UC Davis

UC Davis Previously Published Works

Title

Lead-time and cumulative-error evaluation of NBM-derived reference evapotranspiration forecasts using CIMIS records in California

Permalink

https://escholarship.org/uc/item/6sb9r80k

Journal

Agricultural Water Management, 337

ISSN

0378-3774

Author

Peddinti, SR

Publication Date

2026-12-01

DOI

10.1016/j.agwat.2026.110857

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Agricultural Water Management 337 (2026) 110857

Contents lists available at ScienceDirect

Agricultural Water Management

journal homepage: www.elsevier.com/locate/agwat

Lead-time and cumulative-error evaluation of NBM-derived reference evapotranspiration forecasts using CIMIS records in California

Srinivasa Rao Peddinti *

Department of Land, Air & Water Resources, University of California, Davis, Davis, CA, 95616, USA

ARTICLE INFO

Keywords:
Atmospheric water demand
Forecast verification
Persistence benchmark
Seasonal variability
Meteorological drivers
Water management

ABSTRACT

Reliable reference evapotranspiration (ETo) forecasts can inform planning for atmospheric water demand. This study evaluated National Blend of Models (NBM)-derived daily ETo against California Irrigation Management Information System (CIMIS) station-based ETo records at 132 stations from 1 January 2021 through 11 June 2026. The analysis combined lead-dependent daily performance, seasonal and spatial cumulative bias, and rolling cumulative fixed-lead errors, using 714,041 station–lead–date pairs at leads of 1, 3, and 7 days. Pooled RMSE was 0.70, 0.79, and 0.95 mm day^{-1} at 1-, 3-, and 7-day leads, respectively, while R^2 decreased from 0.91 to 0.88 and 0.83. Station-clustered confidence intervals supported the increase in absolute error with lead time. Compared with a prospective persistence benchmark based on the most recent fully available CIMIS ETo, NBM reduced RMSE by 36%, 36%, and 27% at 1-, 3-, and 7-day leads, respectively. Relative to leave-one-year-out station-month climatology, RMSE reductions were 32%, 24%, and 8%. Errors were smallest in winter and largest in summer. Cumulative station bias varied substantially among complete years, but the overall ordering of forecast performance with lead time remained unchanged when the highest-error year was excluded. For 3-day rolling windows, absolute cumulative ETo error exceeded 5 mm in 1.89%, 2.54%, and 4.35% of windows at 1-, 3-, and 7-day leads, respectively; for 14-day windows, the corresponding frequencies were 36.18%, 42.37%, and 46.36%. Because each rolling window combined same-lead forecasts for consecutive target dates from different initializations, these values describe accumulated fixed-lead forecast error rather than a single multi-day forecast sequence. NBM-derived ETo provided useful prospective guidance on atmospheric water demand, with the strongest performance at shorter lead times, and can serve as an input to local irrigation planning when combined with crop- and field-specific information.

1. Introduction

Reference evapotranspiration (ETo) is a foundational variable for quantifying atmospheric water demand, estimating crop water requirements, supporting irrigation scheduling, and assessing drought severity. ETo represents the atmospheric evaporative demand from a standardized well-watered reference surface and is commonly combined with crop coefficients to estimate crop evapotranspiration. The FAO-56 Penman-Monteith approach established a widely adopted framework for computing ETo from meteorological variables, including air temperature, humidity, wind speed, and radiation (Allen et al., 1998). The ASCE standardized ETo equation further refined computational procedures for short and tall reference surfaces, providing a consistent basis for operational and research applications (Allen et al., 2005). Because ETo integrates multiple weather drivers, accurate estimation and forecasting of ETo are particularly important in irrigated agricultural regions where water applications must be planned ahead of time.

In California, the California Irrigation Management Information System (CIMIS) operates a statewide network of automated weather stations that provide meteorological observations and calculated ETo estimates to support irrigation scheduling and water-resource planning (CIMIS, 2026). CIMIS has been widely used by growers, consultants, water agencies, and researchers because it provides standardized weather and ETo information across major agricultural regions of the state. In addition to station-based ETo, spatial CIMIS products have been developed to extend ETo estimates across California by combining station measurements, satellite-derived radiation, and interpolation of meteorological variables (Hart et al., 2009). These products support regional water-demand estimation, irrigation planning, and crop water-use studies; however, prospective decisions require reliable

* Correspondence to: Department of Land, Air & Water Resources, University of California, Davis, One Shields Avenue, Davis, CA 95616-5270, USA.
E-mail address: speddinti@ucdavis.edu.

https://doi.org/10.1016/j.agwat.2026.110857
Received 4 July 2026; Received in revised form 9 October 2026; Accepted 11 October 2026
Available online 14 October 2026
0378-3774/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

information on future atmospheric demand rather than only current or historical ETo.

This study addresses the need for a statewide evaluation of NBM-derived ETo across California's agricultural station network, spanning multiple forecast leads and years and both daily and cumulative time scales. It complements earlier regional forecast-ETo studies by combining benchmark comparisons with seasonal, spatial, and meteorological-error diagnostics. Previous studies have evaluated gridded meteorological and ETo products for agricultural applications. Abatzoglou (2013) developed the gridMET dataset to provide spatially complete gridded meteorological data for ecological and hydrological applications, and subsequent work evaluated gridded weather datasets for estimating standardized reference evapotranspiration against weather-station records (Blankenau et al., 2020). These studies emphasized that ETo estimates derived from gridded meteorological data depend strongly on the quality of the underlying weather variables and the representativeness of station measurements. This provides an important basis for evaluating forecast-derived ETo because errors in temperature, humidity, radiation, and wind can propagate into the final ETo estimate (Vanella et al., 2020). Early studies explored the use of public weather forecast data with the FAO Penman-Monteith equation to estimate daily ETo (Cai et al., 2007). Other work evaluated short-term ETo forecasting using public weather forecasts and Penman-Monteith-based approaches across multiple climatic regions (Yang et al., 2016). More recent studies compared public weather forecasts, numerical weather prediction outputs, and different ETo modeling approaches for daily forecasting (Liu et al., 2020; Vanella et al., 2020). Collectively, these studies established that meteorological forecasts could provide prospective ETo information, while also demonstrating the importance of local evaluation.

Numerical weather prediction (NWP) products have also been used to forecast ETo from regional to national scales. Perera et al. (2014) evaluated daily ETo forecasts derived from NWP outputs across Australia. Tian and Martinez (2014) developed GEFS-based daily ETo forecasts for water-management applications in the southeastern United States. Medina et al. (2018) extended this concept to medium-range ETo forecasts across the contiguous United States using multi-model numerical weather predictions, and Medina and Tian (2020) investigated probabilistic post-processing approaches for NWP-based daily and weekly ETo forecasts. These studies demonstrate that NWP-based ETo forecasting is technically feasible and potentially useful, but performance depends on forecast lead time, model source, climate region, and post-processing approach.

In the United States, the National Weather Service has produced Forecast Reference Evapotranspiration (FRET), an operational product based on Penman-Monteith ETo calculations using forecast meteorological inputs (U.S. Department of Commerce, 2026). FRET has been evaluated in western U.S. agricultural settings. Ben Hamouda et al. (2022) compared FRET ETo forecasts with CIMIS station-based ETo across selected California microclimate regions, while McEvoy et al. (2022) evaluated FRET and National Digital Forecast Database inputs against agricultural weather-station records in Nevada and examined bias-correction approaches. These studies are directly relevant to forecast ETo applications in the western United States, but their scope differs from a statewide, multi-year evaluation of ETo derived directly from National Blend of Models (NBM) meteorological forecasts across a dense agricultural station network and multiple forecast leads.

The NBM provides a nationally consistent suite of calibrated gridded forecast guidance based on blended National Weather Service and non-National Weather Service model output (Hamill et al., 2017; NOAA, 2026). Its meteorological fields can be used to calculate standardized daily ETo over multiple forecast horizons. Deriving ETo directly from the underlying NBM meteorological variables provides several analytical advantages for evaluation: the same standardized ETo formulation can be applied consistently across leads and locations, gridded forecast inputs can be paired with a statewide station network, and errors in individual meteorological drivers can be examined in relation to ETo error. This approach is complementary to operational ETo products such as FRET and provides a transparent framework for evaluating how NBM meteorological forecast errors propagate into reference ETo.

A remaining need is a statewide California assessment that combines a dense agricultural station network, multiple years and forecast leads, seasonal and regional comparisons, and accumulated ETo errors. California provides coastal, valley, desert, mountain, and northern climatic settings within one agricultural monitoring network. A statewide evaluation can determine whether lead-dependent performance is consistent across seasons and regions and whether small daily errors accumulate into larger water-depth differences. Benchmark forecasts distinguish agreement from forecast skill, while anomaly analysis assesses agreement after removal of the shared seasonal cycle. Spatial summaries and meteorological-driver diagnostics further identify where and under what atmospheric conditions errors are most pronounced. The contribution is this integrated evaluation of NBM-derived ETo, rather than the introduction of forecast-based ETo itself.

Accordingly, this study evaluated NBM-derived daily ETo against CIMIS station-based ETo records across California during 2021–2026 at forecast leads of 1, 3, and 7 days. The analysis was designed to combine daily forecast evaluation with benchmark-relative skill, seasonal and interannual variability, anomaly performance, cumulative ETo behavior, spatial variability, and rolling cumulative fixed-lead forecast errors. It was hypothesized that NBM-derived ETo would reproduce the general daily variability of CIMIS ETo, but that performance would deteriorate with increasing lead time; that forecast errors would vary seasonally and spatially, with larger departures during high-ETo periods; and that modest daily errors would accumulate over consecutive target dates.

The specific objectives of this study were to:

1. Evaluate the daily performance of NBM-derived ETo against CIMIS station-based ETo at 1-, 3-, and 7-day forecast leads using standard error and agreement statistics, station-clustered uncertainty estimates, and benchmark-relative skill.

2. Assess seasonal and interannual variation in forecast performance and determine how much daily agreement remains after removal of station-specific monthly climatology.

3. Quantify cumulative ETo agreement and bias using common station-date support, including annual station-level and regional spatial variability across the CIMIS network.

4. Characterize rolling cumulative fixed-lead ETo forecast errors over 3-, 7-, and 14-day windows, including the frequency and sign of accumulated errors exceeding diagnostic water-depth thresholds.

5. Examine associations between errors in the meteorological variables used to calculate NBM-derived ETo and the resulting ETo error, with particular attention to year- and season-dependent behavior.

2. Materials and methods

2.1. Study area, CIMIS reference ETo, and study period

The evaluation covered California from 1 January 2021 through 11 June 2026 and used the California Irrigation Management Information System (CIMIS) as the station-based reference dataset (Fig. 1). CIMIS is operated by the California Department of Water Resources and reports meteorological observations together with calculated reference evapotranspiration for irrigation and water-management applications (CIMIS, 2026). Daily ASCE standardized short-reference ETo was obtained from the CIMIS application programming interface (API; data item day-asce-eto). Station metadata included identifier, name, county, latitude, longitude, elevation, activation period, and ETo-station status. CIMIS ETo is calculated from meteorological measurements and is not a direct measurement of evapotranspiration.

The final paired dataset contained 132 CIMIS stations, 238,414 unique station-date CIMIS ETo records, and 714,041 valid

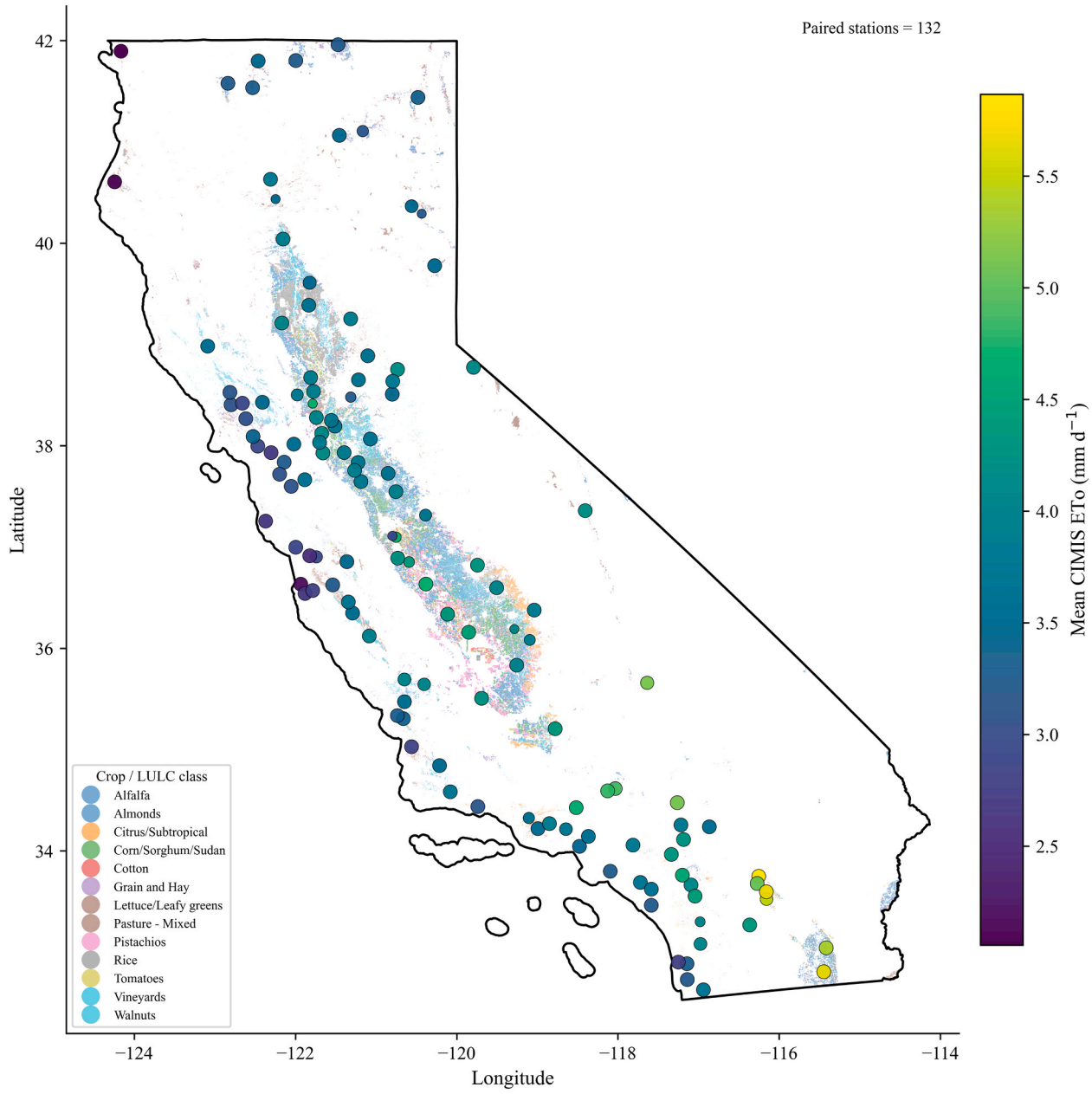


Fig. 1. California study area, CIMIS stations, and selected 2023 Department of Water Resources crop/land-use classes. Station color indicates mean CIMIS ETo over the study period, and marker size reflects the number of valid dates. The map includes the 132 paired stations; the crop layer provides context only.

station–lead–date pairs: 238,293 at lead 1, 238,055 at lead 3, and 237,693 at lead 7. Individual station records contained 137–1988 dates. The calendar coverage was 2021–2025 plus the partial 2026 period ending 11 June; station activation periods and data gaps meant that not every station contributed a complete year. The distribution of stations reflects the monitoring network and is not a probability-based or soil-stratified sampling design. A 2023 California Department of Water Resources crop/land-use layer provided agricultural context in Fig. 1 and was not a predictor or a statistical stratification variable (DWR, 2023).

2.2. *NBM forecast product, operational versions, and lead-time definition*

NBM meteorological guidance was obtained from the operational CONUS core product (Herbie model="nbm", product="co"), distributed as GRIB2 on a Lambert conformal grid with nominal 2.5-km spacing (NCEP, 2026). Historical files were retrieved using Herbie/FastHerbie, with the public Amazon Web Services archive preferred and NOAA NOMADS available as a fallback (Blaylock, 2024). The extracted inventory descriptors were TMP:2 m above ground (K), RH:2 m above ground (%), WIND:10 m above ground (m s^{-1}), and DSWRF: surface (W m^{-2}). The NBM inventory metadata indicated that DSWRF values represented instantaneous forecasts at the stated valid times rather than accumulated radiation energy. A fixed 06 UTC cycle provided a consistent daily initialization and the forecast-hour coverage required for all three leads; other cycles were not compared. For target date D, lead 1 used the 06 UTC run on D and f003–f026 at hourly spacing; lead 3 used the run on D−2 and f051–f072 at 3-hourly spacing; lead 7 used the run on D−6 and f147–f168 at 3-hourly spacing (Table S1). The labels describe target-day products, not exactly 24-, 72-, and 168-h verifying instants. They span near-term, intermediate, and approximately weekly guidance within the archived forecast range.

The study evaluated the evolving operational NBM stream rather than a homogeneous reforecast. The period spans versions 4.0, 4.1, 4.2, 4.3, and 5.0. Published implementation dates were 17 January 2023 for

version 4.1, 15 May 2024 for version 4.2, 27 May 2025 for version 4.3, and 5 May 2026 for version 5.0 (NOAA/MDL, 2026a; NWS, 2024; NOAA/MDL, 2025, 2026b). Version coverage was summarized by initialization date rather than target date. These date-based labels describe operational eras; implementation within a transition day may occur after the 06 UTC cycle. No performance claim was attributed to a particular version, and version labels did not affect the paired ETo metrics.

2.3. Fixed-PST calendar-day alignment and sub-daily aggregation

CIMIS daily reporting uses fixed Pacific Standard Time (PST; UTC−8), without daylight-saving adjustment (CIMIS, 2026). NBM valid times were converted to Etc/GMT + 8. Subtracting 1 s before deriving the local date assigned the midnight value to the preceding target day. A fixed-PST day extends from 08:00 UTC to 08:00 UTC the next day. Lead 1 contains 24 values labeled 01:00 through 00:00 local time. Leads 3 and 7 contain eight values labeled 01:00, 04:00, …, 22:00. Using 1-h weights for lead 1 and 3-h weights for leads 3 and 7 gives a total integration period of 24 h; however, the 3-hourly forecast times do not coincide with both the beginning and end of the fixed-PST day. Boundary reconstruction was therefore evaluated explicitly, rather than assuming that the eight-point forecasts were exact interval means.

For each station, date, and lead, air temperature was converted from kelvin to degrees Celsius. Daily maximum, minimum, and mean temperature were calculated from the available sub-daily samples. Relative humidity was retained in percent. Actual vapor pressure was calculated at each time step from temperature and relative humidity and then averaged over the day. Mean 10-m wind speed was calculated from the sub-daily WIND values. Downward shortwave radiation was converted from Wm^{-2} to MJm^{-2} over the represented time step using $0.0036\Delta t$, where Δt was 1 h for lead 1 and 3 h for leads 3 and 7. A daily record was accepted only when the expected number of samples was available for all four meteorological variables and the represented period totaled 24 h. Relative humidity was constrained to 0–100%. Daily values required all expected samples for each input; missing daily ETo was not interpolated. Negative daily ETo and negative vapor-pressure deficit were constrained to zero in the calculation.

The archived inventory identifies DSWRF as a point-in-time forecast in W m^{-2}, so the primary weighted sum is a rectangular numerical integration, not a conversion of an accumulated-energy variable. Lead-1 sensitivity calculations used f003, f006, …, f024 for 3-hourly temperature, humidity, and wind summaries while retaining the full radiation series integrated over the fixed-PST day. A separate radiation test used bracketing forecasts to reconstruct 00:00 and 24:00 PST and trapezoidal integration between these boundaries. Matched sensitivity comparisons retained 238,293, 237,701, and 237,575 records at leads 1, 3, and 7, respectively. Additional forecast times were required at leads 3 and 7 to reconstruct the fixed-PST day boundaries. The primary daily series was retained for evaluation, with the sensitivity results reported in Section 3.6 and Fig. S2. The lead-1 test isolates meteorological sampling while preserving daily radiation and is not a full 3-hour reconstruction of every input.

2.4. Calculation of NBM-derived reference evapotranspiration

Daily NBM-derived ETo was calculated with the standardized short-reference Penman-Monteith formulation following FAO-56/ASCE procedures (Allen et al., 1998, 2005). Daily soil heat flux was set to zero. The equation was

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma\left(\frac{900}{T+273}\right)u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad (1)$$

where ET_o is reference evapotranspiration (mmd^{-1}), Δ is the slope of the saturation-vapor-pressure curve ($kPa^{\circ}C^{-1}$), R_n is net radiation ($MJm^{-2}d^{-1}$), G is soil heat flux ($MJm^{-2}d^{-1}$), γ is the psychrometric constant ($kPa^{\circ}C^{-1}$), T is mean daily air temperature (°C), u_2 is wind speed at 2 m (ms^{-1}), e_s is saturation vapor pressure (kPa), and e_a is actual vapor pressure (kPa). Negative daily ET_o values, if produced by the equation, were set to zero. The vapor-pressure deficit used in the numerator was constrained to be nonnegative.

Saturation vapor pressure at temperature T was calculated as $e_s(T) = 0.6108\exp\left(\frac{17.27T}{T+237.3}\right)$, and $\Delta = \frac{4098e_s(T)}{(T+237.3)^2}$. Actual vapor pressure at each sub-daily time step was calculated as $e_a = e_s(T)RH/100$ and then averaged to the daily scale. Daily saturation vapor pressure was calculated from daily maximum and minimum temperature as the mean of $e_s(T_{max})$ and $e_s(T_{min})$. Atmospheric pressure was calculated from CIMIS station elevation z (m) as $P = 101.3\left(\frac{293-0.0065z}{293}\right)^{5.26}$, and $\gamma = 0.000665P$. NBM 10-m wind speed was converted to 2 m using $u_2 = u_{10}\frac{4.87}{\ln(67.8\times10-5.42)}$ (Allen et al., 1998).

Extraterrestrial radiation was calculated from latitude and day of year using the FAO-56 solar-geometry equations. Clear-sky radiation was $R_{so} = (0.75 + 2\times10^{-5}z)R_a$. Net shortwave radiation was $R_{ns} = (1-0.23)R_s$, where R_s is daily incoming shortwave radiation and 0.23 is the reference-surface albedo. Net longwave radiation was calculated from T_{max}, T_{min}, e_a, and the R_s/R_{so} cloudiness term using the FAO-56 Stefan-Boltzmann formulation; R_s/R_{so} was constrained to 0–1, the cloudiness term was bound below at 0.05, and the humidity term was constrained to be nonnegative. Net radiation was $R_n = R_{ns} - R_{nl}$. All elevation-dependent terms in the primary calculation used the CIMIS station elevation so that the atmospheric-pressure and clear-sky-radiation calculations represented the station location being verified.

The implemented longwave calculation was

$$R_{nl} = \sigma\frac{(T_{max} + 273.16)^4 + (T_{min} + 273.16)^4}{2}HC \quad (2)$$

where $H = \max(0.34 - 0.14\sqrt{\max(e_a, 0)}, 0)$, $C = \max(1.35q - 0.35, 0.05)$

Here $\sigma = 4.903\times10^{-9}$ MJ K^{-4} m^{-2} day^{-1} , and q is R_s/R_{s_o} constrained to 0–1 (q = 0 when $R_{s_o} \le 0$). Temperatures in this expression are in °C before conversion to kelvin. These bounds reproduce the calculation used for the forecast series. Extraterrestrial radiation follows the FAO-56 latitude/day-of-year solar-geometry equations (Allen et al., 1998).

2.5. Station-to-grid matching and spatial-representation sensitivity

Each CIMIS station was assigned once to the nearest NBM grid center using a k-dimensional tree search on geographic coordinates, and this assignment was reused throughout the record. Station-to-grid-center distances across the 132 paired stations averaged 0.95 km, with a median of 0.99 km, a 95th percentile of 1.48 km, and a maximum of 1.63 km. Nearest-grid extraction preserves a single native-grid forecast at each station and provides a consistent assignment across dates. A four-neighbor inverse-distance-weighted extraction was compared with nearest-grid extraction for one representative date per available meteorological season and year, at all three leads (23 dates per lead). This tests local weighted interpolation but is not literal bilinear interpolation. Because a static NBM terrain-height field was not available in the extracted historical core inventory, U.S. Geological Survey 3D Elevation Program (3DEP) elevation at the nearest grid center served as an external terrain proxy (USGS, 2026). Grid-center minus station elevation had median −0.04 m, mean absolute difference 13.84 m, RMSE 28.50 m, and maximum absolute difference 201.91 m. ETo was recalculated with this elevation in the pressure and clear-sky-radiation terms. No lapse-rate correction to temperature or other meteorological fields was applied; this was a sensitivity test of elevation-dependent ETo terms,

not a correction of model terrain or subgrid microclimate.

2.6. Pairing, quality control, completeness, and season definitions

CIMIS and NBM records were paired by station identifier, target date, and forecast lead. A paired record was retained only when CIMIS ETo and NBM-derived ETo were finite, the NBM daily record passed the complete-day checks described above, and station coordinates were valid. Missing daily values were not interpolated. The daily forecast error was defined consistently throughout the study as

$$e_i = ET_{o,\mathrm{NBM},i} - ET_{o,\mathrm{CIMIS},i} \tag{3}$$

Positive error indicates NBM overestimation and negative error indicates underestimation. Meteorological seasons were winter (December–February), spring (March–May), summer (June–August), and fall (September–November) (NOAA/NCEI, 2016). The partial 2026 record was included in pooled daily, seasonal, station-level, benchmark, anomaly, meteorological-driver, and rolling-error analyses. It was not treated as a complete year: the main annual spatial maps use 2021–2025, while the 2026 cumulative trajectory and station-bias map are identified separately as partial-year results.

CIMIS ETo quality-control (QC) flags were audited. No severe I, M, or S flags occurred among the retained numeric daily ETo values, so excluding these flags did not change the primary evaluation. Numeric ETo values with A, H, R, or E flags were retained. Meteorological-driver QC was applied separately to each variable (Section 2.11). An 85% threshold was used for annual spatial summaries to balance available station coverage against poorly observed station-years; it is a pragmatic inclusion rule, not an agronomic threshold. Expected active days were the overlap of the station activation period, calendar year, and study period. Map inclusion required at least 85% paired completeness for each of the three leads; cumulative map values were then summed on common dates. The threshold was not applied to all pooled daily analyses or to the descriptive cumulative trajectories. Sensitivity analyses at 80%, 90%, and 95% used lead-specific paired station-year sums to assess whether the principal annual-bias pattern depended on the threshold.

2.7. Daily forecast agreement, uncertainty, and lead-time comparisons

Daily agreement was summarized with bias, mean absolute error (MAE), root mean square error (RMSE), Pearson correlation (r), coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), and ordinary least-squares regression slope and intercept (Nash and Sutcliffe, 1970; Legates and McCabe, 1999; Willmott, 1982). For n paired records with error e_i, the principal error measures were

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{6}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(ET_{o,\mathrm{NBM},i} - ET_{o,\mathrm{CIMIS},i}\right)^2}{\sum_{i=1}^{n}\left(ET_{o,\mathrm{CIMIS},i} - \overline{ET_{o,\mathrm{CIMIS}}}\right)^2} \tag{7}$$

Here R^2 is the square of Pearson correlation, not one minus normalized squared error. NSE evaluates squared error relative to the CIMIS mean. Regression slope and intercept describe NBM ETo regressed on CIMIS ETo.

Metrics were calculated overall, by year and lead, and by meteorological season and lead. Because station-day observations are not statistically independent, uncertainty intervals were not based on individual records. Instead, 95% confidence intervals were obtained with a station-clustered nonparametric bootstrap using 1000 replicates, in which CIMIS stations were sampled with replacement and all records belonging to a sampled station were retained together (Efron and Tibshirani, 1993). Percentile intervals were defined by the 2.5th and 97.5th bootstrap percentiles. Pairwise lead-time differences in RMSE, MAE, and bias were evaluated on exact common station-date support with the same station-cluster bootstrap; a difference was considered statistically supported when its 95% interval excluded zero. This resampling retains within-station temporal dependence but does not fully model dependence among stations during shared weather events. Confidence intervals therefore describe uncertainty under the station-cluster design rather than independent station-day sampling.

Metrics were also calculated separately for each station and lead and summarized with equal station weight, including both means and medians. All station-level RMSE values were retained in the analysis and visualization. To assess agreement beyond the shared seasonal cycle, station-specific monthly CIMIS means and corresponding station-month-lead NBM means were removed from their respective series. These means were calculated over the evaluation record. The resulting monthly-anomaly statistics are descriptive diagnostics, not out-of-sample anomaly forecasts or an additional benchmark skill score. As an interannual robustness check, pooled metrics were also recalculated after excluding 2023, the complete year with the largest daily errors, to determine whether the overall lead-time ordering depended on this high-error year.

2.8. Benchmark forecasts and forecast-skill scores

Two benchmarks were evaluated on records available to both the benchmark and NBM. Prospective persistence used the latest completed CIMIS calendar day at the NBM initialization. Because 06 UTC is 22:00 fixed PST on the preceding local date, the persistence source dates were D−2, D−4, and D−8 for leads 1, 3, and 7. This definition excludes the still-incomplete local day at initialization; it assumes that the preceding completed daily record is available and does not reconstruct historical publication delays. The second benchmark was station-specific monthly climatology computed in leave-one-year-out form, excluding the target year. Because other years can occur after the target year, this is a retrospective cross-validated climatological benchmark, not an issuance-time operational hindcast. Missing benchmark values were omitted from the corresponding comparison.

NBM skill relative to each benchmark was summarized with RMSE and MAE skill scores

$$\mathrm{Skill}_X = 1 - \frac{X_{\mathrm{NBM}}}{X_{\mathrm{benchmark}}} \tag{8}$$

where X is RMSE or MAE. Positive values indicate lower error than the benchmark, zero indicates equal error, and negative values indicate poorer performance. Station-clustered bootstrap confidence intervals (1000 replicates) were calculated for the benchmark-relative skill scores. In this study, the term forecast skill is reserved for these benchmark-relative quantities; otherwise, NBM-CIMIS performance is described as agreement or forecast error.

2.9. Cumulative ETo, regional patterns, and spatial diagnostics

Cumulative analyses used identical station-date support across CIMIS and all three NBM leads. A station date was retained only when CIMIS ETo and all three forecasts were available. Sums were calculated in date order without imputation: accumulation resumed after gaps using the next common available date, so missing-day ETo was not added to any series. Consequently, cumulative values are totals over matched available dates rather than gap-filled full-year estimates. Fig. 3

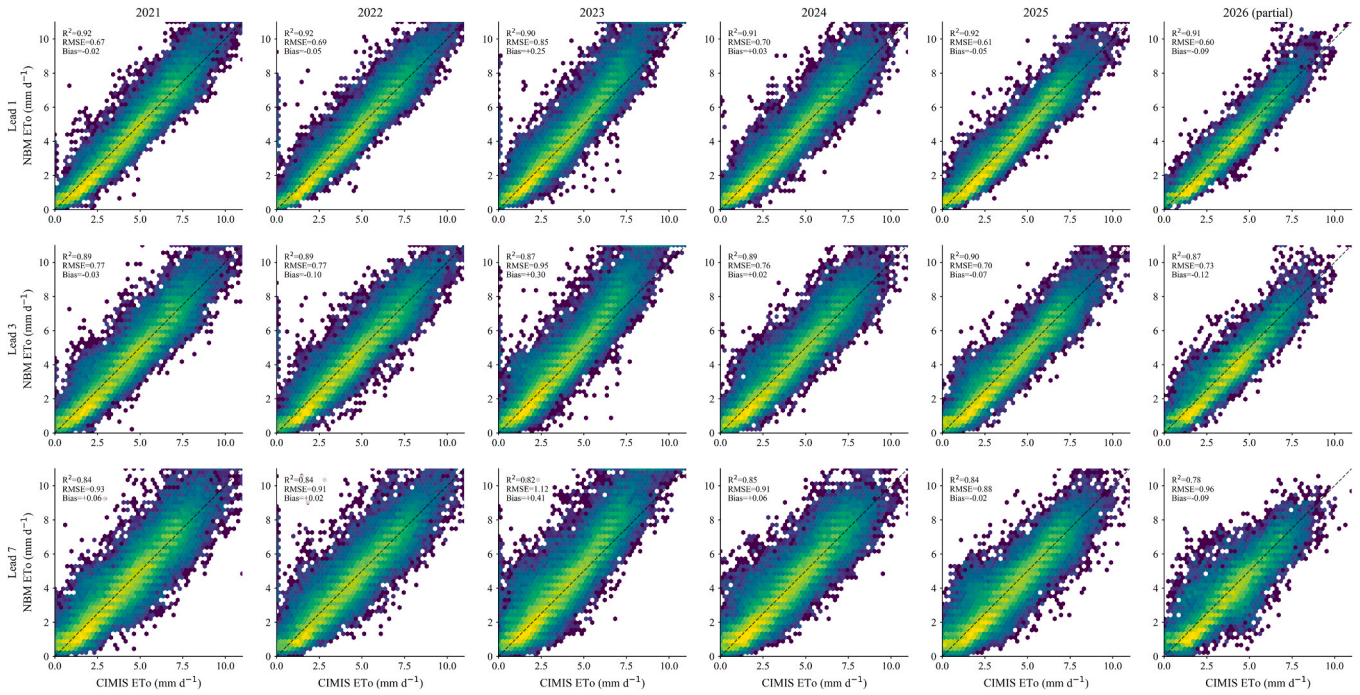


Fig. 2. Year-wise comparison of NBM-derived daily ETo with CIMIS station-based ETo at leads 1, 3, and 7. Columns show years and rows show leads. Hexagon color indicates record density; the diagonal is the 1:1 line. Annotations report R^2, RMSE, and bias. The 2026 column covers 1 January–11 June only.

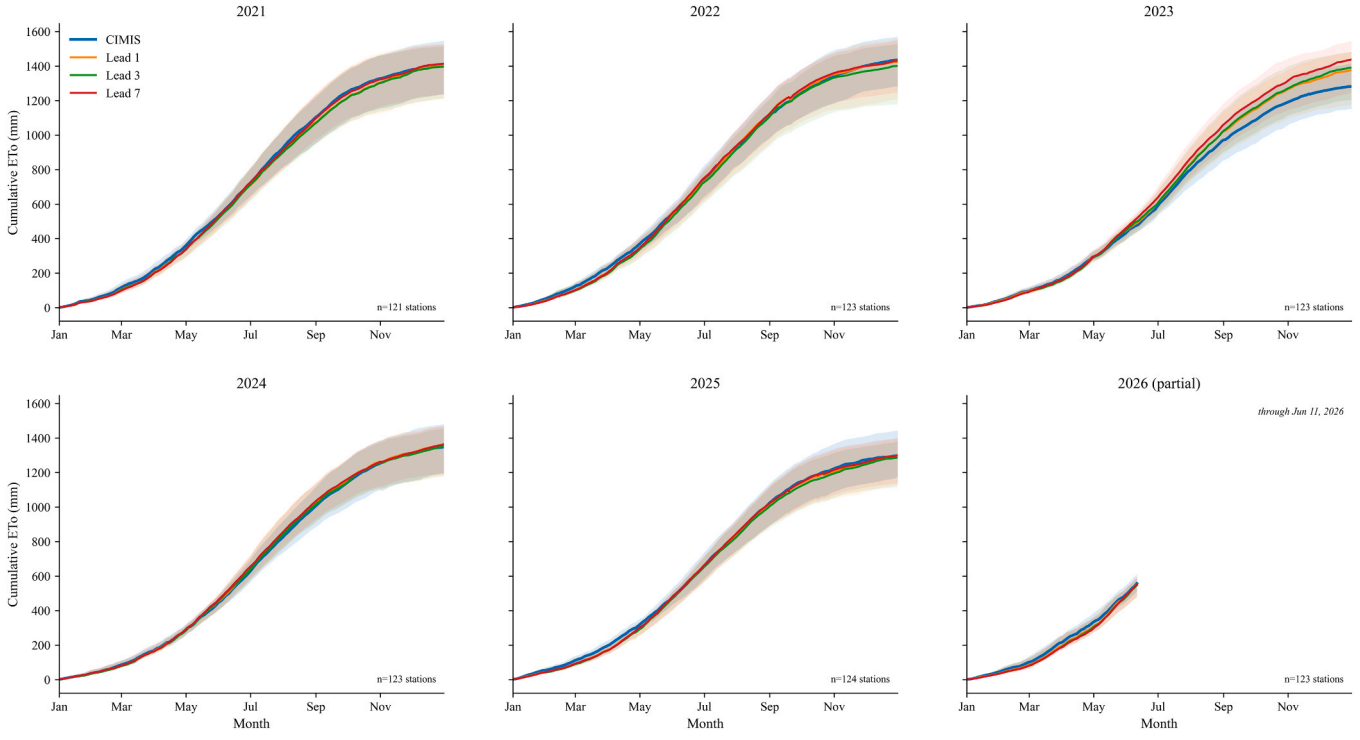


Fig. 3. Common-support cumulative ETo trajectories for CIMIS and the three forecast leads. All four series use the same available dates within a station-year; missing days are not imputed. Lines show daily across-station medians, and shading shows the interquartile ranges. The station set can vary by date. The 2026 panel ends on 11 June and is not an annual total. Table S2 separately summarizes medians of station-specific period totals.

summarizes all available common-support trajectories using the across-station median and interquartile range on each date; the contributing station set can change through a year. Station-specific period totals were also summarized. The difference between the medians of NBM and CIMIS station totals was kept distinct from the median of paired station-level differences (Table S2). These quantities need not equal one another or the endpoint difference of daily median curves when station coverage varies.

Annual cumulative station bias for station s, year y, and lead l was calculated on the common-date support as

$$B_{s,y,l} = \sum_{d} \left[ET_{o,\mathrm{NBM},s,y,l,d} - ET_{o,\mathrm{CIMIS},s,y,d} \right] \quad (9)$$

Only station-years meeting the 85% criterion for each lead were included in the main annual spatial summaries. Fig. 5 presents calculated station-level cumulative bias for 2021–2025; Fig. S1 presents the separate partial 2026 period. No spatial interpolation or IDW fallback was used for these maps. Fig. 5 uses a common ±500 mm color scale; values outside the limits are color-saturated but retained. Regional daily metrics and station-clustered intervals were calculated for the seven California NOAA/NCEI climate divisions after point-in-polygon station assignment (Vose et al., 2014). The paired network contained 15 North Coast, 22 Sacramento, 3 Northeast Interior, 24 Central Coast, 28 San Joaquin, 27 South Coast, and 13 Southeast Desert stations (Table S3). These counts refer to the 132-station paired dataset.

Supplementary spatial diagnostics used lead-specific paired annual station sums, distinct from the all-lead common-date map values. Moran's I used all available station-year sums, with a symmetrized six-nearest-neighbor matrix followed by row standardization and 999 permutations (Moran, 1950). The two-sided permutation comparison was centered on the randomization expectation $-1/(n-1)$, where n is the number of stations. Ordinary kriging additionally required 85% station-year completeness and used a spherical variogram in California Albers coordinates (EPSG:3310). PyKrige automatically fitted each variogram using its default six lag bins, unweighted soft-L1 fitting, no detrending, and no anisotropy adjustment; the variogram was refitted after withholding each station during cross-validation (Murphy, 2014; PyKrige developers, 2026). Diagnostics include squared Pearson correlation between withheld and predicted bias, RMSE, and the median and 95th-percentile conditional kriging standard errors (Fig. S4; Table S4). These standard errors do not incorporate all variogram, sampling, or representativeness uncertainty and are not verified prediction-interval coverage.

2.10. Rolling cumulative fixed-lead ETo forecast error

To determine whether relatively small daily ETo forecast errors accumulated over consecutive days and whether this accumulation increased with forecast lead, rolling cumulative fixed-lead ETo errors were evaluated over 3-, 7-, and 14-day windows. For each station and forecast lead, daily NBM-minus-CIMIS ETo errors were ordered chronologically by target date. A window was retained only when all target dates were consecutive; therefore, a window containing w daily records spanned w − 1 calendar days between its first and last target dates.

Using the daily errors defined in Eq. (3), the rolling cumulative error for station *i*, lead *L*, and window length *w* ending at target-date index *t* was

$$\mathrm{CE}_{i,L,w,t} = \sum_{k=t-w+1}^{t} e_{i,L,k} \quad (10)$$

where positive values indicate cumulative NBM overestimation of reference ETo and negative values indicate cumulative underestimation.

Each rolling window combined forecasts at the same lead across consecutive target dates. Because those forecasts originated from different initialization dates, the resulting values were treated as descriptive measures of accumulated fixed-lead forecast error rather than as a single forecast sequence from one initialization. Overlapping windows were retained and were not treated as statistically independent observations. For each window length and forecast lead, exceedance frequencies were calculated separately for cumulative errors greater than + 5, + 10, and + 15 mm and less than −5, −10, and −15 mm. Corresponding absolute-error exceedance frequencies were also calculated. These thresholds were selected as simple, evenly spaced diagnostic levels in water-depth units to summarize how often accumulated ETo forecast errors reached progressively larger magnitudes across forecast leads and window lengths. They were not intended to represent crop- or irrigation-system-specific decision thresholds. For each lead, window length, and threshold, the signed exceedance frequency was the number of qualifying windows divided by all valid windows, multiplied by 100; absolute exceedance equals positive plus negative exceedance before rounding. Frequencies were pooled over station-windows, with no binomial significance test based on independence.

2.11. Meteorological-driver diagnostics

Daily CIMIS meteorological records were obtained for maximum, minimum, and mean air temperature, mean relative humidity, vapor pressure, solar radiation, and mean wind speed. English-unit values were converted to the analysis units: °C = (°F−32) × 5/9, kPa = millibars× 0.1, m s^{-1} = miles h^{-1} ×0.44704, and MJ m^{-2} day^{-1} = langleys day^{-1} ×0.04184. Variable-specific I, M, and S flags were excluded. Solar-radiation records with QC flag R were additionally excluded; the audit found that all numeric daily solar values above 40 MJ m^{-2} day^{-1} carried this flag. This screening was confined to the driver analysis and did not alter the CIMIS ETo reference series or the primary ETo evaluation.

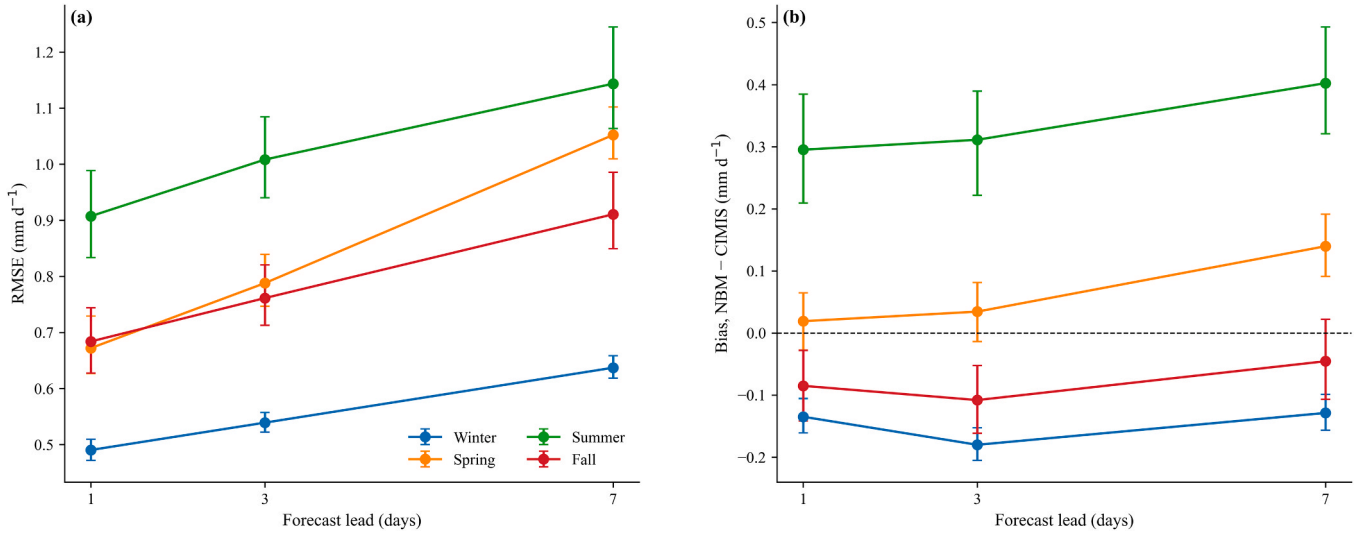


Fig. 4. Seasonal (a) RMSE and (b) mean bias by forecast lead. Colors distinguish winter, spring, summer, and fall; error bars show station-clustered 95% bootstrap intervals. The horizontal dashed line marks zero bias. Available 2026 records are included.

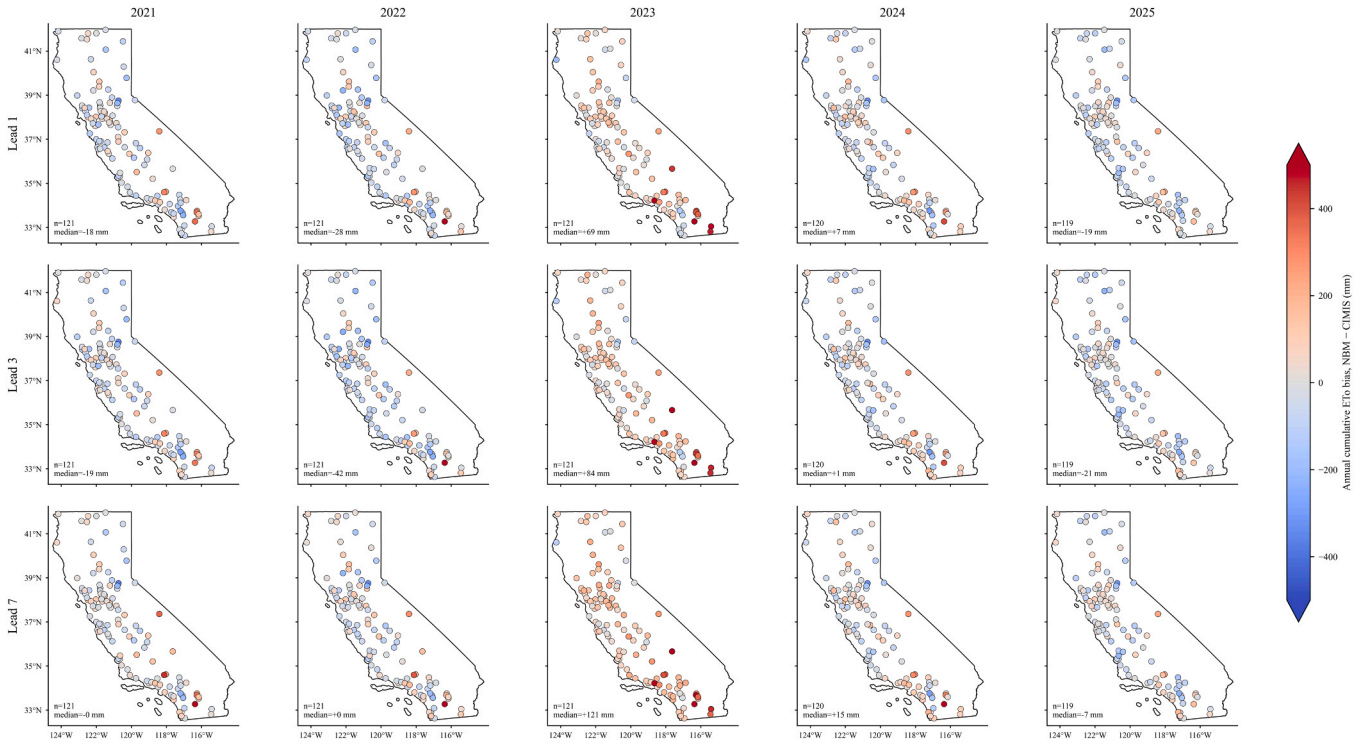


Fig. 5. Station-level cumulative NBM minus CIMIS ETo bias for 2021–2025. Columns show years and rows show forecast leads. Inclusion requires at least 85% expected active-day paired completeness for each lead; values sum common dates across all leads and CIMIS. Colors show calculated station bias without spatial interpolation. All panels use ±500 mm limits; more extreme values are color-saturated but retained. Printed station medians are rounded to whole millimeters; Table S2 gives two-decimal values.

NBM meteorological values were taken from the daily calculation inputs, with wind converted from 10 to 2 m. The radiation comparison used the boundary-reconstructed trapezoidal series where available and the primary radiation integral otherwise. Driver errors were NBM minus CIMIS. Bias, MAE, RMSE, R^2, and Pearson correlations with ETo error were calculated on variable-specific matched records. Multiple linear regressions used complete cases and standardized both daily ETo errors and the errors in mean temperature, mean relative humidity, 2-m wind speed, and solar radiation. Models were fitted for all dates and all summer dates; additional 2023 and summer-2023 fits were used as year-specific diagnostics for the complete year with the largest forecast errors. These diagnostic models were fitted and assessed on the same data; their coefficients and explained variance describe associations and are not independent causal contributions or out-of-sample correction performance.

2.12. Computational implementation and reproducibility

Data acquisition, processing, statistical analysis, and visualization were performed in Python 3.10.6. NBM forecast discovery and subsetting were conducted using Herbie/FastHerbie (herbie-data 2025.12.0) (Blaylock, 2024). GRIB2 data were decoded and organized using cfgrib 0.9.15.1, ecCodes 2.47.0, and xarray 2025.6.1. Numerical and tabular processing was performed with NumPy 2.2.6 and pandas 2.3.3, while SciPy 1.15.3 supported statistical calculations and spatial nearest-neighbor operations. Geospatial processing used PyProj 3.7.1, GeoPandas 1.1.3, and Shapely 2.1.2. PyKrige 1.7.3 was used for the supplementary ordinary-kriging diagnostics, and Matplotlib 3.10.7 was used for figure generation (Blaylock, 2024; Hoyer and Hamman, 2017; Murphy, 2014). Requests 2.34.2 supported programmatic retrieval of CIMIS meteorological records and ancillary web-service data. The processed daily datasets, analysis scripts, and derived analysis tables were retained, allowing the statistical results and manuscript figures to be regenerated without repeating the download and processing of the full NBM GRIB2 forecast archive.

3. Results

3.1. Daily performance, uncertainty, and benchmark-relative skill

Unless otherwise stated, pooled statistics reported below include the partial 2026 record through 11 June 2026; complete-year annual cumulative comparisons are restricted to 2021–2025. Across the full record, NBM-derived daily ETo agreed strongly with CIMIS ETo, while error increased with lead time (Fig. 2; Table 1). Pooled RMSE was 0.70, 0.79, and 0.95 mm day^{-1} at leads 1, 3, and 7; R^2 was 0.91, 0.88, and 0.83; and MAE was 0.51, 0.58, and 0.70 mm day^{-1} . Lead-1 NSE was 0.89. Station-clustered 95% RMSE intervals were 0.66–0.75, 0.75–0.83, and 0.91–1.01 mm day^{-1} . Bias was + 0.02, + 0.01, and + 0.09 mm day^{-1} , with intervals spanning zero at leads 1 and 3 and remaining positive at lead 7 (+0.04 to +0.14 mm day^{-1}). Values throughout the text and tables were rounded directly from the full-precision estimates.

The degradation in absolute error with lead time was statistically distinguishable on common station-date support. Relative to lead 1, RMSE increased by 0.09 mm day^{-1} at lead 3 (95% CI, 0.08–0.10 mm day^{-1}) and by 0.26 mm day^{-1} at lead 7 (95% CI, 0.24–0.27 mm day^{-1}). The lead-3 to lead-7 RMSE increase was 0.17 mm day^{-1} (95% CI, 0.15–0.19 mm day^{-1}). MAE showed the same ordering, with pairwise increases of 0.07, 0.19, and 0.12 mm day^{-1} for lead 1 to 3, lead 1 to 7, and lead 3 to 7, respectively; all 95% confidence intervals excluded zero. In contrast, the small bias difference between leads 1 and 3 was not distinguishable from zero. Station-equal-weight summaries gave the same performance ranking: median station RMSE increased from 0.62 mm day^{-1} at lead 1 to 0.73 mm day^{-1} at lead 3 and 0.89 mm day^{-1} at lead 7, indicating that the pooled result was not caused by stations with longer records receiving greater weight.

Table 1
Daily performance by year and forecast lead. Confidence intervals are station-clustered 95% bootstrap intervals. Skill scores use RMSE and matched benchmark support.

Year	Lead (d)	n	Bias	MAE	RMSE	R^2	NSE	Slope	RMSE 95% CI	Skill vs persistence	Skill vs climatology
2021	1	43,645	-0.02	0.50	0.67	0.92	0.91	1.03	0.63–0.71	0.38	0.34
2021	3	43,526	-0.03	0.58	0.77	0.89	0.88	1.00	0.73–0.81	0.38	0.25
2021	7	43,286	+0.06	0.69	0.93	0.84	0.82	0.97	0.88–0.99	0.30	0.09
2022	1	43,864	-0.05	0.50	0.69	0.92	0.90	1.02	0.62–0.78	0.35	0.33
2022	3	43,745	-0.10	0.56	0.77	0.89	0.88	1.00	0.71–0.85	0.37	0.26
2022	7	43,745	+0.02	0.66	0.91	0.84	0.83	0.97	0.83–1.02	0.30	0.11
2023	1	43,898	+0.25	0.57	0.85	0.90	0.83	1.10	0.75–0.94	0.24	0.24
2023	3	43,898	+0.30	0.65	0.95	0.87	0.79	1.10	0.87–1.04	0.24	0.15
2023	7	43,898	+0.41	0.79	1.12	0.82	0.71	1.07	1.03–1.23	0.17	0.00
2024	1	43,768	+0.03	0.51	0.70	0.91	0.90	1.03	0.66–0.74	0.34	0.28
2024	3	43,768	+0.02	0.56	0.76	0.89	0.88	1.01	0.73–0.80	0.35	0.21
2024	7	43,768	+0.06	0.66	0.91	0.85	0.83	0.97	0.87–0.95	0.25	0.07
2025	1	43,331	-0.05	0.47	0.61	0.92	0.92	1.00	0.58–0.64	0.43	0.38
2025	3	43,331	-0.07	0.54	0.70	0.90	0.89	0.98	0.67–0.72	0.43	0.29
2025	7	43,331	-0.02	0.67	0.88	0.84	0.83	0.94	0.86–0.91	0.31	0.10
2026*	1	19,787	-0.09	0.47	0.60	0.91	0.90	1.00	0.58–0.63	0.50	0.46
2026*	3	19,787	-0.12	0.56	0.73	0.87	0.86	0.97	0.71–0.74	0.48	0.35
2026*	7	19,665	-0.09	0.71	0.96	0.78	0.75	0.90	0.93–0.98	0.36	0.15

* 2026 covers 1 January–11 June only. Bias, MAE, RMSE, and RMSE intervals are in mm day^{-1} . Positive skill indicates lower error than the benchmark. Climatology excludes the target year but may use later years; persistence uses the latest completed daily record at initialization. Metrics and skill can have different sample support when a benchmark is missing.

Interannual performance was generally consistent with the pooled pattern (Table 1). Lead 1 produced the lowest RMSE and lead 7 the highest in every complete year. Complete-year RMSE varied among years; the highest values occurred in 2023 (0.85, 0.95, and 1.12 mm day^{-1} at leads 1, 3, and 7), whereas in the other complete years lead-1 RMSE ranged from 0.61 to 0.70 mm day^{-1} and lead-7 RMSE ranged from 0.88 to 0.94 mm day^{-1} . The 2026 results represent only 1 January–11 June and were therefore interpreted as partial-year daily performance rather than complete annual performance; over this available period, RMSE was 0.60, 0.73, and 0.96 mm day^{-1} at leads 1, 3, and 7. The 2026 records were retained in pooled daily, seasonal, station-level, benchmark, anomaly, and rolling-error analyses, but were not combined with 2021–2025 when comparing complete-year annual cumulative totals or the main annual spatial-bias maps.

Benchmark-relative evaluation showed that NBM-derived ETo provided positive forecast skill beyond simple reference forecasts (Fig. 8a; Table 1). Against a prospective persistence benchmark based on the most recently completed CIMIS day available at forecast initialization, pooled RMSE skill scores were 0.36, 0.36, and 0.27 at leads 1, 3, and 7, respectively. Station-clustered 95% confidence intervals were 0.31–0.40, 0.32–0.40, and 0.23–0.31, remaining above zero at all three leads. Relative to leave-one-year-out station-month climatology, pooled RMSE skill decreased from 0.32 at lead 1 to 0.24 at lead 3 and 0.08 at lead 7; the corresponding 95% intervals were 0.28–0.36, 0.20–0.27, and 0.02–0.12. Year-specific skill varied across the record, but the pooled comparison remained positive at all three leads. Thus, NBM-derived ETo added information relative to both benchmarks overall, while the margin over climatology became small at the longest lead.

3.2. Common-support cumulative ETo and interannual bias

Over the complete years, common-support cumulative trajectories showed clear interannual variation (Fig. 3; Table S2). Differences between the medians of station-specific NBM and CIMIS period totals were within approximately ±31 mm in 2021, 2022, 2024, and 2025, whereas 2023 showed a persistent positive departure during the high-ETo season. In that year, the median CIMIS station total was 1276.35 mm, and median NBM totals exceeded this value by 97.90, 111.89, and 161.15 mm at leads 1, 3, and 7. These are differences between medians of station totals calculated across the common-support station set; they are not asserted to be exact differences at the endpoints of the plotted daily median curves.

Because the high-bias year yields several related cumulative summaries, the corresponding statistics are distinguished explicitly. Using paired differences at the 123 stations with common dates in 2023, median cumulative bias was + 70.69, + 84.53, and + 123.81 mm at leads 1, 3, and 7; corresponding means were + 89.57, + 106.23, and + 146.73 mm. Applying the map-completeness criterion retained 121 stations and gave medians of + 68.93, + 83.84, and + 121.39 mm and means of + 88.29, + 105.10, and + 145.29 mm (Fig. 5; Table S2). The different values arise from the statistic and station subset being summarized. In particular, the lead-7 mean of + 146.73 mm for all common-support stations and + 145.29 mm for map-eligible stations should not be interchanged with either median or with the + 161.15 mm difference of median totals.

The 2026 panel in Fig. 3 covers only 1 January–11 June. Across station-specific totals over available common dates, median CIMIS ETo was 554.99 mm; the differences between median NBM and CIMIS totals were −2.62, −7.85, and −3.12 mm at leads 1, 3, and 7. These partial-period values were not treated as annual totals or combined with the complete-year spatial maps.

3.3. Seasonal performance, anomaly agreement, and interannual robustness

Forecast error varied with season and lead (Fig. 4). Winter RMSE was 0.49, 0.54, and 0.64 mm day^{-1} at leads 1, 3, and 7, whereas summer RMSE was 0.91, 1.01, and 1.14 mm day^{-1} . Spring and fall were intermediate, with lead-7 RMSE of 1.05 and 0.91 mm day^{-1} . Summer lead-7 RMSE had a station-clustered 95% interval of 1.06–1.24 mm day^{-1} , compared with 0.62–0.66 mm day^{-1} in winter.

Winter bias was negative at all leads (−0.13, −0.18, and −0.13 mm day^{-1}), whereas summer bias was positive (+0.30, +0.31, and +0.40 mm day^{-1}). Spring bias increased from + 0.02 to + 0.14 mm day^{-1} across leads 1–7; fall remained slightly negative. Thus, the largest absolute errors occurred during high atmospheric demand, although these absolute-error comparisons do not by themselves establish a larger relative error in summer.

Removing station-specific monthly means reduced R^2 from the raw pooled values to 0.67, 0.56, and 0.33 at leads 1, 3, and 7, respectively (Fig. 8b). Anomaly RMSE was 0.58, 0.67, and 0.84 mm day^{-1} . The common seasonal cycle therefore contributed to the high raw-value

correlation, while shorter leads retained stronger agreement in daily departures from the respective monthly means. Because both series were centered using the evaluation record, these values are diagnostic anomaly agreement rather than prospective anomaly-forecast skill. As an interannual robustness test, excluding the highest-error complete year (2023) reduced RMSE from 0.70 to 0.66 mm day^{-1} at lead 1, from 0.79 to 0.75 mm day^{-1} at lead 3, and from 0.95 to 0.91 mm day^{-1} at lead 7. Mean bias outside that year was −0.03, −0.05, and + 0.02 mm day^{-1} . The ordering of lead 1 < lead 3 < lead 7 persisted, with R^2 of 0.92, 0.89, and 0.84. The pooled lead-time pattern was therefore not dependent on a single high-error year.

3.4. Station-level, regional, and spatial variation in forecast error

The annual station-bias maps show marked interannual and local variability (Fig. 5). After requiring at least 85% completeness for each lead, 121 stations were retained in each of 2021–2023, 120 in 2024, and 119 in 2025.

Over the complete years, mapped median cumulative bias ranged from −19.48 to + 68.93 mm at lead 1, from −42.18 to + 83.84 mm at lead 3, and from −6.69 to + 121.39 mm at lead 7 (Table S2). The strongest coherent positive pattern occurred in 2023, while the other complete years contained both positive and negative station biases centered closer to zero. Values are sums over common available dates, with no interpolation between stations. The separate 2026 partial-period maps are shown in Fig. S1.

Station-level RMSE distributions shifted upward with lead time (Fig. 6). Across 132 stations, median RMSE was 0.62 mm day^{-1} at lead 1 (IQR 0.55–0.69), 0.73 mm day^{-1} at lead 3 (IQR 0.65–0.80), and 0.89 mm day^{-1} at lead 7 (IQR 0.82–0.99). Equal weight mean station RMSE was 0.67, 0.77, and 0.93 mm day^{-1} , preserving the ranking. Maximum station RMSE was 1.79, 1.80, and 2.46 mm day^{-1} . All station values, including these extremes, were retained.

Performance also differed among California climate divisions (Fig. 8c; Table S3). The lowest regional RMSE occurred in the Central Coast, where RMSE increased from 0.57 to 0.67 and 0.80 mm day^{-1} across leads 1, 3, and 7. North Coast, Sacramento, Northeast Interior, San Joaquin, and South Coast showed the same lead-time degradation, with lead-7 RMSE ranging from approximately 0.83 to 0.98 mm day^{-1} . The Southeast Desert had the largest errors at every lead, with RMSE of 1.07, 1.12, and 1.33 mm day^{-1} . Regional estimates for sparsely represented divisions, particularly Northeast Interior, should be interpreted cautiously because only three CIMIS stations were available there. Annual station bias showed modest positive spatial autocorrelation, with Moran's I approximately 0.08–0.36 (Fig. S4). Nevertheless, leave-one-station-out kriging agreement was weak across the complete-year analysis. Squared Pearson correlation was close to zero in most year-lead combinations and did not exceed 0.35; cross-validation RMSE ranged from approximately 85 to 138 mm (Table S4). Conditional kriging standard errors were also large relative to the annual bias magnitudes. Spatial autocorrelation therefore did not establish reliable interpolation, supporting presentation of calculated station values in the main maps.

3.5. Rolling cumulative fixed-lead forecast error

Accumulating daily errors across consecutive target dates increased the frequency of management-scale ETo discrepancies, particularly for

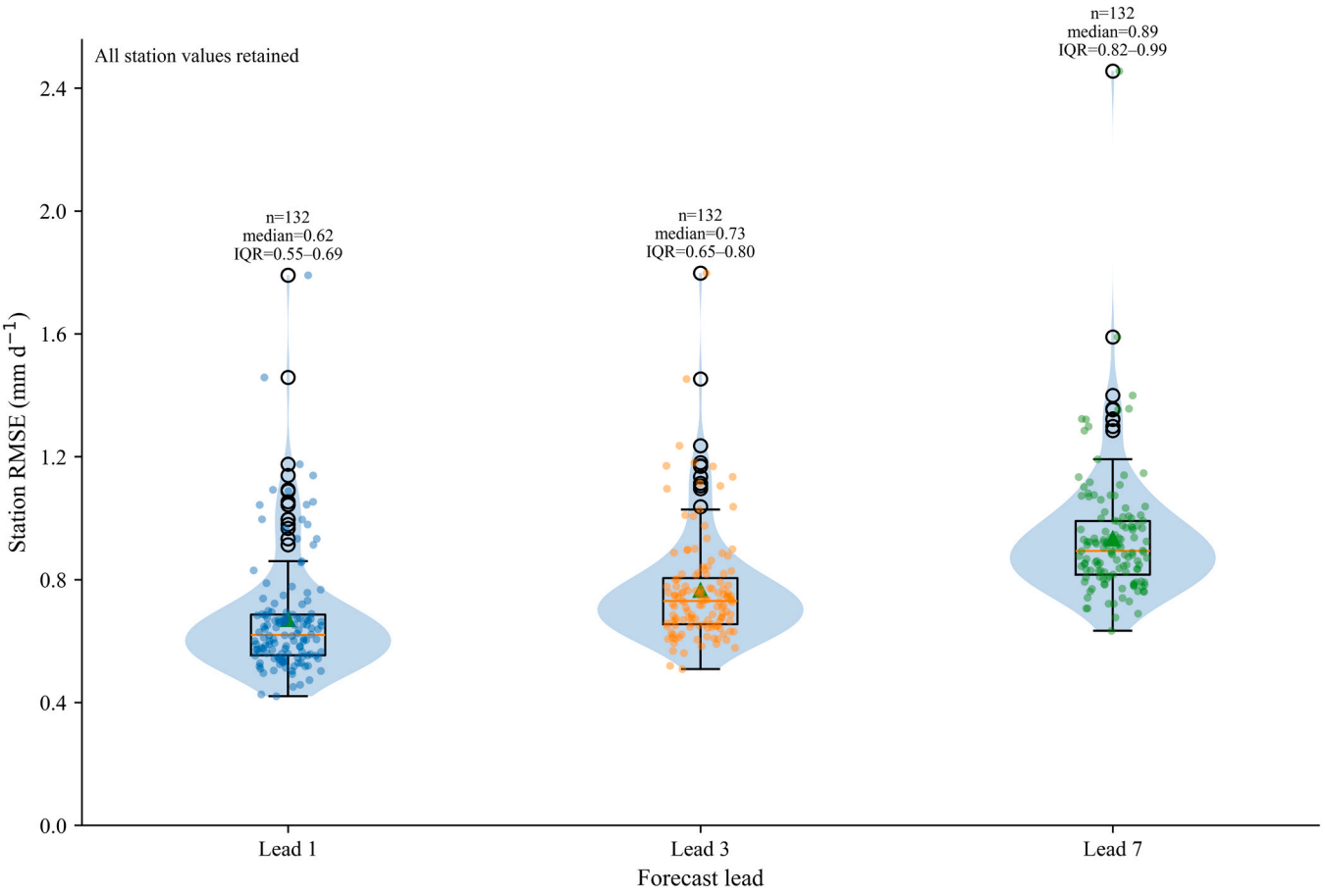


Fig. 6. Distribution of station-level RMSE at each forecast lead, including the partial 2026 record. Violin densities, boxplots, and individual values summarize all 132 stations. No station or outlier was removed. Annotations give the station count, median, and interquartile range.

longer windows and longer forecast leads (Fig. 7). These calculations are rolling cumulative fixed-lead errors: each window sums errors from the same lead across consecutive target dates and therefore does not represent a multi-day forecast sequence issued from one initialization. For 3-day windows, the absolute exceedance frequency above 5 mm was 1.89%, 2.54%, and 4.35% for leads 1, 3, and 7, respectively. Exceedance of 10 or 15 mm was uncommon over 3 days, remaining below 0.32% and 0.08%, respectively.

Exceedance became more frequent as the accumulation window increased. For 7-day windows, absolute exceedance of 5 mm increased to 12.76%, 16.54%, and 21.95% for leads 1, 3, and 7. For 14-day windows, the corresponding frequencies were 36.18%, 42.37%, and 46.36%. At the 10-mm threshold over 14 days, exceedance occurred in 10.88%, 13.56%, and 17.74% of windows, while the 15-mm threshold was exceeded in 4.42%, 5.25%, and 7.48%.

Separating the sign of accumulated error showed that overestimation and underestimation were not symmetric. At the 5-mm threshold for 14-day windows, overestimation frequencies were 18.39%, 20.37%, and 26.55% for leads 1, 3, and 7, compared with underestimation frequencies of 17.79%, 21.99%, and 19.81%. At larger thresholds, positive accumulated errors were more dominant; for example, at 15 mm over 14 days, overestimation occurred in 3.84%, 4.47%, and 6.52% of windows, whereas underestimation occurred in only 0.58%, 0.78%, and 0.96%. Because successive rolling windows overlap, these frequencies are descriptive exceedance rates rather than independent event probabilities, and they quantify accumulated ETo forecast error rather than crop-specific irrigation risk.

3.6. *Meteorological-error associations and robustness of the processing choices*

Solar-radiation error had the largest absolute standardized regression coefficient at all leads (Fig. 8d): 0.49, 0.51, and 0.55 for leads 1, 3, and 7. Wind coefficients were 0.38, 0.35, and 0.29; relative-humidity coefficients were −0.25, −0.29, and −0.31; and mean-temperature coefficients were 0.07, 0.11, and 0.16. Wind ranked second in absolute magnitude at leads 1 and 3, while relative humidity ranked second at lead 7. Complete-case sample sizes were 214,337, 214,157, and 213,876. The fitted models explained 52.16%, 58.45%, and 67.77% of ETo-error variance in these samples. These are in-sample statistical associations, not independent causal contributions.

A year-specific diagnostic for the highest-error complete year (2023) showed solar-radiation biases after R-flag exclusion of + 1.30, + 1.03, and + 1.31 MJ m^{-2} day^{-1} at leads 1, 3, and 7; 2-m-equivalent wind biases were + 0.72, + 0.84, and + 0.98 m s^{-1}. Mean-temperature biases were + 0.37, + 0.35, and + 0.33 °C, and relative-humidity biases were −0.07, −0.30, and −0.65 percentage points. Solar radiation had the largest standardized coefficient in those annual regressions. Across all summer dates, wind ranked first at lead 1 and radiation at leads 3 and 7. These diagnostics support a physical interpretation involving radiation, aerodynamic demand, and humidity without attributing the interannual anomaly to untested hydrologic events.

The sensitivity analyses preserved the lead-time ordering (Figs. S2–S3). Lead-1 RMSE was 0.70 mm day^{-1} for both hourly and 3-hourly meteorological sampling when rounded to two decimals; the increase was less than 0.01 mm day^{-1}, and the direct difference between the resulting ETo series had RMSE 0.06 mm day^{-1}. Fixed-PST boundary reconstruction and trapezoidal radiation integration changed lead-3 and lead-7 forecast RMSE by less than 0.01 mm day^{-1} on their matched sensitivity subsets. Four-neighbor weighted extraction differed from nearest-grid ETo by 0.05–0.06 mm day^{-1} RMSE over the sampled dates. Replacing station elevation in the ETo pressure and clear-sky terms with grid-center 3DEP elevation changed forecast RMSE by less than 0.01 mm day^{-1}. The annual-bias pattern was also stable across 80–95% completeness thresholds, and exclusion of I, M, and S flags changed no primary CIMIS ETo records. These checks constrain the tested processing effects; the limited-date spatial test and the lead-1 test with radiation held complete do not establish insensitivity to every possible sampling or downscaling method.

4. Discussion

NBM-derived daily ETo captured much of the variability in CIMIS station-based ETo across California, with pooled RMSE increasing from 0.70 mm day^{-1} at lead 1 to 0.95 mm day^{-1} at lead 7 and R^2 declining from 0.91 to 0.83. Lead-dependent deterioration is physically plausible because uncertainty in the evolution of temperature, humidity, wind, and cloud-controlled radiation propagates through the Penman–Monteith calculation. Similar patterns have been reported in Australia and the United States (Perera et al., 2014; Tian and Martinez, 2014; Medina et al., 2018). The positive skill scores show improvement over the tested persistence and monthly-climatology benchmarks, although the margin over climatology was only 0.08 at lead 7. Common-support paired comparisons and equal-weight station summaries preserved the lead ranking. Temporal and radiation sensitivities were small relative to the lead differences, but the lead-1 sampling test

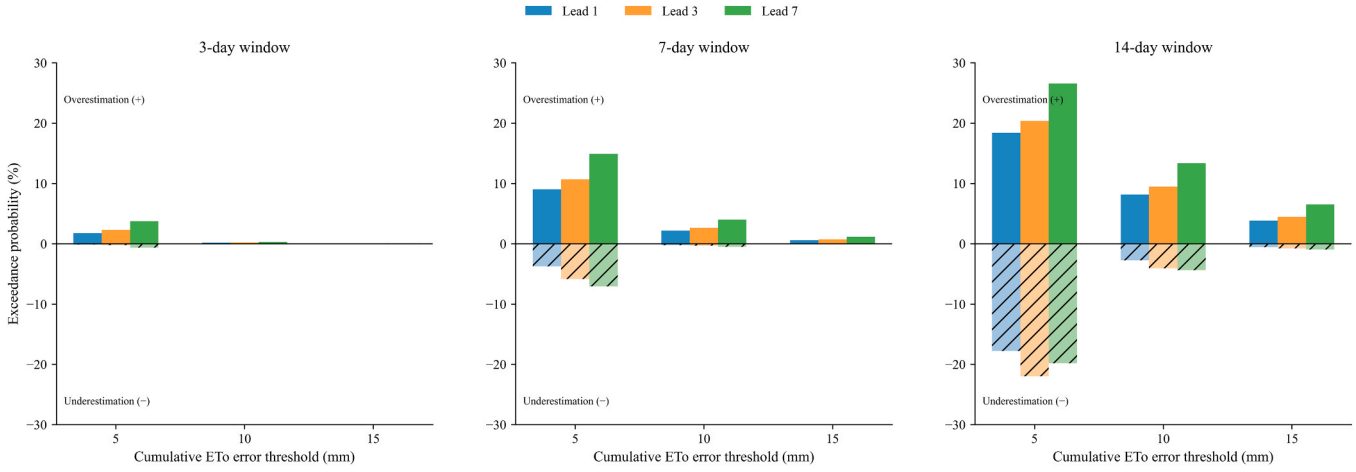


Fig. 7. Signed exceedance frequencies of rolling cumulative fixed-lead ETo error for 3-, 7-, and 14-day windows. Positive bars denote overestimation beyond the indicated threshold; negative, hatched bars display underestimation frequencies by magnitude below zero. All panels use the same symmetric scale. Frequencies are descriptive fractions of valid overlapping windows, not independent-event or crop-risk probabilities. Each window combines a fixed lead over consecutive target dates from different initializations.

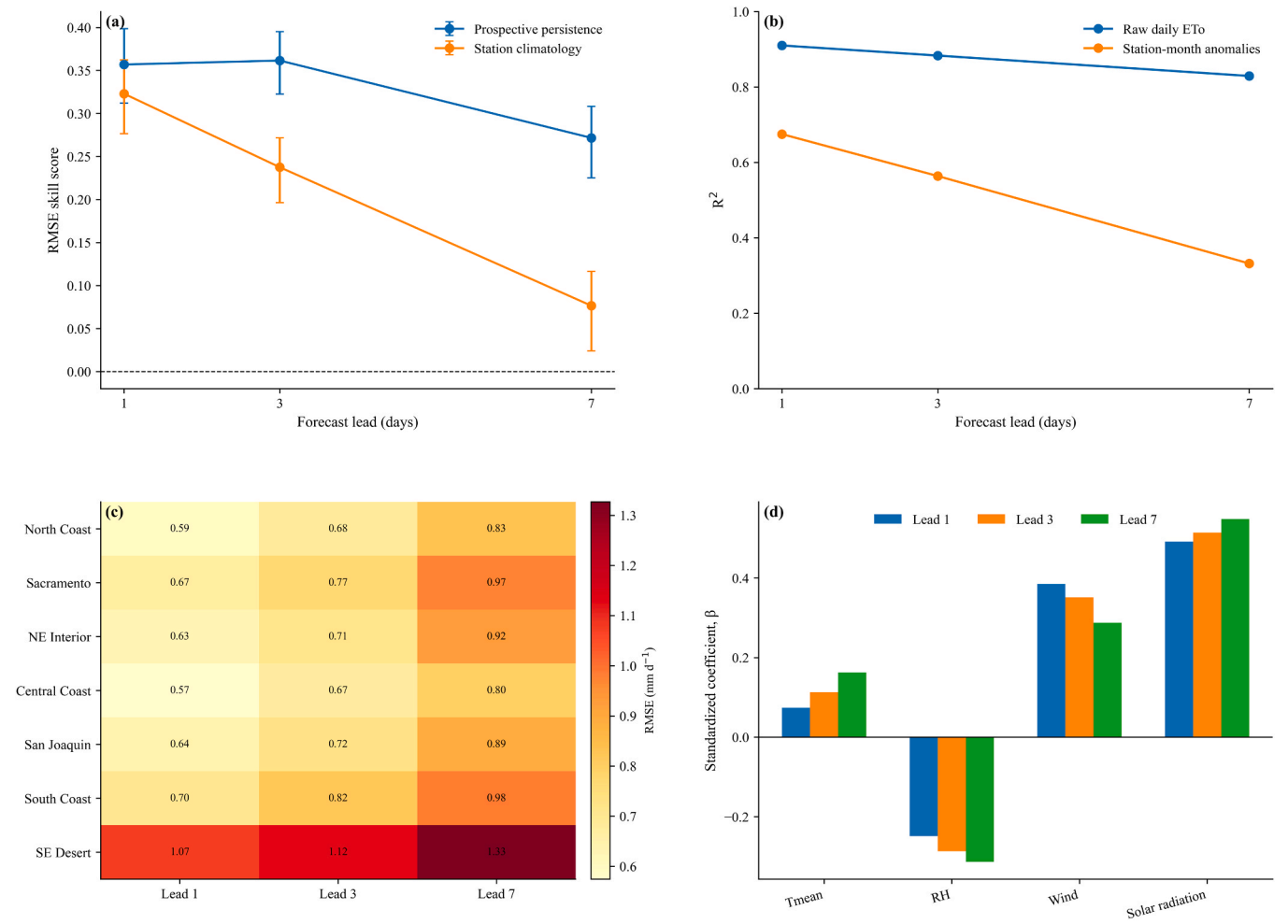


Fig. 8. Benchmark, anomaly, regional, and meteorological-error diagnostics. (a) RMSE skill relative to prospective persistence and leave-one-year-out monthly climatology, with station-clustered 95% intervals. (b) Raw R^2 and R^2 after centering each series by its station-month mean. (c) Regional RMSE by California climate division. (d) Standardized regression coefficients for meteorological errors, excluding CIMIS solar-radiation records flagged R. Coefficients describe in-sample associations. Available 2026 records are included.

retained a complete radiation integral, and the spatial extraction test covered representative dates rather than every day.

Daily and cumulative evaluations describe complementary aspects of forecast performance. Small systematic daily biases can accumulate into substantial seasonal differences, as shown by the interannual cumulative results. Common-date matching prevents one forecast from accumulating over more dates than another but does not restore missing-day water demand. Cumulative totals must therefore be interpreted with their contributing dates and station sets. Likewise, subtracting monthly means reduced R^2 to 0.67, 0.56, and 0.33 across leads, showing that part of the raw correlation arises from the seasonal cycle. Remaining anomaly agreement is informative, but removing each series' in-sample monthly mean also removes its mean seasonal bias and is not equivalent to validating an operational anomaly forecast.

Seasonal differences were pronounced and were consistent with previous evaluations of forecast ETo in the western United States. Winter RMSE ranged from 0.49 to 0.64 mm day^{-1} across leads, whereas summer RMSE increased from 0.91 mm day^{-1} at lead 1 to 1.14 mm day^{-1} at lead 7. Summer bias was also consistently positive, increasing from + 0.30 to + 0.40 mm day^{-1} . Ben Hamouda et al. (2022) reported generally strong agreement between forecast reference ETo and CIMIS at selected California stations, while McEvoy et al. (2022) found larger warm-season errors and reduced summer forecast performance in Nevada. The larger summer errors observed here are therefore consistent with the greater difficulty of forecasting ETo during periods of high atmospheric demand, when radiation, vapor-pressure conditions, wind, and temperature are all important contributors to the Penman-Monteith calculation. The meteorological-driver analysis supports this interpretation: solar-radiation error had the largest standardized association at all leads, while wind and relative humidity also contributed substantial associations and their relative ordering varied with lead. These relationships are statistical associations rather than independent causal effects, but they provide a physical basis for the observed warm-season degradation.

Interannual variability was also substantial. The highest-error complete year (2023) showed a pronounced positive cumulative bias, and the year-specific diagnostics indicated positive shortwave-radiation and wind-speed errors, with smaller mean-temperature and relative-humidity departures. These associations provide physically consistent context for the elevated ETo bias but do not uniquely explain the anomaly or isolate model-version effects. Excluding this high-error year reduced overall error but preserved the ordering lead 1 < lead 3 < lead 7, demonstrating that the statewide lead-time conclusion was robust to this interannual anomaly.

Performance varied across stations and climate divisions. The Central Coast had the lowest regional RMSE, whereas the Southeast Desert reached 1.07, 1.12, and 1.33 mm day^{-1} across leads. Station placement, local advection, elevation, coastal influences, and desert conditions can affect the correspondence between gridded guidance and a reference station (Ben Hamouda et al., 2022; McEvoy et al., 2022). The network

does not sample California uniformly: only three paired stations represent the Northeast Interior division, so its regional estimates and bootstrap intervals warrant particular caution. Weak kriging cross-validation despite positive spatial autocorrelation further limits claims about unmonitored locations. The California results provide a geographically diverse evaluation but should not be transferred quantitatively to other regions without local verification.

The rolling cumulative analysis provides a useful water-depth interpretation of forecast error, but it should be interpreted as cumulative fixed-lead error rather than as a direct measure of irrigation scheduling performance. Over 3-day windows, only 1.89–4.35% of cases exceeded 5 mm absolute cumulative ETo error across the three leads, whereas the corresponding frequency increased to 36.18–46.36% over 14-day windows. At the 14-day, 10-mm threshold, exceedance ranged from 10.88% to 17.74%. Separating the sign of accumulated error further showed that positive and negative departures were not always symmetric, particularly at the larger thresholds. These results demonstrate how modest daily forecast differences can accumulate over longer periods, but each window sums errors from the same forecast lead across consecutive target dates. It therefore does not represent a multi-day forecast sequence issued from a single initialization. In addition, the 5-, 10-, and 15-mm thresholds are reference-ETo error thresholds expressed in water-depth units; they are not crop-specific irrigation thresholds and should not be interpreted directly as irrigation-application error.

The results should be considered in the context of the study design and data characteristics. CIMIS ETo is calculated from station meteorological observations and therefore remains subject to measurement and representativeness uncertainty (Allen et al., 1998, 2005; Hart et al., 2009). The CIMIS network is not spatially uniform across California, which limits inference in sparsely represented regions and at unmonitored locations. The study evaluated the evolving operational NBM rather than a homogeneous reforecast system; therefore, interannual differences cannot be attributed solely to specific NBM version changes. Station-cluster bootstrapping preserves the temporal sequence within stations but does not fully account for dependence among stations during shared weather events. The climatology benchmark was evaluated retrospectively, while the persistence benchmark assumed availability of the most recently completed CIMIS day without reconstructing historical data-release latency. Finally, the 2026 record extends only through 11 June and therefore has incomplete seasonal coverage relative to the complete calendar years.

Reference ETo describes atmospheric demand over a standardized surface. Crop evapotranspiration depends on crop coefficients and, where relevant, crop stress; net irrigation requirement additionally depends on effective precipitation, root-zone water storage, and other water-balance terms; applied depth further depends on irrigation efficiency and management (Allen et al., 1998; Peddinti and Kambhammettu, 2019; Peddinti et al., 2025). Under a constant crop coefficient, an ETo error contributes that coefficient times the error to estimated crop ET, but there is no universal conversion to applied irrigation. Lead 1 had the lowest error and is best supported for near-term atmospheric-demand guidance; lead 3 offers earlier notice with increased error and lead 7 provides a longer outlook with limited improvement over climatology. No optimal management trade-off was established. Future evaluation should compare operational products directly, verify multi-lead sequences from a common initialization, and test local bias correction, probabilistic processing, and crop-specific water-balance decisions (McEvoy et al., 2022; Medina and Tian, 2020).

5. Conclusions

Across 132 CIMIS stations during 1 January 2021–11 June 2026, NBM-derived ETo reproduced broad daily and seasonal variability, but daily absolute error increased with forecast lead. Pooled RMSE was 0.70, 0.79, and 0.95 mm day^{-1} at leads 1, 3, and 7, and the differences were supported by paired station-clustered intervals. Positive skill relative to persistence and cross-validated monthly climatology demonstrated added information for reference ETo, with the smallest margin over climatology at lead 7. Summer and Southeast Desert errors were larger, while cumulative bias showed substantial interannual variability. The lead-time ordering persisted when the highest-error complete year was excluded. Rolling cumulative fixed-lead errors increased with accumulation-window length, showing that modest daily forecast errors can become more consequential over longer periods. Reference ETo represents atmospheric demand and should be combined with crop- and field-specific information when used for irrigation planning. NBM-derived ETo therefore provides useful short-lead guidance on future atmospheric water demand, with stronger performance at shorter forecast leads, and can support local irrigation planning and water-management applications when appropriately evaluated for local conditions.

CRediT authorship contribution statement

Srinivasa Rao Peddinti: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.agwat.2026.110857.

Data availability

Data will be made available on request.

References

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33, 121–131.
Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper, 56. Food and Agriculture Organization of the United Nations, Rome.
Allen, R.G., Walter, I.A., Elliott, R.L., Howell, T.A., Itenfisu, D., Jensen, M.E., Snyder, R. L., 2005. The ASCE Standardized Reference Evapotranspiration Equation. American Society of Civil Engineers, Reston, VA. https://doi.org/10.1061/9780784408056.
Ben Hamouda, G., Zaccaria, D., Bali, K., Snyder, R.L., Ventura, F., 2022. Evaluation of Forecast Reference Evapotranspiration for Different Microclimate Regions in California to Enable Prospective Irrigation Scheduling. J. Irrig. Drain. Eng. 148, 04021061. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001632.
Blankenau, P.A., Kilic, A., Allen, R., 2020. An evaluation of gridded weather data sets for the purpose of estimating reference evapotranspiration in the United States. Agric. Water Manag. 242, 106376.
Blaylock, B.K., 2024. Herbie: Retrieve Numerical Weather Prediction Model Data [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4567540.
Cai, J., Liu, Y., Lei, T., Pereira, L.S., 2007. Estimating reference evapotranspiration with the FAO Penman–Monteith equation using daily weather forecast messages. Agric. For. Meteorol. 145, 22–35.
CIMIS, 2026. CIMIS - California Irrigation Management Information System [WWW Document]. URL https://cimis.water.ca.gov/ (accessed 6.23.26).
DWR, 2023. Statewide Crop Mapping - California Natural Resources Agency Open Data [WWW Document]. URL https://data.cnra.ca.gov/dataset/statewide-crop-mapping (accessed 7.4.26).
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Hamill, T.M., Engle, E., Myrick, D., Peroutka, M., Finan, C., Scheuerer, M., 2017. The US National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. Mon. Weather. Rev. 145, 3441–3463.

Hart, Q.J., Brugnach, M., Temesgen, B., Rueda, C., Ustin, S.L., Frame, K., 2009. Daily reference evapotranspiration for California using satellite imagery and weather station measurement interpolation. Civil. Eng. Environ. Syst. 26, 19–33. https://doi.org/10.1080/10286600802003500.

Hoyer, S., Hamman, J., 2017. xarray: N-D labeled arrays and datasets in Python. J. Open. Res. Softw. 5, 10. https://doi.org/10.5334/jors.148.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018.

Liu, B., Liu, M., Cui, Y., Shao, D., Mao, Z., Zhang, L., Khan, S., Luo, Y., 2020. Assessing forecasting performance of daily reference evapotranspiration using public weather forecast and numerical weather prediction. J. Hydrol. 590, 125547.

McEvoy, D.J., Roj, S., Dunkerly, C., McGraw, D., Huntington, J.L., Hobbins, M.T., Ott, T., 2022. Validation and Bias Correction of Forecast Reference Evapotranspiration for Agricultural Applications in Nevada. J. Water Resour. Plan. Manag. 148, 04022057. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001595.

Medina, H., Tian, D., 2020. Comparison of probabilistic post-processing approaches for improving numerical weather prediction-based daily and weekly reference evapotranspiration forecasts. Hydrol. Earth Syst. Sci. 24, 1011–1030.

Medina, H., Tian, D., Srivastava, P., Pelosi, A., Chirico, G.B., 2018. Medium-range reference evapotranspiration forecasts for the contiguous United States based on multi-model numerical weather predictions. J. Hydrol. 562, 502–517.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

Murphy, B.S., 2014. PyKrige: development of a kriging toolkit for Python. AGU Fall Meet. Abstr. pp. H51K-0753.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10, 282–290.

NCEP, 2026. National Blend of Models (NBM) Products [WWW Document]. National Centers for Environmental Prediction. URL https://www.nco.ncep.noaa.gov/pmb/products/blend/ (accessed 9.11.26).

NOAA, 2026. National Blend of Models. Meteorological Development Laboratory (accessed 13 September 2026). https://vlab.noaa.gov/web/mdl/nbm.

NOAA/MDL, 2025. National Blend of Models (NBM) Upgraded to Version 4.3 [WWW Document]. Meteorological Development Laboratory. URL https://vlab.noaa.gov/web/mdl/-/nbm-upgraded-to-version-4.3 (accessed 9.11.26).

NOAA/MDL, 2026a. NBM Versions. Meteorological Development Laboratory. https://vlab.noaa.gov/web/mdl/nbm-versions (accessed 13 September 2026).

NOAA/MDL, 2026b. National Blend of Models operational upgrade to version 5.0. Meteorological Development Laboratory news and announcements, 5 May 2026. https://vlab.noaa.gov/web/mdl/events (accessed 13 September 2026).

NOAA/NCEI, 2016. Meteorological Versus Astronomical Seasons. National Centers for Environmental Information. https://www.ncei.noaa.gov/news/meteorological-versus-astronomical-seasons (accessed 13 September 2026).

NWS, 2024. Service Change Notice 24-41: Upgrade of National Blend of Models Guidance, effective May 15, 2024 [WWW Document]. National Weather Service. URL https://www.weather.gov/media/notification/pdf_2023_24/scn24-41_nbm_v4.2.pdf (accessed 9.11.26).

Peddinti, S.R., Kambhammettu, B.P., 2019. Dynamics of crop coefficients for citrus orchards of central India using water balance and eddy covariance flux partition techniques. Agric. Water Manag. 212, 68–77.

Peddinti, S.R., Nicolas, F., Raij-Hoffman, I., Kisekka, I., 2025. Measurement and Modeling of Evapotranspiration for Processing Tomatoes Under Subsurface Drip Irrigation Using Eddy Covariance, pyfao56, and Particle Swarm Optimization. J. Nat. Resour. Agric. Ecosyst. 3, 149–163.

Perera, K.C., Western, A.W., Nawarathna, B., George, B., 2014. Forecasting daily reference evapotranspiration for Australia using numerical weather prediction outputs. Agric. For. Meteorol. 194, 50–63.

PyKrige developers, 2026. pykrige.ok.OrdinaryKriging API documentation. https://geostat-framework.readthedocs.io/projects/pykrige/en/stable/generated/pykrige.ok.OrdinaryKriging.html (accessed 13 September 2026).

Tian, D., Martinez, C.J., 2014. The GEFS-based daily reference evapotranspiration (ETo) forecast and its implication for water management in the southeastern United States. J. Hydrometeorol. 15, 1152–1165.

U.S. Department of Commerce, 2026. Experimental Forecast Reference Evapotranspiration (FRET) Information. National Weather Service. https://www.weather.gov/cae/fretinfo.html (accessed 3 July 2026).

USGS, 2026. 3D Elevation Program (3DEP) [WWW Document]. U.S. Geological Survey. URL https://www.usgs.gov/3d-elevation-program (accessed 9.11.26).

Vanella, D., Intrigliolo, D.S., Consoli, S., Longo-Minnolo, G., Lizzio, G., Dumitrache, R.C., Mateescu, E., Deelstra, J., Ramírez-Cuesta, J.M., 2020. Comparing the use of past and forecast weather data for estimating reference evapotranspiration. Agric. For. Meteorol. 295, 108196.

Vose, R.S., Applequist, S., Squires, M., Durre, I., Menne, M.J., Williams Jr., C.N., Fenimore, C., Gleason, K., Arndt, D., 2014. Improved historical temperature and precipitation time series for U.S. climate divisions. J. Appl. Meteorol. Climatol. 53, 1232–1251. https://doi.org/10.1175/JAMC-D-13-0248.1.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63, 1309–1313.

Yang, Y., Cui, Y., Luo, Y., Lyu, X., Traore, S., Khan, S., Wang, W., 2016. Short-term forecasting of daily reference evapotranspiration using the Penman-Monteith model and public weather forecasts. Agric. Water Manag. 177, 329–339.