinement or error structure specification. Unjustified complexity can increase parameter uncertainty without improving predictions (Kim et al., 2023; Dubova et al., 2025), making this distinction critical for operational deployment.

The objective of this study is to quantify the relative contribution of calibration architecture and error structure specification to the performance of a satellite-based ET ensemble for citrus orchards, and to identify an ensemble configuration that can be implemented operationally with transparent uncertainty quantification. To this end, we implement a factorial design that varies both calibration architecture and error structure,

combining six OpenET models with SEB-PV within a Bayesian Model Averaging (BMA) framework. Two parallel error specifications are considered: conditional independence, in which residuals are treated as independent, and multivariate spatiotemporal, in which residuals are modeled with joint spatial correlation and temporal autocorrelation. Within each framework, we evaluate global, hierarchical, and state-dependent calibration variants that range from simple weight-only schemes to models with state-dependent bias corrections.

The analysis uses six years (2019–2024) of daily actual ET derived from eddy covariance energy balance residuals at three commercial citrus orchards in California's San Joaquin Valley. Model performance is evaluated with forward-chaining temporal cross-validation that respects time ordering, thereby mimicking an operational context where models calibrated on past data predict future conditions. Within this design, we address two research questions:

1. Does predictive improvement in ET ensembles arise primarily from calibration complexity or from explicit modeling of spatiotemporal error structure?

2. Under a multivariate spatiotemporal error model, which calibration architecture (global, hierarchical, state-dependent) provides the most reliable out-of-sample predictions for operational water management?

By separating calibration and error structure in a single framework, this study provides a basis for ensemble design choices in field-scale ET estimation and clarifies the conditions under which added model complexity leads to meaningful gains in predictive skill and uncertainty characterization.

3.2 Materials and Methods

Figure 3.1 summarizes the complete methodological framework. The following subsections detail data collection, model specifications, and evaluation protocols.

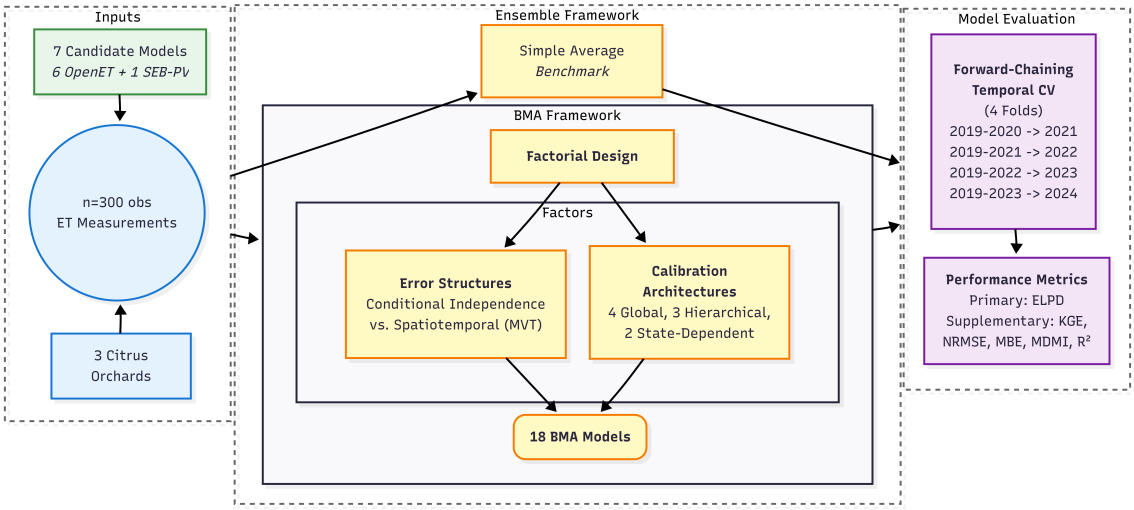

Figure 3.1: Methodological framework. Seven base ET models are integrated into both a simple average benchmark and a Bayesian Model Averaging (BMA) framework.

3.2.1 Study sites

The study was conducted at three commercial citrus orchards in Tulare County, within California's southern San Joaquin Valley, USA (Fig. 3.2; Table 3.1). The region lies in the semi-arid to Mediterranean transition zone (Köppen BSk/BSh/Csa; Beck et al., 2018), a major U.S. citrus-producing area. All orchards were under standard commercial management during 2019–2024, with irrigation via microsprinkler systems. Non-overlapping eddy covariance (EC) flux footprints ensured spatially distinct measurements (Kljun et al., 2015).

Site selection criteria: (1) location within the core citrus production zone, (2) continuous multiyear measurements enabling assessment of seasonal and interannual variability, and (3) representation of operational contrasts including crop type (Navel oranges vs. Page mandarins) and row orientation (east–west vs. north–south).

The regional climate is characterized by hot, dry summers and mild, wet winters. Annual precipitation during 2019–2024 ranged from 200 to 500 mm, concentrated between November and March, while summer temperatures frequently exceeded 30°C. Reference

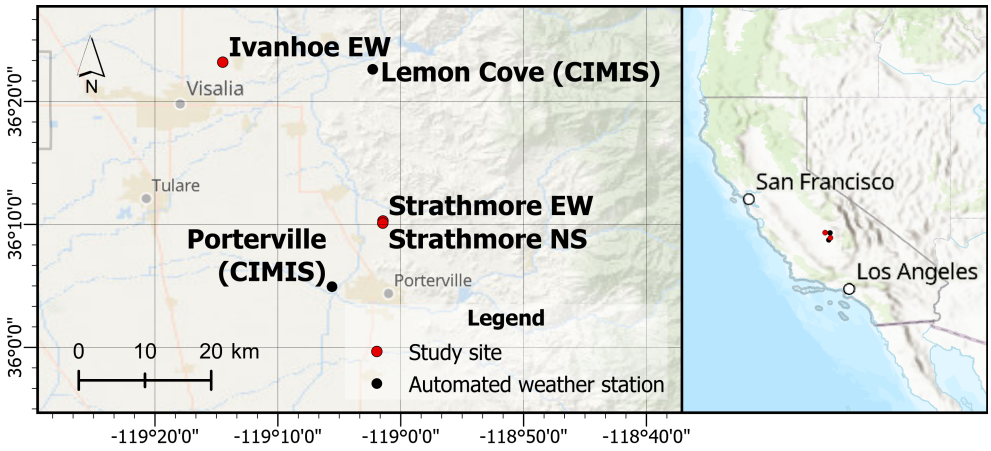

Figure 3.2: Geographic locations of the three citrus orchard study sites in Tulare County, California's southern San Joaquin Valley.

ET ranged from 1,360 to 1,440 mm yr^{-1}, exceeding precipitation and requiring irrigation throughout the growing season.

Table 3.1: Geographic coordinates and elevation of study sites.

Station name	Crop type	Latitude (°N)	Longitude (°W)	Elevation (m a.s.l.)
IvanhoeEW	Navel oranges	36.387	119.241	104
StrathmoreEW	Page mandarins	36.171	119.024	145
StrathmoreNS	Page mandarins	36.168	119.024	145

Site 1: IvanhoeEW (Navel oranges) A mature 20-year-old Navel orange orchard with east–west row orientation. Tree spacing is 6.4 m $\times$ 3.0 m (rows $\times$ trees), corresponding to 521 trees ha^{-1}. Mean canopy height is 3.9 m, with fractional canopy cover near 60%.[1]

Site 2: StrathmoreEW (Page mandarins) A 20-year-old Page mandarin orchard with east–west row orientation. Tree spacing is 6.1 m $\times$ 3.1 m, yielding 529 trees ha^{-1}. Mean canopy height is 3.9 m and fractional cover about 60%. Soil profile consists of loam (0–0.30 m) over sandy loam, with estimated rooting depth of 0.41 m.

[1] Detailed soil profile characterization was not available for this site.

Site 3: StrathmoreNS (Page mandarins) A Page mandarin orchard comparable in age to Site 2 but with north–south orientation. Tree spacing is 5.5 m $\times$ 2.7 m, corresponding to 673 trees ha^{-1}. Mean canopy height is 3.8 m and fractional cover approximately 65%. The soil profile is clay loam throughout (0–0.60 m), with estimated rooting depth near 0.60 m.

3.2.2 In situ measurements and instrumentation

Actual ET was estimated using the residual surface energy balance method,

$$\lambda E = R_n - H - G,$$

where R_n is net radiation, H is sensible heat flux, and G is soil heat flux, all in W m^{-2}. All sites were instrumented with identical or functionally equivalent systems.

Net radiation (R_n) was measured with NRLite2 net radiometers (Kipp & Zonen Inc., Delft, Netherlands) mounted 1.5 m above the canopy. Sensible heat flux (H) was measured with RM 81000 three-dimensional sonic anemometers (R.M. Young Inc., Traverse City, MI, USA) mounted at 4.7 m height at the Strathmore sites and 5.4 m at IvanhoeEW.

Soil heat flux (G) was measured using soil sensor packages at each site. Each package contained soil heat flux plates (HFT3 sensors, REBS, Bellevue, WA, USA) installed at 0.05 m depth, and soil temperature and moisture sensors (TEROS11) at 0.05 m. Three packages were installed along a transect perpendicular to the tree rows, and G was computed as the mean across packages.

Additional meteorological variables were obtained using TR-525M rain gauges (Texas Electronics, Dallas, TX, USA) and temperature and humidity probes mounted approximately 1.0 m above the canopy. All measurements were recorded at 30-minute intervals with CR3000 data loggers (Campbell Scientific Inc., Logan, UT, USA).

3.2.3 Data processing, quality control, and footprint analysis

Eddy covariance (EC) data were processed following established EC guidelines (Vickers and Mahrt, 1997; Mauder et al., 2013; Pastorello et al., 2020). Field operations included regular sensor maintenance, calibration checks, and inspection of system performance. Meteorological and soil variables were screened using physical range tests, statistical outlier detection, and temporal consistency checks (Papale et al., 2006; Pastorello et al., 2014). High-frequency turbulence data were processed using standard filtering and coordinate rotation procedures (Kaimal and Finnigan, 1994). Multi-threshold statistical screening with magnitude-dependent filters was applied to remove values that did not meet EC plausibility criteria. Detailed protocols are provided in Supplementary Material S1.

Short gaps in the EC time series were filled using gradient boosted tree models (Chen and Guestrin, 2016) implemented within established flux-processing workflows (Supplementary Material S2). Model training included randomized hyperparameter search and automated generation of interaction and transformation features. Gap filling followed a sequential procedure: first meteorological variables, then energy balance components, and finally sensible heat flux. Filling was limited to gaps of up to six consecutive observations (3 h) to preserve the natural temporal structure required for Bayesian spatiotemporal modeling. Longer gaps were retained as missing.

Spatial representativeness of the EC measurements was evaluated using the Flux Footprint Prediction model (Kljun et al., 2015). Planetary boundary layer height was obtained from ERA5 reanalysis (Hersbach et al., 2020), and zero-plane displacement was estimated as two-thirds of canopy height (Stull, 1988). Footprints were computed within a 1,000 m domain at 2 m resolution to adequately represent flux source areas at field scale. Daily footprints were generated using latent heat flux magnitude as weights, and analyses were restricted to the 80 % cumulative footprint region, a standard threshold in flux analyses (Klosterhalfen et al., 2023). Resulting maps are provided in Supplementary Material S3.

3.2.4 Candidate ET models for ensemble development

The ensemble was constructed from seven satellite-based ET models representing widely used operational algorithms in agricultural systems. These models served as inputs to the Bayesian frameworks in Section 3.2.6 and span surface energy balance methods, vegetation index–based parameterizations, and crop coefficient formulations.

OpenET models Daily ET estimates were obtained from six algorithms available through the OpenET platform (Melton et al., 2022): DisALEXI, eeMETRIC, geeSEBAL, PT–JPL, SIMS, and SSEBop. Five models use surface energy balance formulations, whereas SIMS applies a crop coefficient approach derived from vegetation density. All six models are driven primarily by Landsat Collection 2 imagery and gridded meteorological datasets (Spatial CIMIS in California and gridMET elsewhere), with bias-corrected forcing from a network of agricultural weather stations.

For each day in 2019–2024, model outputs were extracted within the dynamic 80 % EC footprint boundaries described in Section 3.2.3. Retrievals were obtained using the OpenET API, ensuring consistent spatial support between remote-sensing estimates and in situ measurements.

SEB-PV model and custom diurnal temporal upscaling To increase structural diversity in the ensemble, the Surface Energy Balance for Partially Vegetated surfaces (SEB–PV) model (Lagos et al., 2012) was included. Following Cigarra-Guíñez et al. (2025), SEB–PV was implemented with Landsat Collection 2 imagery, global soil property datasets, surface soil moisture products, and automated weather station data. The algorithm computes instantaneous ET at satellite overpass time using a four-layer energy balance representation that partitions energy exchange between soil and vegetation.

A total of 110 Landsat scenes (path/row 042/035) were processed for 2019–2024. Scenes with more than 1 % cloud cover within a 200 m buffer around each footprint were

excluded, and only dates with corresponding EC observations were retained. Because SEB–PV produces instantaneous ET, a temporal upscaling procedure was used to obtain daily values. We evaluated published approaches and implemented a hybrid method combining (i) ratios of daily to instantaneous solar irradiance (Jackson et al., 1983), (ii) ratios of grass-reference ET (Colaizzi et al., 2006; ASCE-EWRI, 2005), and (iii) air temperature scalars (Samani et al., 2007). The final daily ET estimate was obtained as a weighted combination of these ratios; full details are provided in Supplementary Material S4.

3.2.5 Simple average ensemble benchmark

A simple arithmetic mean of the seven ET models was used as the primary non-Bayesian benchmark. The OpenET platform provides an ensemble product that applies Median Absolute Deviation (MAD) filtering to identify and exclude outlier predictions before averaging the remaining base models. A preliminary comparison revealed that a simple arithmetic mean of all six OpenET base models—without MAD filtering—outperformed the official OpenET ensemble product (KGE: 0.72 vs. 0.69; NRMSE: 28.2% vs. 30.3%; Supplementary Material S5). This indicates that MAD filtering discarded predictions that contributed useful information in citrus orchards. Given its superior performance, transparency, and reproducibility, the arithmetic mean of all seven candidate models (six OpenET algorithms plus SEB-PV) was adopted as the benchmark.

3.2.6 Probabilistic ensemble framework

To quantify uncertainty and assess how predictive performance depends on calibration structure and error specification, a Bayesian Model Averaging framework was implemented. A factorial design allowed each calibration architecture to be evaluated under two likelihood assumptions: conditional independence and spatiotemporal dependence.

Model architectures: a systematic progression of complexity

Three classes of calibration architecture were considered, moving from global adjustments to models that adapt to local and meteorological variation.

Global calibration These models apply a single correction across all sites and dates. Four variants were tested: (i) *weights-only* ($\mu = \sum w_k x_k$), (ii) *intercept* ($\mu = \alpha + \sum w_k x_k$), (iii) *scale* ($\mu = \beta \sum w_k x_k$), and (iv) *full* ($\mu = \alpha + \beta \sum w_k x_k$), where μ is the expected ET (mm d^{-1}), x_k are individual model predictions (mm d^{-1}), w_k are weights (dimensionless), α is an intercept (mm d^{-1}), and β is a scale factor (dimensionless).

Hierarchical calibration Parameters vary by site using partial pooling, capturing systematic site-level differences while sharing information across locations when data are limited (Gelman and Hill, 2006).

State-dependent calibration Calibration parameters are expressed as functions of daily meteorological variables (temperature, radiation, reference ET), allowing the ensemble to respond to varying atmospheric conditions.

Full model specifications for all architectures are provided in Supplementary Material S6.

Error structure formulation

Each calibration architecture was fitted under two error structures.

Conditional independence (CI) Residuals are treated as independent after conditioning on predictors. This serves as the benchmark error formulation.

Spatiotemporal dependence A multivariate Student's t likelihood jointly models residuals across sites and days, with three components: (i) spatial correlation between

sites on the same day, (ii) AR(1) temporal persistence within each site, and (iii) observation-level variance. Student's t distributions accommodate occasional extreme deviations (Kruschke, 2015). LKJ priors for spatial correlation and AR(1) parameters are specified in Supplementary Material S6.

3.2.7 Prior specifications and model implementation

Weakly informative priors constrained parameters to physically plausible ranges while allowing the data to dominate inference (Gelman et al., 2020). For hierarchical models, more regularizing priors were defined using posterior summaries from simpler architectures to stabilize sampling. All models were implemented in Stan using Hamiltonian Monte Carlo, a Markov Chain Monte Carlo (MCMC) algorithm (Hoffman and Gelman, 2014; Stan Development Team, 2025). Prior specifications and hyperparameters are documented in Supplementary Material S7. Prior predictive checks are provided in Data Availability.

3.2.8 Model diagnostics and performance evaluation

Convergence diagnostics

Convergence was assessed using rank-normalized split-$\widehat{R}$ (all values < 1.01), effective sample sizes exceeding 100 per chain, and inspection of divergent transitions after sampler tuning (Vehtari et al., 2021).

Model evaluation

Model performance was evaluated using two complementary approaches. Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO-CV; Vehtari et al., 2017) was applied as a diagnostic tool to identify influential observations and assess model misspecification. For diagnostic purposes, models were fitted with eight MCMC chains. Observations with Pareto shape parameters $\hat{k} > 0.7$ indicate unreliable

approximations and may signal influential data points requiring further investigation.

To address the research questions on operational forecasting performance, forward-chaining temporal cross-validation was implemented (Bergmeir et al., 2018). This design evaluates whether models trained on historical data can reliably predict future years with potentially different climatic conditions. The 2019–2024 dataset was partitioned into four sequential folds. Models were trained on 2019–2020 (66 observations) and tested on 2021 (32 observations), then iteratively expanded to predict 2022, 2023, and 2024. Training sets contained 66, 98, 162, and 232 observations; test sets contained 32, 64, 70, and 67 observations, respectively. Due to overlapping training windows, fold-level results are not independent. For computational efficiency in the repeated model fitting across folds, models were fitted with four MCMC chains.

Bayesian model comparison

Predictive accuracy was evaluated using expected log predictive density (ELPD), which quantifies out-of-sample predictive performance while accounting for uncertainty (Vehtari et al., 2017). Models were ranked based on paired ELPD differences across folds, with standard errors computed using paired estimates (Supplementary Material S8).

Supplementary performance metrics

Posterior predictive distributions for each test fold were used to compute MAE, MBE, NRMSE, R^2, and Kling–Gupta Efficiency (KGE). The Model Decision Making Indicator (MDMI),

$$\text{MDMI} = 100 \times \text{KGE} - \text{NRMSE},$$

summarized operational skill by combining efficiency and normalized error. Benchmark (non-Bayesian) models were evaluated using the same folds to ensure comparability.

Interpretive model similarity analysis

To contextualize the posterior BMA weights, an interpretive analysis of base model relationships was conducted using distance correlation and energy distance. This analysis is descriptive only; BMA weights are determined solely through likelihood-based inference. Results are presented in Appendix A, with full protocols in Supplementary Material S8.

3.3 Results

This section first establishes baseline performance of individual models and a non-Bayesian benchmark, then evaluates the effect of error structure specification, compares calibration architectures within the spatiotemporal framework, and characterizes the optimal model.

3.3.1 Baseline performance of individual models and simple average

Baseline evaluation over the full 2019–2024 dataset ($n = 300$) revealed substantial differences among the seven candidate ET models (Table 3.2). SIMS and SSEBop achieved the highest Kling–Gupta Efficiency (KGE = 0.74) with Normalized RMSE (NRMSE) below 29%. DisALEXI and geeSEBAL showed intermediate performance (KGE $\approx$ 0.63–0.64; NRMSE 30–36%), while PT–JPL, eeMETRIC, and SEB-PV performed worse (KGE 0.45–0.58; NRMSE 31–39%). SEB-PV combined relatively low KGE (0.45) with negligible mean bias (MBE = 0.0 mm d^{-1}).

The simple arithmetic average of all seven models, used as the primary non-Bayesian benchmark, obtained KGE = 0.70 and the lowest NRMSE (26.1%), with MBE = 0.6 mm d^{-1} (Table 3.2). Averaging reduced random error but did not eliminate systematic overestimation.

Diversity analysis based on distance correlation and energy distance (Figure 3.9) revealed that candidate models possess complementary error structures. Distance correlation identified one high-similarity cluster (PT–JPL, SIMS, SSEBop; distance

correlation > 0.70) and more distinct models including SEB-PV and eeMETRIC, indicating that optimal weighting can leverage these structural differences. Details appear in Appendix A and Supplementary Material S9.

Table 3.2: Descriptive performance metrics for individual ET models and the simple average ensemble across the full 2019–2024 dataset.

Model	KGE	NRMSE (%)	RMSE	R^2	MDMI	MAE	MBE
Individual Models							
SSEBop	0.74	28.6	0.9	0.70	45.3	0.7	0.6
SIMS	0.74	28.6	0.9	0.67	45.3	0.7	0.6
DisALEXI	0.63	30.0	1.0	0.48	33.0	0.8	0.5
geeSEBAL	0.64	35.6	1.2	0.56	28.8	0.9	0.8
PT-JPL	0.58	34.7	1.1	0.68	22.9	1.0	0.9
eeMETRIC	0.55	39.4	1.3	0.43	15.1	0.9	0.7
SEB-PV	0.45	31.8	1.0	0.26	13.2	0.8	0.0
Ensemble Benchmark							
Simple Average	0.70	26.1	0.9	0.73	43.8	0.7	0.6

KGE = Kling-Gupta Efficiency; NRMSE = Normalized Root Mean Square Error; RMSE = Root Mean Square Error (mm d^{-1}); R^2 = Coefficient of Determination; MDMI = Model Decision Making Indicator; MAE = Mean Absolute Error (mm d^{-1}); MBE = Mean Bias Error (mm d^{-1}).

3.3.2 Effect of error structure specification

All Bayesian configurations satisfied convergence criteria, with rank-normalized split-$\widehat{R} < 1.01$, adequate effective sample sizes, and limited influential observations (Supplementary Material S10). The hierarchical scale model exhibited the highest number of highly influential observations (n = 19 with Pareto $\hat{k} > 0.7$), reflecting parameter non-identifiability between scale factors and weights. With limited spatial replication (n = 3 sites), hierarchical models offer minimal advantage over simpler alternatives, consistent with theoretical expectations for multilevel models with few groups (Gelman and Hill, 2006). Details on convergence diagnostics and identifiability issues appear in Supplementary Material S10.

Comparison between multivariate t (MVT) and conditional independence (CI) likelihoods showed consistent gains in expected log predictive density (ELPD) for MVT

across all calibration architectures (Table 3.3). Differences in ELPD (MVT minus CI) were positive in all cases, ranging from 37 (hierarchical intercept) to 128 units (weights-only). Even the simplest MVT configuration (weights-only, ELPD = -179) outperformed the best CI model (hierarchical full, ELPD = -184). This result directly addresses the first research question: explicit modeling of spatiotemporal dependence improves predictive performance more than changes in calibration architecture.

Table 3.3: Comparison of Multivariate-t (MVT) vs Conditional Independence (CI) model structures using ELPD from temporal cross-validation.

Model architecture	**MVT ELPD**[a]	**CI ELPD**[a]	**$\triangle$ ELPD**[b]	**SE($\triangle$)**[c]	**$\lvert\triangle\rvert$/SE**[d]
Global Calibration					
Weights only	-179.2	-306.8	127.6	19.8	6.5
Intercept	-150.8	-215.7	64.9	11.6	5.6
Scale	-152.5	-229.2	76.7	11.5	6.7
Full	-150.7	-216.4	65.6	10.8	6.1
State-dependent					
Intercept	-134.5	-195.2	60.7	11.3	5.4
Intercept and weights	-137.9	-198.2	60.2	12.1	5.0
Hierarchical					
Intercept	-147.4	-184.6	37.2	6.5	5.8
Scale	-151.6	-195.9	44.3	7.0	6.4
Full	-139.5	-184.0	44.5	7.2	6.2

[a] Expected log predictive density from 4-fold temporal cross-validation (higher values indicate better predictive performance).
[b] Difference in ELPD (MVT $-$ CI); positive values favor MVT.
[c] Standard error of the ELPD difference.
[d] Ratio of absolute difference to standard error.

Residual variance decomposition under the MVT framework (Figure 3.3) indicates that spatial and temporal components dominate total residual variance, accounting collectively for 95–98% across all architectures. Observational variance remained between 1 and 3%, while temporal and spatial components accounted for 56–81% and 18–53%, respectively. Global models concentrated variance in the temporal component, whereas hierarchical

models shifted a larger fraction into the spatial term as site-specific parameters absorbed cross-site heterogeneity.

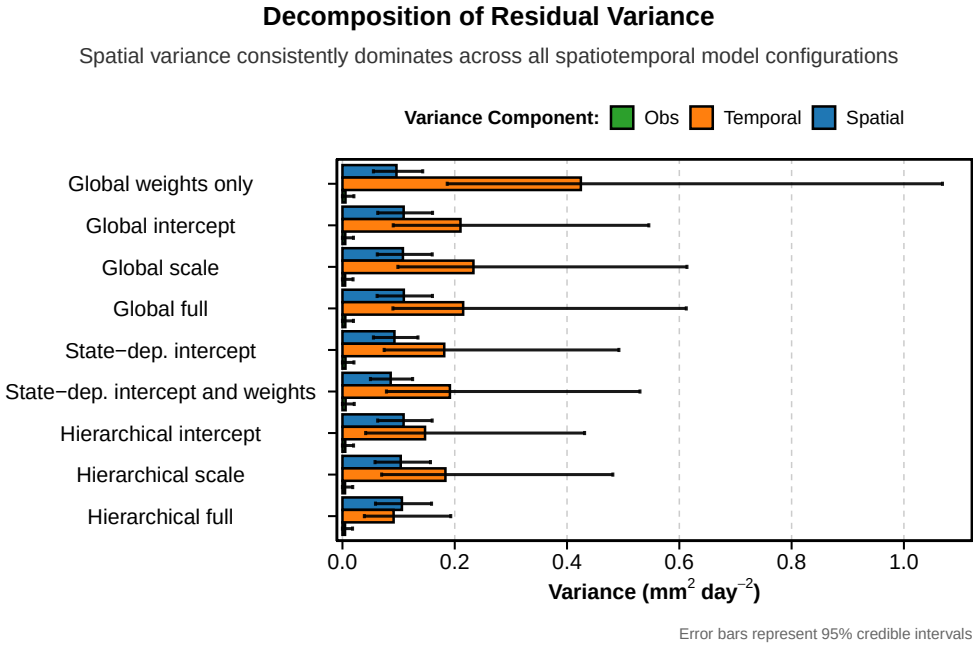

Figure 3.3: Posterior mean decomposition of total residual variance into observational, temporal, and spatial components for each MVT calibration architecture.

3.3.3 Calibration architectures under spatiotemporal dependence

Within the MVT framework, calibration architectures differed in predictive performance under forward-chaining temporal cross-validation (Table 3.4, Figure 3.4). The state-dependent intercept model achieved the highest ELPD and was used as reference. The state-dependent intercept and weights model and the hierarchical full model showed similar performance, with ΔELPD of -3.4 and -5.0, respectively.

Global and hierarchical models without state dependence had lower ELPD, with ΔELPD between -12.9 and -18.0 for most cases, and -44.7 for the weights-only model. All models lost predictive skill when extrapolating to the 2023 test period (average annual precipitation

500 mm vs. 260 mm in 2019–2022), but losses were smallest for the state-dependent intercept architecture, indicating more stable behavior under conditions that differed from the training period.

Comparison with CI-based rankings (Table 3.4) shows that ignoring spatiotemporal dependence inflates the apparent advantage of hierarchical models. Under CI, hierarchical full appears markedly superior to global full (ΔELPD = 32.4); under MVT, this advantage reduces to 11.2 ΔELPD points. This occurs because the spatial correlation matrix in MVT explicitly models dependencies that hierarchical parameters were implicitly absorbing under CI assumptions. Model selection frameworks that ignore spatiotemporal dependencies will systematically overestimate the benefits of hierarchical calibrations.

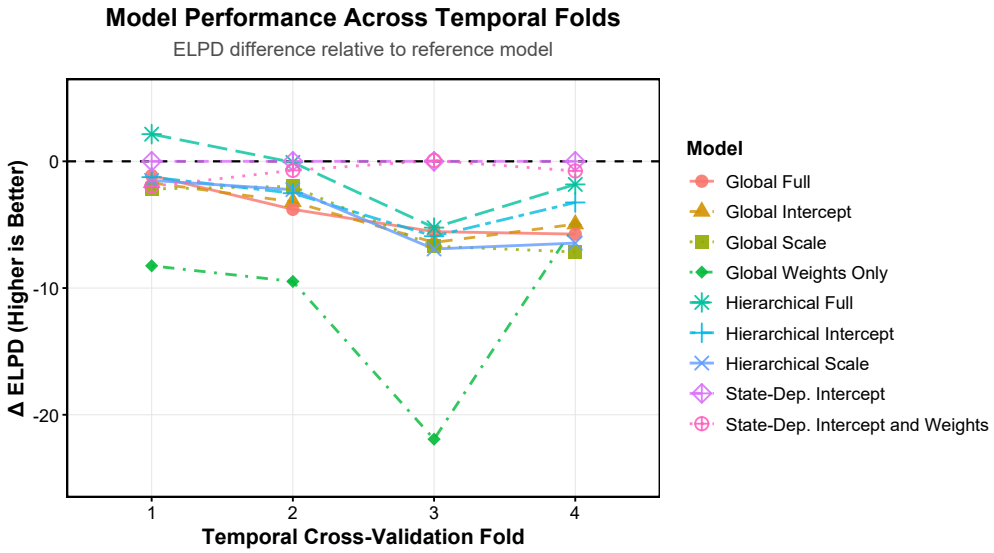

Figure 3.4: ΔELPD relative to the state-dependent intercept model for each of the four temporal cross-validation folds.

3.3.4 Characteristics and predictive skill of the optimal model

The state-dependent intercept model combines MVT error structure with an intercept that varies with daily meteorological conditions. Posterior weights for the seven base models (Figure 3.5a) show PT–JPL as the main contributor (48% mean weight), followed by SIMS

Table 3.4: Out-of-sample predictive performance using forward-chaining temporal cross-validation (4 folds, 2019–2024). Models are ranked by ELPD within the MVT framework.

Model	**ELPD**	**$\triangle$ELPD**	**SE($\triangle$ELPD)**
Spatiotemporal (MVT) Framework			
State-dependent intercept[a]	-134.5 ± 9.7	0.0	—
State-dependent weights and intercept	-137.9 ± 10.0	-3.4	1.7
Hierarchical full	-139.5 ± 9.5	-5.0	6.2
Hierarchical intercept	-147.4 ± 9.7	-12.9	3.9
Full	-150.7 ± 9.4	-16.2	4.3
Intercept	-150.8 ± 9.6	-16.3	4.1
Hierarchical scale	-151.6 ± 9.4	-17.1	5.6
Scale	-152.5 ± 9.4	-18.0	5.6
Weights only	-179.2 ± 12.0	-44.7	14.8
Conditional Independence (CI) Framework			
Hierarchical full	-184.0 ± 12.5	-49.5	11.3
Hierarchical intercept	-184.6 ± 13.0	-50.1	8.8
State-dependent intercept	-195.2 ± 10.8	-60.7	11.3
Hierarchical scale	-195.9 ± 13.0	-61.4	8.4
State-dependent weights and intercept	-198.2 ± 10.8	-63.6	12.0
Intercept	-215.7 ± 11.2	-81.2	12.1
Full	-216.4 ± 11.3	-81.9	12.5
Scale	-229.2 ± 11.2	-94.7	11.6
Weights only	-306.8 ± 17.6	-172.3	25.5

ELPD = Expected Log Predictive Density; SE = standard error of ELPD difference. The state-dependent intercept model is the reference (best MVT model).

(19%), with the remaining models receiving smaller but non-zero weights. This weighting differs from the baseline KGE ranking and reflects both individual skill and cross-model dependence.

The global intercept ($\alpha_0 = -0.48$ mm d^{-1}; Figure 3.5b) indicates systematic negative bias correction under average conditions. State-dependent effects (Figure 3.5c) show that minimum temperature produces the largest positive adjustment (0.44 mm d^{-1} per standard deviation), while reference ET and maximum temperature introduce negative corrections (-0.20 and -0.11 mm d^{-1}); wind speed and solar radiation have smaller effects. The resulting dynamic intercept (Figure 3.5d) varies across representative environmental scenarios, with stronger corrections under cold, humid conditions and weaker corrections during warm, dry periods.

Posterior predictive checks indicate that uncertainty depends on satellite revisit frequency and climatic variability. Prediction intervals are wider before the launch of Landsat 9, when gaps between overpasses were longer, and narrower under dual-satellite operation (Figure 3.14). Residual time series (Figure 3.6) show that the state-dependent intercept model reduces residual variance and bias relative to the simple average across sites and years, with residuals closer to zero and less temporal persistence.

Across temporal folds, the state-dependent intercept model reaches KGE = 0.85 and NRMSE = 15%, while the simple average remains at KGE = 0.65 and NRMSE = 28% (Figure 3.7). Individual satellite products show KGE between 0.36 and 0.69 and NRMSE between 31 and 42%. Mean bias error decreases from 0.68 mm d^{-1} for the simple average to 0.15 mm d^{-1} for the state-dependent intercept model, with higher Model Decision Making Indicator values (Figure 3.8).

A sensitivity analysis using SEB-PV configured with in situ soil moisture and net radiation for temporal upscaling (Appendix C, Figure 3.11) shows that SEB-PV can reach performance similar to the best individual models. However, its low ensemble weight and the limited change in fold-averaged BMA metrics indicate that improving a single

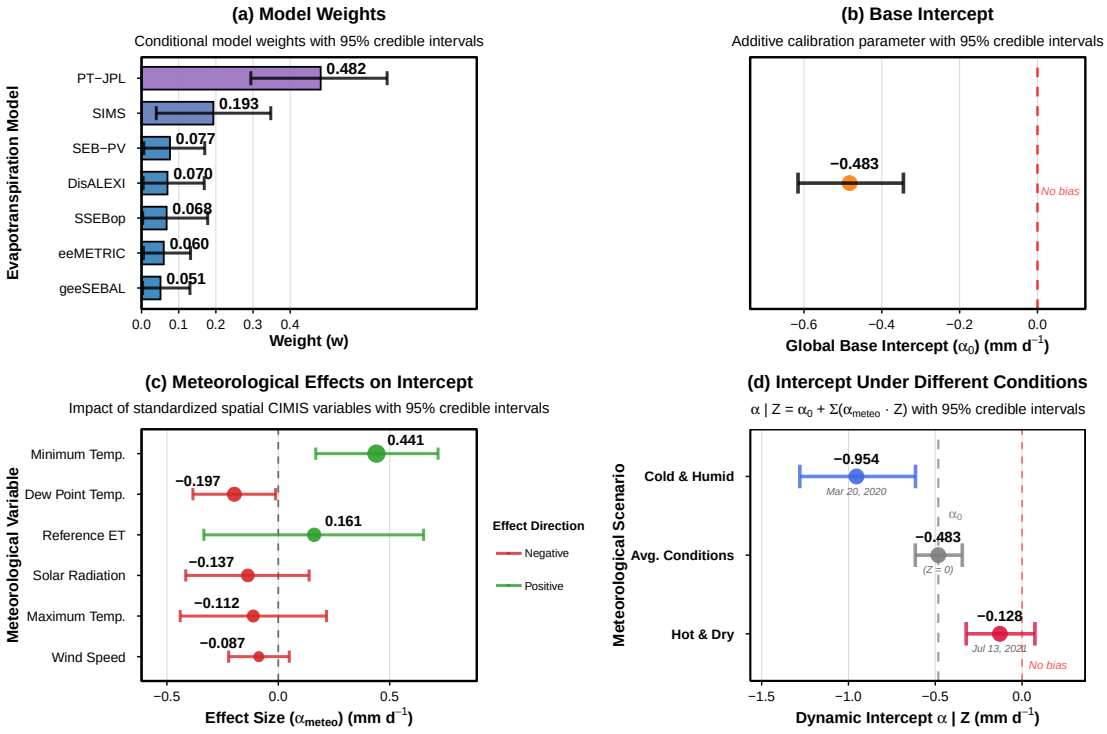

Figure 3.5: Posterior estimates for the state-dependent intercept model. (a) Weights for base ET models. (b) Global intercept. (c) Effects of standardized meteorological covariates on the intercept. (d) Dynamic intercept under representative environmental scenarios.

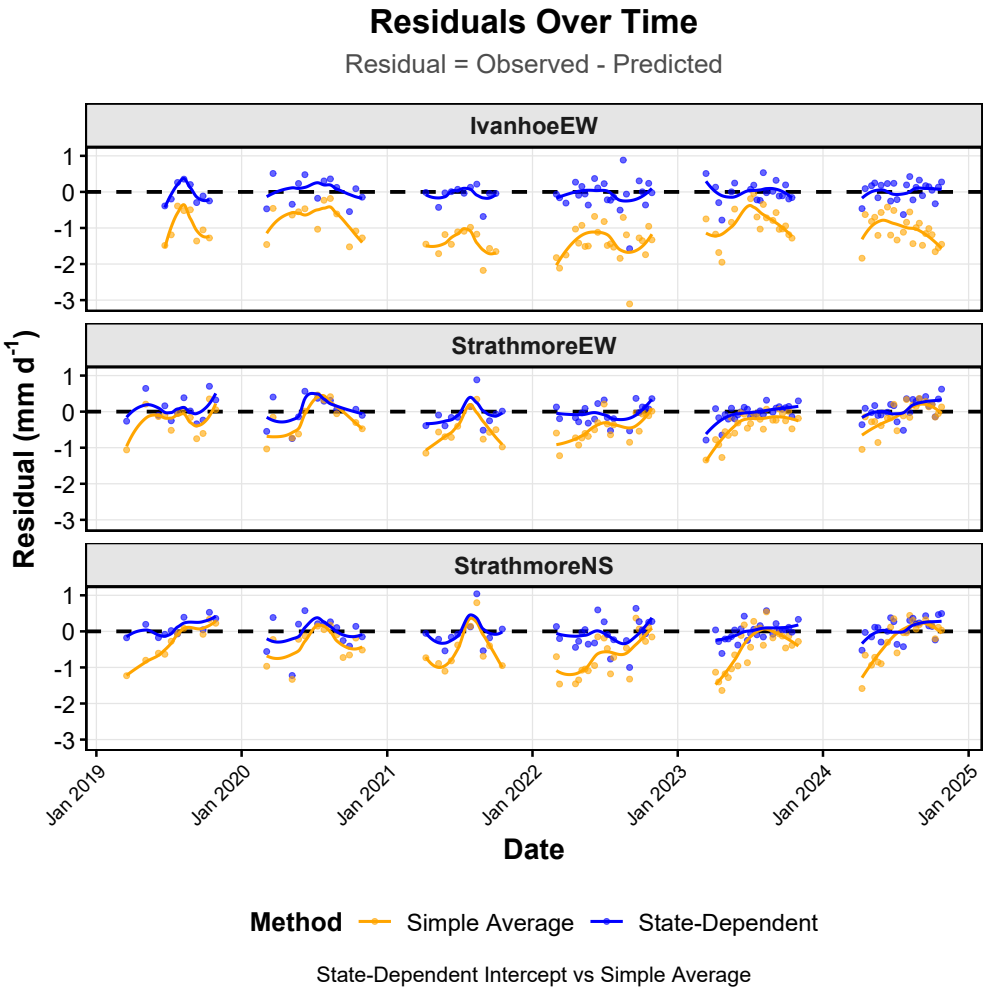

Figure 3.6: Temporal evolution of prediction residuals for the state-dependent intercept model (blue) and the simple average ensemble (orange) over 2019–2024.

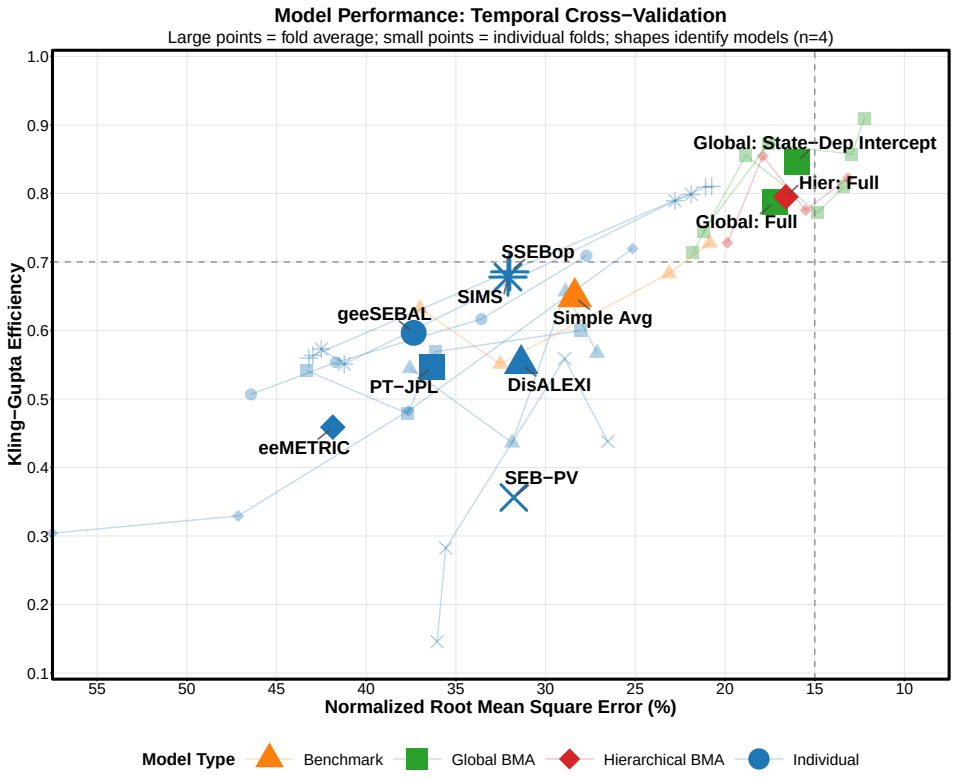

Figure 3.7: Kling–Gupta Efficiency versus Normalized RMSE averaged across four temporal folds (large points) with individual fold realizations (small points).

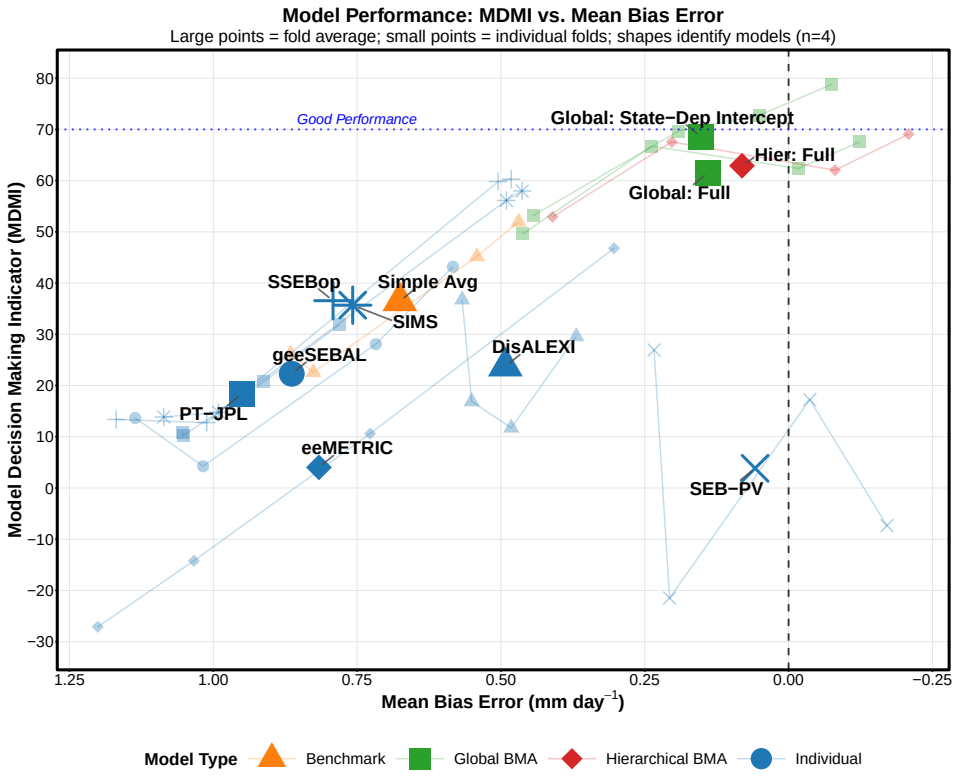

Figure 3.8: Model Decision Making Indicator versus Mean Bias Error averaged across four temporal folds (large points) with individual fold realizations (small points).

component model produces modest changes at the ensemble level. NRMSE for the BMA ensembles remains close to 15% with KGE between 0.8 and 0.9, while required bias correction decreases but does not vanish. These results indicate that gains arise from the interaction between spatiotemporal Bayesian calibration and the diversity and quality of the underlying satellite-based ET products.

3.4 Discussion

The primary determinant of ensemble performance is explicit modeling of spatiotemporal error structure. Across all calibration architectures, models that represent spatial and temporal dependence consistently achieved higher predictive accuracy than those assuming conditional independence. This finding challenges weighting approaches that do not account for residual dependence. Simply moving from conditional independence to multivariate spatiotemporal specification—while maintaining the same weights-only architecture—yielded substantial predictive gains (Table 3.3), demonstrating that weighting schemes alone cannot address the dominant sources of error when spatiotemporal correlations remain unmodeled.

This contrasts with performance-based weighting strategies (Reitz et al., 2025), where spatially optimized weights reduce errors but operate under independence assumptions. Our results show that in citrus systems, residual dependence persists even after optimal weighting, and that explicit calibration is required to eliminate systematic bias. Within the spatiotemporal framework, the state-dependent intercept architecture substantially outperformed weights-only models, indicating that weighting redistributes bias but does not remove it.

The comparison with global-scale BMA-ET implementations (Wu et al., 2025) further clarifies the implications of independence assumptions. Their framework applies canonical Bayesian Model Averaging under conditional independence assumptions and uses spatial

holdout validation for evaluation. This strategy effectively assesses structural differences across global ET products but treats temporally concatenated observations as independent. Our forward-chaining temporal validation demonstrates that temporal non-stationarity substantially alters predictive performance under independence assumptions. The state-dependent intercept architecture improved performance under unseen weather conditions (2023 wet year) by adjusting systematic bias as meteorological drivers shifted, confirming that temporal generalization benefits from explicit modeling of both dependence and dynamic bias.

The study also clarifies how ensemble performance relates to characteristics of underlying ET products. Previous work documented seasonal biases in OpenET models for citrus systems, particularly during spring (Dhungel et al., 2024). Our results confirm this and show that calibration architectures assuming conditional independence can amplify these biases 3.6. Similarity analysis revealed that several OpenET models share correlated error structures, limiting ensemble diversity. This aligns with findings that structural similarity among ET algorithms constrains deterministic multi-model aggregation (Petrakis et al., 2024). The inclusion of SEB-PV provided partial improvement due to its distinct error structure, confirming that ensemble gains depend on combining models with complementary error characteristics.

Several methodological limitations indicate directions for further development. The small spatial network (n = 3 sites) restricts identifiability of hierarchical parameters and limits assessment of how calibration complexity scales with spatial heterogeneity. The unstructured spatial correlation matrix captures dependence within the network but cannot generalize to unmonitored locations. Extending the framework with structured spatial priors or Gaussian Process covariance models (Stein, 1999; Banerjee et al., 2014) would enable regional-scale prediction. Computational cost increases with multivariate calibration and may pose challenges for real-time applications requiring rapid updates. Finally, ensemble gains remain bounded by similarity among current ET models; increasing diversity by

incorporating algorithms with distinct physical assumptions or additional sensing inputs may yield further improvements.

The findings demonstrate that reliable ET ensemble prediction depends on modeling spatiotemporal dependence and incorporating adaptive bias correction. Weighting alone is insufficient when systematic biases and residual correlations remain unaddressed. The Bayesian framework developed here provides predictive distributions and uncertainty estimates that support irrigation scheduling, allocation planning, and risk assessment. Future work should expand spatial coverage, introduce structured spatial covariance for regional generalization, increase ensemble diversity, and integrate near-real-time inputs for operational forecasting.

3.5 Conclusions

This study demonstrates that ensemble evapotranspiration prediction in agricultural systems depends primarily on explicit modeling of spatiotemporal error structure rather than on calibration complexity. Through systematic comparison of parallel model suites under conditional independence and multivariate spatiotemporal assumptions, we addressed two fundamental questions using six years of residual energy balance data from three commercial citrus orchards.

First, explicit modeling of spatiotemporal dependence is substantially more important than refining calibration architecture. The multivariate Student's t framework outperformed conditional independence counterparts across all calibration variants, with ELPD differences ranging from 37 to 128 units. Variance decomposition reveals that spatiotemporal structure accounts for 95–98% of residual variance, confirming that weighting or calibration schemes ignoring residual dependence cannot address the dominant sources of error. Second, state-dependent calibration provides robust temporal generalization. The state-dependent intercept model, which allows bias correction to

vary with meteorological conditions, achieved the best out-of-sample performance and demonstrated stable extrapolation to unseen weather conditions where static calibration approaches exhibited substantial degradation.

These findings have implications for operational ensemble design. Methods relying solely on deterministic weighting or simple averaging ignore spatiotemporal dependencies and cannot quantify prediction uncertainty. The Bayesian framework developed here generates full posterior predictive distributions that support risk-informed water management decisions. While our implementation used a small spatial network (n = 3 sites), limiting assessment of hierarchical calibration benefits, the core finding that spatiotemporal error structure dominates calibration complexity applies generally to agricultural monitoring systems.

Future work should expand spatial coverage to enable structured spatial priors for regional generalization, increase ensemble diversity through models with distinct physical assumptions, and integrate near-real-time inputs for operational forecasting.

Data availability

The Stan source codes for the BMA models, along with the Prior Predictive Checks and the Supplementary Material, are openly available at the Zenodo repository: `https://doi.org/10.5281/zenodo.17654893`. The datasets used in this study are available from the corresponding author upon reasonable request.

Acknowledgments

We express our gratitude to Samuel G. Metcalf for his valuable assistance with flux tower data collection and field measurements.

A Appendix. Base model diversity analysis

We evaluated model diversity using distance correlation (dCor) between prediction series and energy distance (ED) between error distributions. dCor captures both linear and non-linear dependence; ED measures differences between error distributions.

Figure 3.9 displays models in dCor–ED space. SIMS and PT-JPL form a pair with high prediction similarity and low error dissimilarity. geeSEBAL, eeMETRIC, DisALEXI, SSEBop and SIMS have low ED ($<$0.11), which indicates similar error patterns. SEB-PV has higher ED and lies farther from the cluster. Hierarchical clustering based on these metrics is presented in Supplementary Material S9.

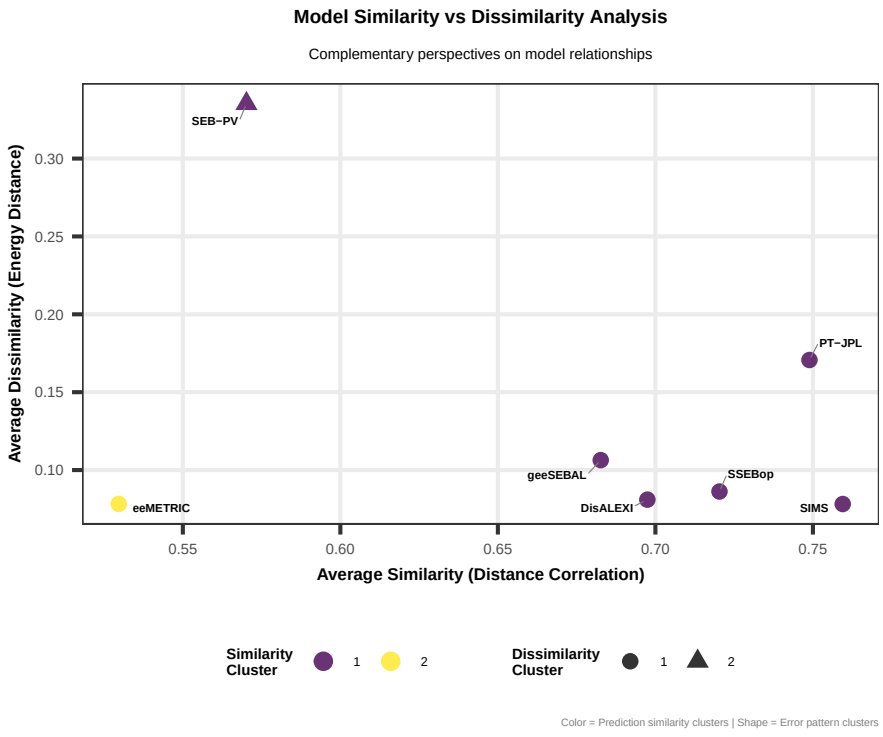

Figure 3.9: Distance correlation vs energy distance for base ET models. Colors and shapes indicate clusters based on similarity and dissimilarity.

B Appendix. BMA weights only posterior

Figure 3.10 shows posterior means and credible intervals for base model weights under the simplest MVT model (weights-only). PT-JPL, SIMS and SSEBop receive higher weights, consistent with their baseline performance and error similarity.

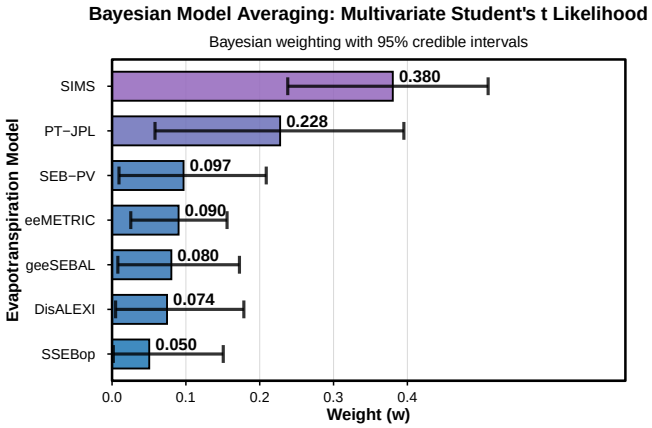

Figure 3.10: Posterior weights for the simplest multivariate weights-only model.

C Appendix. Sensitivity analysis with field-measured SEB-PV inputs

Table 3.5 reports ELPD for all models when SEB-PV uses field-measured radiation and soil moisture, while other models remain unchanged. The best MVT state-dependent intercept model improves from ELPD -134.5 to -129.1. Relative rankings remain stable.

Table 3.5: Temporal cross-validation results with SEB-PV driven by field-measured radiation and soil moisture.

Model	**ELPD**	**$\triangle$ELPD**	**SE($\triangle$ELPD)**
Spatiotemporal (MVT) Framework			
State-dependent intercept[a]	-129.1 ± 10.1	0.0	—
State-dependent weights and intercept	-131.2 ± 9.9	-2.1	1.9
Hierarchical full	-139.2 ± 10.1	-10.1	14.6
Hierarchical intercept	-145.9 ± 9.6	-16.8	7.8
Hierarchical scale	-148.0 ± 10.0	-18.9	11.8
Intercept	-148.9 ± 9.4	-19.8	7.3
Full	-150.2 ± 9.6	-21.1	7.4
Scale	-151.3 ± 9.7	-22.2	7.8
Weights only	-156.7 ± 9.8	-27.6	7.8
Conditional Independence (CI) Framework			
Hierarchical full	-179.6 ± 13.4	-50.5	21.8
Hierarchical intercept	-181.2 ± 13.8	-52.1	18.0
Hierarchical scale	-186.9 ± 14.2	-57.8	22.6
State-dependent weights and intercept	-193.5 ± 10.3	-64.4	13.9
State-dependent intercept	-195.9 ± 10.6	-66.8	15.2
Intercept	-216.2 ± 10.8	-87.1	17.1
Full	-217.1 ± 11.0	-88.0	18.2
Scale	-227.9 ± 10.8	-98.8	18.1
Weights only	-245.4 ± 10.8	-116.3	25.6

Reference model. All models converged with $\widehat{R} < 1.01$ and Bulk/Tail-ESS > 400.

ELPD = Expected Log Predictive Density; SE = standard error.

Figure 3.11 shows the KGE–NRMSE space for this sensitivity run, and Figure 3.12 presents updated state-dependent parameters. Figure 3.13 shows ΔELPD by fold.

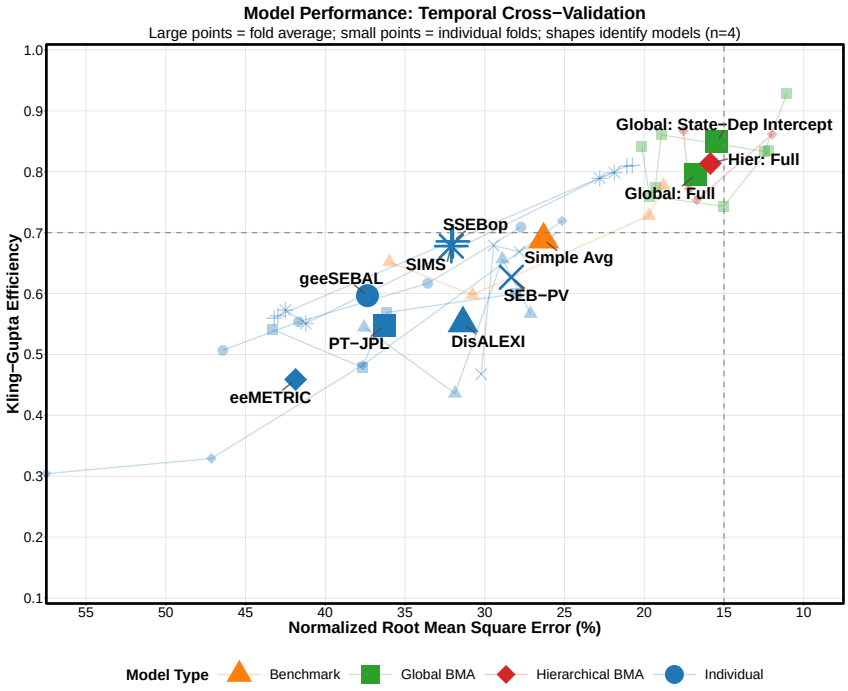

Figure 3.11: KGE vs NRMSE with SEB-PV driven by field measurements.

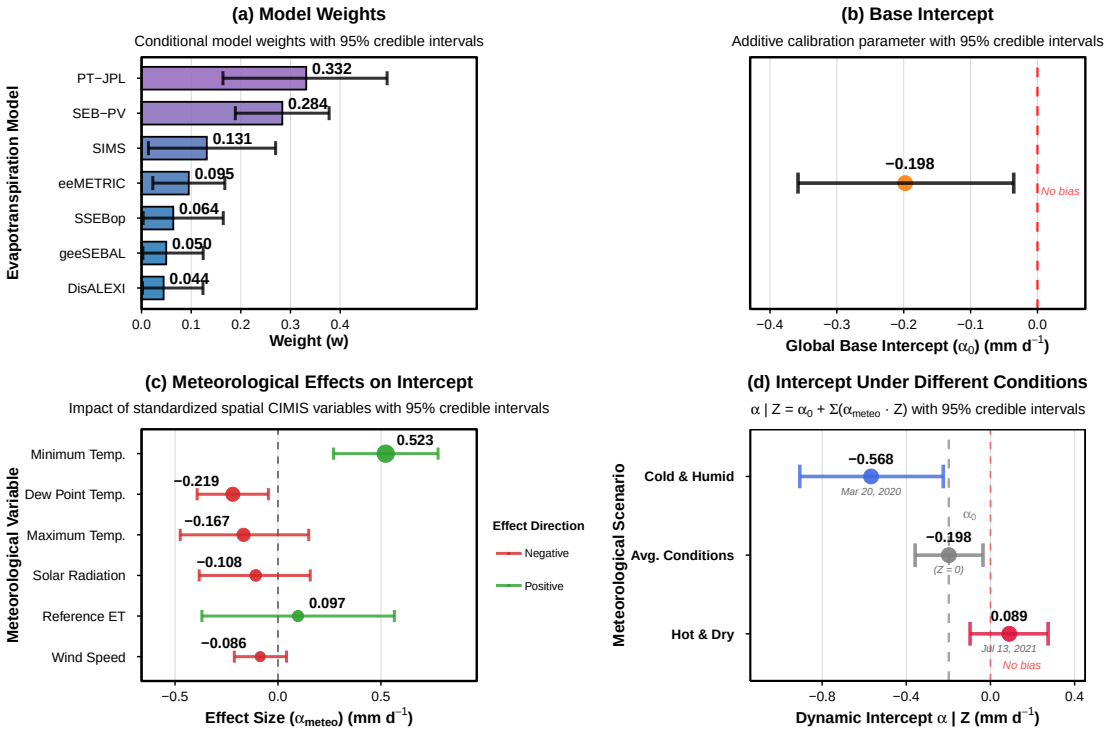

Figure 3.12: Posterior summaries for state-dependent intercept model with SEB-PV using field measurements.

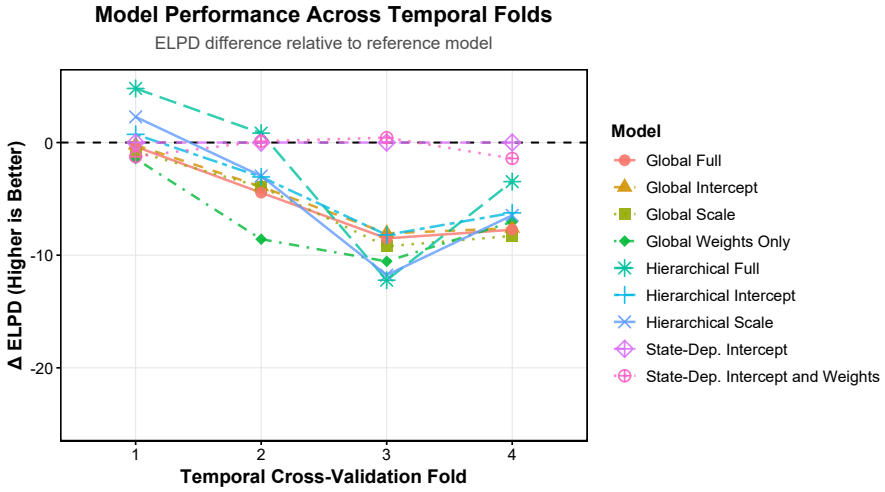

Figure 3.13: ΔELPD by fold with SEB-PV driven by field measurements.

D Appendix. Posterior predictive check

Figure 3.14 shows observed ET, median predictions and 50% and 95% credible intervals for the state-dependent intercept MVT model, along with the simple average ensemble for comparison.

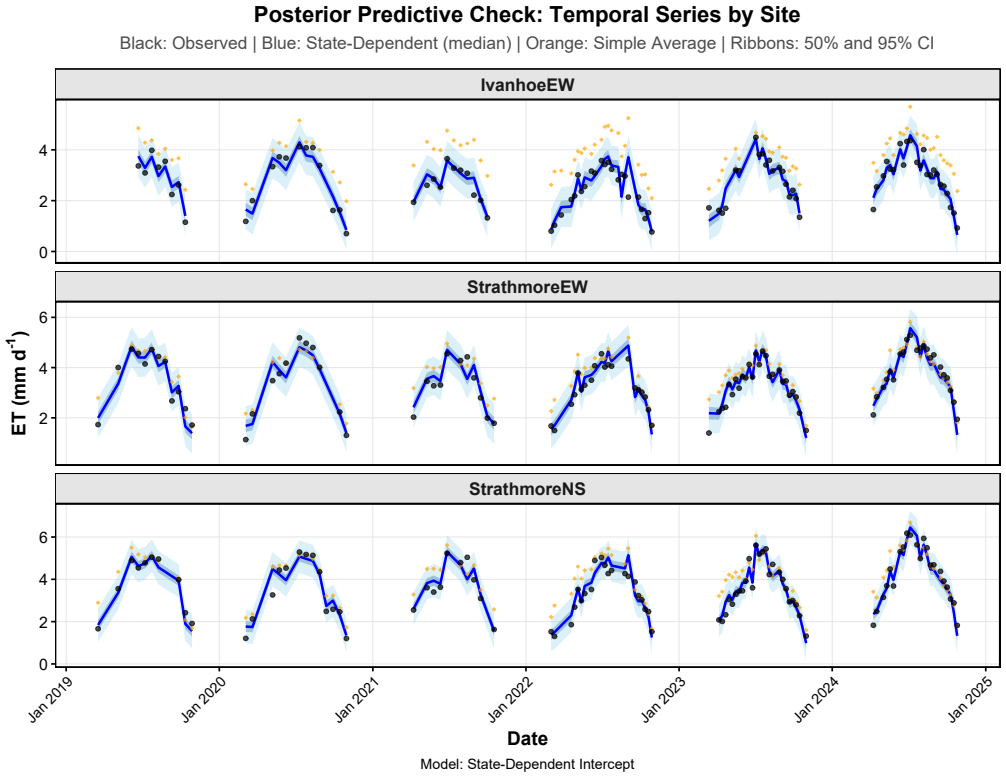

Figure 3.14: Observed ET (points), posterior predictive medians and intervals (blue) and simple average ensemble (orange) for three sites over 2019–2024.

CAPÍTULO 4

DISCUSIÓN GENERAL

Los resultados presentados en esta tesis establecen que para los sistemas estudiados, el diseño de monitoreo de evapotranspiración debe priorizar la especificación explícita de estructuras de error espaciotemporales sobre el refinamiento de arquitecturas de calibración. Esta conclusión emerge de un programa de investigación secuencial implementado en dos fases. Primero, la adaptación operacional del modelo SEB-PV mediante productos de humedad del suelo grillados y datos de estaciones meteorológicas, junto con la evaluación de algoritmos de escalamiento temporal bajo contextos climáticos contrastantes. Segundo, la integración de este modelo físico especializado en un marco de Promedio de Modelos Bayesiano junto con seis algoritmos satelitales de OpenET, evaluando arquitecturas de calibración bajo especificaciones de error alternativas mediante validación cruzada temporal rigurosa.

El Capítulo 2 demostró que SEB-PV mantiene precisión operacional (NRMSE 21–29%) utilizando exclusivamente datos globalmente disponibles, eliminando la barrera de instrumentación in situ. Este resultado depende de dos desarrollos metodológicos. El parámetro $beta$ de resistencia superficial del suelo funciona como factor de corrección efectivo cuando se emplean productos grillados, pero los ajustes requeridos varían según condiciones climáticas. Los sitios en clima mediterráneo demandaron ajustes menores que los sitios bajo condiciones semiáridas, indicando que la calibración compensa diferencias en soporte espacial entre productos grillados (9–25 km) y procesos a escala de huerto.

La evaluación de siete métodos de escalamiento temporal reveló dependencia clara del contexto climático y estructural. Métodos basados en radiación neta (ETd2med) alcanzaron RMSE de 0.8 mm d^{-1} en avellanos bajo clima mediterráneo, donde disponibilidad energética controla ET. Métodos basados en variables meteorológicas (ETd4) alcanzaron RMSE de 1.3 mm d^{-1} en pistachos bajo condiciones advectivas, donde déficit de presión de vapor domina. Esta dicotomía refleja diferencias en los controladores físicos dominantes, consistente con estudios previos que han documentado variabilidad en el desempeño de algoritmos de escalamiento según tipo de ecosistema y condiciones atmosféricas (Jiang et al., 2021; Cammalleri et al., 2014; Chávez et al., 2008; Colaizzi et al., 2006), e informa la selección de algoritmos en implementaciones operacionales para estos sistemas.

La integración de SEB-PV con seis modelos OpenET permitió abordar en el Capítulo 3 la pregunta fundamental sobre diseño de ensambles: si las ganancias predictivas emergen de calibraciones sofisticadas o de la modelación explícita de dependencias espaciotemporales en residuos. El diseño factorial evaluó arquitecturas de calibración (global, jerárquica y dependiente del estado) bajo dos especificaciones de verosimilitud: independencia condicional versus estructura espaciotemporal multivariada. Los marcos multivariados superaron a contrapartes de independencia condicional por 37–128 unidades de densidad predictiva logarítmica esperada (ELPD), equivalente a 5–7 errores estándar de la diferencia de ELPD, a través de todas las arquitecturas evaluadas. La descomposición de varianza residual confirmó que componentes espaciales y temporales explican colectivamente 95–98% de la varianza, mientras que el término observacional contribuye solo 1–2%. Estos resultados indican que, en los sistemas estudiados, invertir en arquitecturas de calibración más complejas sin antes especificar explícitamente la estructura de error espaciotemporal es metodológicamente infructuoso: la ganancia máxima obtenida por refinamiento de calibración dentro del marco de independencia condicional no alcanza la ganancia mínima obtenida por el modelo más simple bajo

estructura espaciotemporal.

La arquitectura de calibración óptima resultó ser el intercepto dependiente del estado dentro del marco espaciotemporal multivariado, logrando un NRMSE de 15% frente a 28% del promedio simple determinístico (reducción del 46%). Este modelo ajusta sesgo como función de variables meteorológicas diarias estandarizadas, permitiendo que correcciones varíen con temperatura mínima (efecto positivo: 0.44 mm d^{-1} por desviación estándar), ET de referencia y temperatura máxima (efectos negativos: -0.20 y -0.11 mm d^{-1}), mientras velocidad del viento y radiación solar ejercen influencias menores. La robustez de este enfoque se manifestó durante validación cruzada temporal, donde el modelo mantuvo desempeño estable bajo el año húmedo 2023 (precipitación anual 500 mm versus 260 mm promedio 2019–2022), mientras que calibraciones estáticas exhibieron degradación sustancial. Este resultado demuestra que sistemas que no adaptan correcciones de sesgo a condiciones meteorológicas no observadas durante calibración fallan bajo variabilidad climática.

Los hallazgos indican que la ponderación basada en desempeño histórico, si bien útil como punto de partida, es insuficiente como único mecanismo de agregación multi-modelo cuando existen dependencias espaciotemporales no modeladas (Reitz et al., 2025). El modelo de pesos únicamente bajo independencia condicional, análogo conceptual al promedio ponderado por desempeño de Reitz et al. (2025), alcanzó ELPD de -307, mientras que el mismo esquema de ponderación bajo estructura espaciotemporal alcanzó -179, una diferencia de 128 unidades ELPD que excede por mucho las ganancias por refinamiento de calibración (3–45 unidades ELPD).

La especificación de estructura de error espaciotemporal mediante distribuciones t multivariadas con procesos autorregresivos de primer orden y matrices de correlación espacial representa un avance metodológico sobre implementaciones BMA previas que asumen independencia condicional (Wu et al., 2025). El marco desarrollado genera distribuciones predictivas posteriores completas que cuantifican apropiadamente

incertidumbre, facilitando la gestión del agua bajo criterios de riesgo. La reducción del sesgo de 0.68 mm d^{-1} (promedio simple) a 0.15 mm d^{-1} (intercepto dependiente del estado) tiene consecuencias prácticas directas. Al integrar esta diferencia sobre una temporada de riego típica de 180 días, el error acumulado asciende a aproximadamente 95 mm (equivalente a 950 m^3 ha^{-1}), una una magnitud con implicaciones directas para la planificación de los recursos hídricos.

Los resultados tienen consecuencias directas para diseño de sistemas operacionales. Plataformas que reportan métricas de precisión sin intervalos de predicción proporcionan información incompleta para gestión bajo criterios de riesgo, donde la incertidumbre asociada a cada estimación es relevante para la toma de decisiones. En los sistemas estudiados, enfoques basados en radiación neta funcionaron mejor bajo clima mediterráneo donde disponibilidad energética controla ET, mientras que bajo condiciones advectivas donde déficit de presión de vapor domina, métodos basados en ET de referencia fueron superiores. El intercepto dependiente del estado mantuvo precisión bajo condiciones climáticas no observadas durante calibración (año húmedo 2023). Este resultado demuestra que sistemas operacionales deben incorporar covariables meteorológicas explícitamente en lugar de asumir parámetros constantes.

El análisis de diversidad estructural entre modelos candidatos, basado en distancia de correlación y distancia energética entre distribuciones de error, reveló que PT-JPL, SIMS y SSEBop forman un cluster de alta similitud en predicciones (distancia de correlación > 0.70), mientras que SEB-PV presenta la mayor disimilitud en distribuciones de error y eeMETRIC la menor similitud de predicciones respecto al resto. Esta diversidad estructural explica por qué el peso posterior de SEB-PV (11% en el modelo de pesos únicamente) es menor de lo que su disimilitud en error sugeriría, reflejando el balance entre diversidad y precisión individual en la asignación bayesiana de pesos. Incrementar la diversidad del ensamble mediante la incorporación de algoritmos con principios físicos fundamentalmente distintos representa una dirección promisoria para mejorar el desempeño agregado.

Los principios establecidos sobre dominancia de estructura de error sobre calibración, necesidad de corrección de sesgo adaptativa mediante calibración de intercepto dependiente del estado, y la dependencia contextual de algoritmos de escalamiento temporal, se derivan de sistemas específicos pero reflejan características generales de datos ambientales espaciotemporales. La prevalencia de correlación espaciotemporal en datos ambientales es ubicua (Legendre, 1993; F. Dormann et al., 2007), y la tendencia de modelos que comparten forzantes comunes a exhibir errores correlacionados no es específica de ET. El diseño factorial implementado constituye una plantilla aplicable a otros contextos de monitoreo satelital en agricultura. La calidad de los modelos miembros condiciona el desempeño del ensamble con rendimientos decrecientes: mejorar sustancialmente un modelo individual mediante inputs de campo produce ganancias modestas a nivel de ensamble dado que su contribución representa una fracción del conjunto. La incorporación de múltiples tipos de cobertura en una misma red es posible dado que el marco identifica cada sitio independientemente, lo que amplía el alcance potencial del diseño factorial a sistemas agrícolas más heterogéneos.

Limitaciones y trabajos futuros

Las limitaciones de este estudio se describen a continuación. La red de tres sitios, todos ubicados en el valle de San Joaquín, limita la representatividad espacial y restringe la evaluación de si los patrones encontrados se mantienen bajo condiciones distintas de clima, suelo o manejo. La matriz de correlación espacial estima la relación entre sitios observados simultáneamente, por lo que no puede inferirse la correlación hacia locaciones no monitoreadas. El período de estudio (2019–2024) incluye variabilidad interanual moderada; bajo condiciones sostenidamente fuera del rango histórico, la extrapolación de los efectos de las covariables meteorológicas podría introducir incertidumbre adicional. La disponibilidad de imágenes satelitales libres de nubosidad limita la frecuencia de

observaciones para calibración, afectando la estimación de la persistencia temporal de los errores y la amplitud de los intervalos de credibilidad. Adicionalmente, la subestimación del flujo de calor del suelo en sitios californianos afecta la partición del balance energético instantáneo en SEB-PV.

Los trabajos futuros deberían abordar estas limitaciones en seis direcciones. Primero, expandir la red espacial hacia otras regiones climáticas y tipos de cultivo, lo cual es posible dado que el marco asigna parámetros independientemente a cada sitio. Segundo, reemplazar la matriz de correlación no estructurada por modelos de procesos Gaussianos (Stein, 1999; Banerjee et al., 2014) que modelen la correlación espacial como función continua de la distancia entre sitios, habilitando estimaciones hacia locaciones no monitoreadas. Tercero, el período 2019–2024 no incluye condiciones climáticas extremas históricamente documentadas en la región. Evaluar el marco utilizando datos de años históricamente extremos anteriores al período de estudio, o mediante escenarios sintéticos, permitiría caracterizar los límites de extrapolación de la calibración dependiente del estado. La evaluación del marco bajo escenarios de cambio climático constituye una pregunta abierta que requiere análisis específicos. Cuarto, acoplar el marco con inputs meteorológicos de pronóstico y productos satelitales de baja latencia para transitar hacia despliegue operacional en tiempo real. Quinto, mejorar la diversidad estructural del ensamble mediante la incorporación de algoritmos con principios físicos distintos a los modelos OpenET actuales, que comparten estructuras de error correlacionadas. Sexto, mejorar el desempeño individual de SEB-PV, particularmente la representación de las dinámicas térmicas del suelo en sitios semiáridos.

Conclusiones

La operacionalización de SEB-PV utilizando productos de humedad del suelo grillados y datos de estaciones meteorológicas estándar alcanzó precisión operacional (NRMSE 21–

29%) mediante ajuste paramétrico apropiado. Los algoritmos de escalamiento temporal exhiben dependencia del contexto climático: métodos basados en radiación neta resultaron superiores en avellanos bajo clima mediterráneo, mientras que métodos basados en variables meteorológicas fueron superiores en pistachos bajo condiciones semiáridas. Esta diferenciación refleja controladores físicos dominantes, disponibilidad energética versus demanda atmosférica, con implicaciones directas para implementación operacional.

La evaluación factorial demuestra que la estructura de error domina sobre la complejidad de calibración: componentes espaciales y temporales explicaron 95–98% de la varianza residual. La arquitectura óptima redujo el error normalizado en un 46% respecto al promedio determinístico simple, validando la hipótesis central, y mostró mayor estabilidad bajo condiciones climáticas no observadas durante calibración.

Los resultados establecen principios para diseño de sistemas de monitoreo satelital en agricultura. La inversión metodológica debe priorizar la modelación de dependencias espaciotemporales antes que el refinamiento de calibraciones. La selección de algoritmos de escalamiento temporal depende del contexto climático dominante. El marco probabilístico desarrollado produce distribuciones predictivas posteriores que cuantifican la incertidumbre asociada a cada estimación, proveyendo a usuarios y gestores una herramienta para discernir cuándo confiar y cuándo cuestionar los productos satelitales disponibles.

La transición desde enfoques determinísticos hacia marcos probabilísticos con estructura de error espaciotemporal explícita representa un paso metodológico concreto hacia la adopción operacional de estimaciones satelitales de ET en agricultura bajo presión hídrica. Las herramientas computacionales están disponibles públicamente.

BIBLIOGRAFÍA

Allen, R. G., Morse, A., Tasumi, M., Trezza, R., Bastiaanssen, W. G. M., Wright, J. L., and Kramber, W. (2002). Evapotranspiration from a satellite-based surface energy balance for the snake plain aquifer in idaho. In *Proceedings of the 2002 USCID/EWRI Conference on Energy, Climate, Environment, and Water*, pages 167–178, San Luis Obispo, California. U.S. Committee on Irrigation and Drainage. Accessed: 2025-06-04.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements-FAO*, volume 56 of *Irrigation and Drainage Paper*. FAO, Rome. 300: D05109.

Allen, R. G., Tasumi, M., and Trezza, R. (2007). Satellite-based energy balance for mapping evapotranspiration with internalized calibration (metric)—model. *J. Irrig. Drain. Eng.*, 133(4):380–394.

Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A., and Kustas, W. P. (2007). A climatological study of evapotranspiration and moisture stress across the continental united states based on thermal remote sensing: 1. model formulation. *J. Geophys. Res. Atmos.*, 112(D10).

ASCE-EWRI (2005). *The ASCE Standardized Reference Evapotranspiration Equation*. American Society of Civil Engineers.

ASTM (2000). Standard solar constant and zero air mass solar spectral irradiance tables. Standard E490-00, American Society for Testing and Materials.

Bai, Y., Zhang, S., Bhattarai, N., Mallick, K., Liu, Q., Tang, L., Im, J., Guo, L., and Zhang, J. (2021). On the use of machine learning based ensemble approaches to improve evapotranspiration estimates from croplands across a wide environmental gradient. *Agric. For. Meteorol.*, 298-299:108308.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, New York, 2nd edition.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F. (2018). Present and future köppen-geiger climate classification maps at 1-km resolution. *Sci. Data*, 5(1):180214.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.*, 120:70–83.

Bhattarai, N., Mallick, K., Stuart, J., Vishwakarma, B. D., Niraula, R., Sen, S., and Jain, M. (2019). An automated multi-model evapotranspiration mapping framework using remotely sensed and reanalysis data. *Remote Sens. Environ.*, 229:69–92.

Brutsaert, W. and Sugita, M. (1992). Application of self-preservation in the diurnal evolution of the surface energy budget to determine daily evaporation. *J. Geophys. Res. Atmos.*, 97(D17):18377–18382.

Cammalleri, C., Anderson, M. C., and Kustas, W. P. (2014). Upscaling of evapotranspiration fluxes from instantaneous to daytime scales for thermal remote sensing applications. *Hydrol. Earth Syst. Sci.*, 18(5):1885–1894.

Chen, J. M. and Liu, J. (2020). Evolution of evapotranspiration models using thermal and shortwave remote sensing data. *Remote Sens. Environ.*, 237:111594.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chávez, J. L., Neale, C. M. U., Prueger, J. H., and Kustas, W. P. (2008). Daily evapotranspiration estimates from extrapolating instantaneous airborne remote sensing et values. *Irrig. Sci.*, 27(1):67–81.

Cigarra-Guíñez, L. E., Lagos, O., Steduto, P., Krogh, S. A., Shapiro, K., Souto, C., Lillo-Saavedra, M., Balbontín, C., and Zaccaria, D. (2025). Enhancing evapotranspiration estimates in orchards with the surface energy balance for partially vegetated surfaces (seb-pv) model through combined use of gridded soil moisture and temporal upscaling methods. *Sci. Total Environ.*, 1002:180569.

Colaizzi, P. D., Evett, S. R., Howell, T. A., and Tolk, J. A. (2006). Comparison of five models to scale daily evapotranspiration from one-time-of-day measurements. *Trans. ASABE*, 49(5):1409–1417.

Dhungel, R., Anderson, R. G., French, A. N., Skaggs, T. H., Ajami, H., and Wang, D. (2024). Intercomparison of citrus evapotranspiration among eddy covariance, openet ensemble models, and the water and energy balance model (baitsss). *Agric. Water Manag.*, 304:109066.

Dubova, M., Chandramouli, S., Gigerenzer, G., Grünwald, P., Holmes, W., Lombrozo, T., Marelli, M., Musslick, S., Nicenboim, B., Ross, L. N., Shiffrin, R., White, M., Wagenmakers, E.-J., Bürkner, P.-C., and Sloman, S. J. (2025). Is ockham's razor losing its edge? new perspectives on the principle of model parsimony. *Proc. Natl. Acad. Sci. U. S. A.*, 122(5):e2401230121.

Er-Raki, S., Chehbouni, A., Guemouria, N., Ezzahar, J., Khabba, S., Boulet, G., and and, L. H. (2009). Citrus orchard evapotranspiration: Comparison between eddy covariance measurements and the fao-56 approach estimates. *Plant Biosyst.*, 143(1):201–208.

F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628.

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., Wang, A., Anderson, R. G., Aragon, B., Arain, M. A., Baldocchi, D. D., Baker, J. M., Barral, H., Bernacchi, C. J., Bernhofer, C., Biraud, S. C., Bohrer, G., Brunsell, N., Cappelaere, B., Castro-Contreras, S., Chun, J., Conrad, B. J., Cremonese, E., Demarty, J., Desai, A. R., De Ligne, A., Foltýnová, L., Goulden, M. L., Griffis, T. J., Grünwald, T., Johnson, M. S., Kang, M., Kelbe, D., Kowalska, N., Lim, J.-H., Maïnassara, I., McCabe, M. F., Missik, J. E., Mohanty, B. P., Moore, C. E., Morillas, L., Morrison, R., Munger, J. W., Posse, G., Richardson, A. D., Russell, E. S., Ryu, Y., Sanchez-Azofeifa, A., Schmidt, M., Schwartz, E., Sharp, I., Šigut, L., Tang, Y., Hulley, G., Anderson, M., Hain, C., French, A., Wood, E., and Hook, S. (2020). Ecostress: Nasa's next generation mission to measure evapotranspiration from the international space station. *Water Resour. Res.*, 56(4):e2019WR026058. e2019WR026058 2019WR026058.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017).

Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.*, 202:18–27. Big Remotely Sensed Data: tools, applications and experiences.

Grisafi, F., DeJong, T. M., and Tombesi, S. (2021). Fruit tree crop models: an update. *Tree Physiol.*, 42(3):441–457.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, 377(1):80–91.

Harrison, L. P. (1965). *Fundamental Concepts and Definitions Relating to Humidity*. Reinhold Publishing Company.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Model Assessment and Selection*, chapter 7, pages 233–235. Springer New York, New York, NY.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The era5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146(730):1999–2049.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(47):1593–1623.

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., and Ames, D. P. (2019). Introductory overview: Error metrics for hydrologic modelling – a review of common practices and an open source library to facilitate use and adoption. *Environ. Model. Softw.*, 119:32–48.

Jackson, R., Hatfield, J., Reginato, R., Idso, S., and Pinter, P. (1983). Estimation of daily evapotranspiration from one time-of-day measurements. *Agric. Water Manag.*, 7(1):351–362. Plant Production and Management Under Drought Conditions.

Jiang, L., Zhang, B., Han, S., Chen, H., and Wei, Z. (2021). Upscaling evapotranspiration from the instantaneous to the daily time scale: Assessing six methods including an optimized coefficient based on worldwide eddy covariance flux network. *J. Hydrol.*, 596:126135.

Jin, Y., He, R., Marino, G., Whiting, M., Kent, E., Sanden, B. L., Culumber, M., Ferguson, L., Little, C., Grattan, S., Paw U, K. T., Lagos, L. O., Snyder, R. L., and Zaccaria, D. (2018). Spatially variable evapotranspiration over salt affected pistachio orchards analyzed with satellite remote sensing estimates. *Agric. For. Meteorol.*, 262:178–191.

Kaimal, J. C. and Finnigan, J. J. (1994). *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford University Press.

Ke, Y., Im, J., Park, S., and Gong, H. (2016). Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches. *Remote Sens.*, 8(3):215.

Kell, G. (1975). Density, thermal expansivity, and compressibility of liquid water from 0℃ to 150℃: correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *J. Chem. Eng. Data*, 20:97–105.

Kim, H.-J., Chandrasekara, S., Kwon, H.-H., Lima, C., and woong Kim, T. (2023). A novel multi-scale parameter estimation approach to the hargreaves-samani equation for estimation of penman-monteith reference evapotranspiration. *Agric. Water Manag.*, 275:108038.

Kljun, N., Calanca, P., Rotach, M. W., and Schmid, H. P. (2015). A simple two-dimensional parameterisation for flux footprint prediction (ffp). *Geosci. Model Dev.*, 8(11):3695–3713.

Klosterhalfen, A., Chi, J., Kljun, N., Lindroth, A., Laudon, H., Nilsson, M. B., and Peichl, M. (2023). Two-level eddy covariance measurements reduce bias in land-atmosphere exchange estimates over a heterogeneous boreal forest landscape. *Agric. For. Meteorol.*, 339:109523.

Kruschke, J. K. (2015). Chapter 16 - metric-predicted variable on one or two groups. In Kruschke, J. K., editor, *Doing Bayesian Data Analysis (Second Edition)*, pages 449–475. Academic Press, Boston, second edition edition.

Lagos, L. O. (2008). *A Modified Surface Energy Balance for Modeling Evapotranspiration and Canopy Resistance*. PhD thesis, University of Nebraska, Lincoln, NE, USA.

Lagos, L. O., Martin, D. L., Verma, S. B., Suyker, A., and Irmak, S. (2009). Surface energy balance model of transpiration from variable canopy cover and evaporation from residue-covered or bare-soil systems. *Irrig. Sci.*, 28(1):51–64.

Lagos, L. O., Merino, G., Martin, D., Verma, S., and Suyker, A. (2012). Evapotranspiration of partially vegetated surfaces. In *Evapotranspiration - Remote Sensing and Modeling*, chapter 13, pages 273–304. IntechOpen.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74(6):1659–1673.

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., and Mukhopadhyay, A. (2021). Rmse is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *J. Atmos. Sol-Terr. Phys.*, 218:105624.

Liu, X., Xu, J., Yang, S., Lv, Y., and Zhuang, Y. (2020). Temporal upscaling of rice evapotranspiration based on canopy resistance in a water-saving irrigated rice field. *J. Hydrometeorol.*, 21(7):1639–1654.

Liu, Z. (2021). The accuracy of temporal upscaling of instantaneous evapotranspiration to daily values with seven upscaling methods. *Hydrol. Earth Syst. Sci.*, 25(8):4417–4433.

Lu, Y., Lu, S., Horton, R., and Ren, T. (2014). An empirical model for estimating soil thermal conductivity from texture, water content, and bulk density. *Soil Sci. Soc. Am. J.*, 78:1859–1868.

Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H. P., Schmidt, M., and Steinbrecher, R. (2013). A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements. *Agric. For. Meteorol.*, 169:122–135.

Melton, F. S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J. B., Kilic, A., Senay, G. B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P. M., Schull, M., Wang, C., Yang, Y., and Anderson, R. G. (2022). Openet: Filling a critical data gap in water management for the western united states. *J. Am. Water Resour. Assoc.*, 58(6):971–994.

Mobe, N., Dzikiti, S., Zirebwa, S., Midgley, S., von Loeper, W., Mazvimavi, D., Ntshidi, Z., and Jovanovic, N. (2020). Estimating crop coefficients for apple orchards with varying canopy cover using measured data from twelve orchards in the western cape province, south africa. *Agric. Water Manag.*, 233:106103.

Odi-Lara, M., Campos, I., Neale, C. M. U., Ortega-Farías, S., Poblete-Echeverría, C., Balbontín, C., and Calera, A. (2016). Estimating evapotranspiration of an apple orchard using a remote sensing-based soil water balance. *Remote Sens.*, 8(3).

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., and Yakir, D. (2006). Towards a standardized

processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, 3(4):571–583.

Pastorello, G., Agarwal, D., Papale, D., Samak, T., Trotta, C., Ribeca, A., Poindexter, C., Faybishenko, B., Gunter, D., Hollowgrass, R., and Canfora, E. (2014). Observational data patterns for time series data quality assessment. In *2014 IEEE 10th International Conference on e-Science*, volume 1, pages 271–278.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., et al. (2020). The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Sci. Data*, 7(1):225.

Pathak, T. B., Maskey, M. L., Dahlberg, J. A., Kearns, F., Bali, K. M., and Zaccaria, D. (2018). Climate change trends and impacts on california agriculture: A detailed review. *Agronomy*, 8(3).

Paw U, K. T., Qiu, J., Su, H.-B., Watanabe, T., and Brunet, Y. (1995). Surface renewal analysis: a new method to obtain scalar fluxes. *Agric. For. Meteorol.*, 74(1):119–137.

Petrakis, R. E., Norman, L. M., Villarreal, M. L., Senay, G. B., Friedrichs, M. O., Cassassuce, F., Gomis, F., and Nagler, P. L. (2024). An ensemble mean method for remote sensing of actual evapotranspiration to estimate water budget response across a restoration landscape. *Remote Sens.*, 16(12).

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7:217–240.

Purdy, A. J., Fisher, J. B., Goulden, M. L., Colliander, A., Halverson, G., Tu, K., and Famiglietti, J. S. (2018). Smap soil moisture improves global evapotranspiration. *Remote Sens. Environ.*, 219:1–14.

Quezada, L., Holzapfel, E., Kuschel-Otárola, M., Lillo-Saavedra, M., Rivera, D., Souto, C., Lagos, O., and Palma, D. (2025). Optimization of water and land allocation in fruit orchards over a 20-year period. *Water*, 17(1).

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133(5):1155 – 1174.

Rahmati, M., Amelung, W., Brogi, C., Dari, J., Flammini, A., Bogena, H., Brocca, L., Chen, H., Groh, J., Koster, R. D., McColl, K. A., Montzka, C., Moradi, S., Rahi, A., Sharghi S., F., and Vereecken, H. (2024). Soil moisture memory: State-of-the-art and the way forward. *Reviews of Geophysics*, 62(2):e2023RG000828. e2023RG000828 2023RG000828.

Reichle, R., De Lannoy, G., Koster, R., Crow, W., Kimball, J., Liu, Q., and Bechtold, M. (2022). Smap L4 global 3-hourly 9 km ease-grid surface and root zone soil moisture geophysical data, version 7.

Reichle, R. H., Lannoy, G. J. M. D., Liu, Q., Koster, R. D., Kimball, J. S., Crow, W. T., Ardizzone, J. V., Chakraborty, P., Collins, D. W., Conaty, A. L., Girotto, M., Jones, L. A., Kolassa, J., Lievens, H., Lucchesi, R. A., and Smith, E. B. (2017). Global assessment of the smap Level-4 surface and root-zone soil moisture product using assimilation diagnostics. *J. Hydrometeorol.*, 18(12):3217 – 3237.

Reitz, M., Volk, J. M., Ott, T., Anderson, M., Senay, G. B., Melton, F., Kilic, A., Allen, R., Fisher, J. B., Ruhoff, A., Purdy, A. J., and Huntington, J. (2025). Performance mapping and weighting for the evapotranspiration models of the openet ensemble. *Water Resour. Res.*, 61(8):e2024WR038899. e2024WR038899 2024WR038899.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y., Chuang, H., Iredell, M., et al. (2014). The NCEP climate forecast system version 2. *J. Clim.*, 27(6):2185–2208.

Samani, Z., Bawazir, S. A., Bleiweiss, M., Skaggs, R., and Tran, V. D. (2007). Estimating daily net radiation over vegetation canopy through remote sensing and climatic data. *J. Irrig. Drain. Eng.*, 133(4):291–297.

Sandoval-Solis, S. (2020). Water resources management in california. In Vieira, E. d. O., Sandoval-Solis, S., Pedrosa, V. d. A., and Ortiz-Partida, J. P., editors, *Integrated Water Resource Management: Cases from Africa, Asia, Australia, Latin America and USA*, pages 35–44. Springer International Publishing, Cham.

Shuttleworth, W., Gurney, R., Hsu, A., and Ormsby, J. (1989). *FIFE: The Variation in Energy Partition at Surface Flux Sites*. IAHS, Wallingford.

Shuttleworth, W. J. and Wallace, J. S. (1985). Evaporation from sparse crops-an energy combination theory. *Q. J. R. Meteorol. Soc.*, 111(469):839–855.

Singh, L. B. (1948). Studies in biennial bearing ii. a review of the literature. *J. Hortic. Sci.*, 24(1):45–65.

Souto, C., Lagos, O., Holzapfel, E., Maskey, M. L., Wunderlich, L., Shapiro, K., Marino, G., Snyder, R., and Zaccaria, D. (2019). A modified surface energy balance to estimate crop transpiration and soil evaporation in micro-irrigated orchards. *Water*, 11(9).

Souto, C., Lagos, O., Holzapfel, E., Ruybal, C., Bryla, D. R., and Vidal, G. (2022). Evaluating a surface energy balance model for partially wetted surfaces: Drip and micro-sprinkler systems in hazelnut orchards (corylus avellana l.). *Water*, 14(24).

Stan Development Team (2025). RStan: the R interface to Stan. R package version 2.32.7.

Steduto, P., Allen, R. G., Kilic, A., et al. (2023). *Remote sensing determination of evapotranspiration – Algorithms, strengths, weaknesses, uncertainty and best fit-for-purpose*. FAO, Cairo, Egypt. Lead authors: Pasquale Steduto, Rick G. Allen, Ayse Kilic.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stojanović, O., Siegmann, B., Jarmer, T., Pipa, G., and Leugering, J. (2022). Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets. *Front. Environ. Sci.*, Volume 9 - 2021.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*, volume 13 of *Atmospheric and Oceanographic Sciences Library*. Kluwer Academic Publishers, Dordrecht.

Sun, H., Yang, Y., Wu, R., Gui, D., Xue, J., Liu, Y., and Yan, D. (2019). Improving estimation of cropland evapotranspiration by the bayesian model averaging method with surface energy balance models. *Atmosphere*, 10(4).

Tran, B. N., van der Kwast, J., Seyoum, S., Uijlenhoet, R., Jewitt, G., and Mul, M. (2023). Uncertainty assessment of satellite remote-sensing-based evapotranspiration estimates: a systematic review of methods and gaps. *Hydrol. Earth Syst. Sci.*, 27(24):4505–4528.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Stat. Comput.*, 27(5):1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Anal.*, 16(2):667 – 718.

Vickers, D. and Mahrt, L. (1997). Quality control and flux sampling problems for tower and aircraft data. *J. Atmos. Ocean. Technol.*, 14(3):512 – 526.

Wang, K. and Dickinson, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Rev. Geophys.*, 50(2):RG2005.

Weber, C., Simnitt, S., Wechsler, S., and Wakefield, H. (2023). Fruit and Tree Nuts Outlook: September 2023. Situation and Outlook Report FTS-377, U.S. Department of Agriculture, Economic Research Service. Accessed on 2023-10-27.

Wu, Y., Miao, C., Wang, Y., Zhang, Q., Ji, J., and Chai, Y. (2025). Multi-spatial scale assessment and multi-dataset fusion of global terrestrial evapotranspiration datasets. *Earth Syst. Sci. Data Discuss.*, 2025:1–27.