This manuscript presents a comprehensive and well-executed analysis of wildfire impacts on evapotranspiration, basin outflow, and water-balance behavior across major California watersheds. The long-term perspective, integration of multiple datasets, and focus on water-supply relevance make this a valuable contribution to the postfire hydrology literature. The results are generally convincing and clearly presented. I have a few comments below that I believe would help strengthen the methodological clarity and interpretation:

We thank Referee #1 for the thoughtful and constructive comments. In the potential revision, we will clarify the role of basin water storage (ΔS) in the annual water balance interpretation, justify key analytical thresholds with references and sensitivity tests, add an independent precipitation comparison to support attribution of closure deviations, and expand discussion of scale mismatch between burned areas, ET aggregation, and downstream FNF stations. Below we respond point-by-point and indicate the changes will be made in the manuscript.

1.The watershed descriptions highlight volcanic terrain, groundwater-fed baseflow, and snowmelt-driven recharge, suggesting substantial subsurface storage capacity in several basins. While FNF integrates both surface and subsurface discharge, the manuscript does not explicitly discuss potential changes in basin water storage (ΔS). Although I am not deeply familiar with California's montane aquifer and soil storage dynamics, postfire changes in infiltration and recharge could plausibly lead to transient storage effects that influence P–ET–FNF residuals. Clarifying whether storage changes are assumed negligible, and over what timescales, would strengthen the interpretation of the water-balance analysis.

Response:

Thank you for this helpful comment. We agree that basin water storage changes (ΔS) may contribute to deviations between $P - ET$ and FNF, particularly in montane watersheds characterized by volcanic terrain, groundwater-fed baseflow, and snowmelt-driven recharge. Wildfire can plausibly influence infiltration, soil-water storage, and groundwater recharge, which in turn may generate transient storage changes that affect annual water-balance residuals.

In the original manuscript, the comparison between $P - ET$ and FNF was framed primarily as a water-balance closure diagnostic. In the revised manuscript we will clarify that the annual residual $R = P - ET - \text{FNF}$ should be interpreted as an apparent residual, which may reflect a combination of (i) basin storage change (ΔS) and (ii) observational or structural uncertainties in precipitation, evapotranspiration, and FNF datasets.

We also clarify the relevant timescale of the analysis. Because our calculations are based on annual water-year totals, short-term storage fluctuations (e.g., seasonal soil-water variations) are partly damped when aggregated to annual scales. However, wildfire may still induce multi-season storage transients through changes in snow accumulation/melt, infiltration capacity, soil moisture retention, and groundwater recharge/discharge. Without independent basin-scale storage observations, it is not possible to uniquely partition the residual between ΔS and measurement uncertainty. We therefore interpret systematic deviations as indicative of combined hydrologic and observational processes rather than as strict closure errors.

These clarifications will be added to the Methods and Discussion sections, and terminology referring to "closure error" or "imbalance" will be revised to "apparent residual" where appropriate.

Revisions:

To address this point, we will revise the description of the water-balance framework in the Methods section (Section 2.3, Lines 186–189). The previous text stated that under steady-state conditions (ΔS ≈ 0), $P - ET$ should closely match FNF and that deviations represent water-balance imbalance. In the revised manuscript, this paragraph will clarify that the annual residual $R = P - ET - \text{FNF}$ is interpreted as an apparent residual, reflecting the combined effects of potential storage change (ΔS) and observational or structural uncertainties in precipitation, evapotranspiration, and FNF datasets.

Lines 186–189 (Methods, Sec. 2.3): We will replace the "steady-state (ΔS ≈ 0)" sentence with an interpretation of the residual. Original text: "Under steady-state conditions (ΔS ≈ 0), P − ET should closely match FNF. Deviations from the 1:1 relationship were therefore used to identify potential imbalance and diagnose hydrologic processes..." will be revised to: "At annual (water-year) timescales, changes in basin storage (ΔS) are often smaller than the dominant flux terms, but they are not necessarily negligible, particularly following major disturbance. Accordingly, deviations between P−ET and FNF are interpreted as an 'apparent non-closure residual', R = P−ET − FNF, which may reflect a combination of ΔS and uncertainties or structural mismatches in P, ET, and/or FNF."

Lines 289–292: Original text: “Under ideal conditions, modeled P and ET should closely reproduce observed runoff, resulting in well-aligned P−ET and FNF curves. In basins where ET remains stable and unaffected by disturbance, a near-linear relationship between P and FNF is expected, with data points clustering along the 1:1 line (reflecting annual storage change) and a slope approaching unity.” will be revised to “Under ideal conditions, modeled P and ET should broadly reproduce observed runoff, resulting in similar temporal behavior between $P - ET$ and FNF. In basins where ET remains stable and disturbance effects are limited, a near-linear relationship between P and FNF is expected, with data points clustering near the 1:1 line and a slope approaching unity. Deviations from this relationship represent the combined effects of transient storage variation, dataset uncertainty, and disturbance-related hydrologic responses.”

Finally, the Discussion section (Section 4.3, Lines 403–410) will be revised to clarify that wildfire-related deviations in $P - ET -$ FNF are interpreted as shifts in apparent residuals, which may reflect a combination of transient storage dynamics and observational or structural uncertainties rather than a strict water-balance imbalance.

Line 403 (Discussion section title)*:* Change heading to avoid “correction.” The original text: “4.3 Diagnosing and correcting postfire water-balance closure errors” will be revised to “4.3 Diagnosing postfire water-balance non-closure and residual structure”

Lines 404–410 (Discussion, wildfire effects)*:* Change language to “correlate” and emphasized combined residual effects. The original text: “Wildfire disturbance emerged as an important driver of deviations from expected water-balance closure... fire years consistently exhibited more negative residuals (P − ET − FNF), indicating that observed basin outflow (FNF) systematically exceeded modeled P − ET... This divergence demonstrates that wildfires can create short-term hydrologic imbalances...” will be revised to: “Wildfire disturbance emerged as an important correlate of deviations from apparent annual water-balance closure across California’s major watersheds. Fire years consistently exhibited more negative residuals (P−ET−FNF), indicating that observed basin outflow (FNF) exceeded P−ET to a greater degree during and immediately after disturbance. Because the annual residual reflects the combined influence of ΔS and uncertainties or structural mismatches in P, ET, and FNF, these results indicate that wildfire years are associated with systematic shifts in the residual structure rather than implying a single underlying mechanism.”

2.Several analytical thresholds appear somewhat arbitrary and would benefit from additional justification or brief sensitivity testing. These include the 75% recovery benchmark, the 3% burned-area threshold used to define high-fire years, and the 500 m buffer for selecting unburned reference areas. Providing a short explanation or supporting references for these choices would improve methodological transparency.

Response:

We thank the reviewer for noting that several analytical thresholds require clearer justification. We agree that these values should be explicitly motivated and will revise the manuscript to provide brief rationale and supporting citations for each threshold.

75% recovery benchmark. We defined recovery as the first year in which post-disturbance ET reached ≥75% of the pre-fire mean, representing substantial but not complete functional return toward baseline conditions. We selected this threshold as an operational criterion that captures meaningful recovery while avoiding the unrealistic requirement of full (100%) return, which may not occur in semi-arid or disturbance-prone ecosystems. Similar fractional recovery thresholds (e.g., 70–80% of pre-disturbance conditions) have been used in remote sensing–based vegetation and post-fire recovery studies to operationalize functional recovery following stand-replacing disturbances (e.g., White et al., 2018; White et al., 2022). We will clarify this rationale in the Methods section.

500 m buffer for unburned reference areas. The 500 m buffer distance was selected to balance two competing considerations, maintaining environmental comparability between burned and reference areas, and reducing spatial dependence and fire-edge effects that could bias recovery estimates. Comparable reference selection distances (on the order of several hundred meters) have been applied in post-fire remote sensing and ecological studies to limit spatial autocorrelation while preserving local environmental similarity (e.g., Jin et al., 2012; Reddy et al., 2015; Dias & Acácio, 2024). We will add clarification and supporting citations accordingly.

3% burned-area threshold. The 3% basin burned-area threshold was used as an operational cutpoint to identify fire years likely to produce detectable basin-scale hydrologic signals while retaining sufficient sample size for analysis. Published basin-scale studies have used thresholds ranging from approximately 1% (e.g., Hallema et al.) to 5% (e.g., Beyene et al., 2022), reflecting different trade-offs between disturbance detectability and statistical robustness. Our choice of 3%

lies within this commonly applied range and was intended to balance signal detectability against sample retention. To further evaluate the robustness of this choice, we will conduct a brief sensitivity analysis using alternative thresholds and report whether the main qualitative conclusions regarding wildfire effects on basin water balance remain consistent.

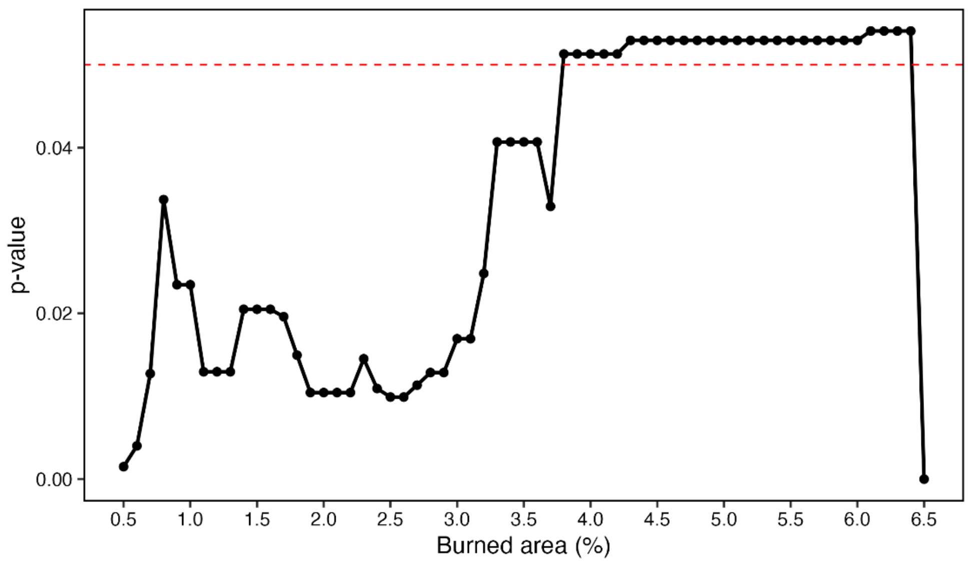


Figure S. Sensitivity analysis of the selection of the threshold for the percentage of area burned defined as high-fire years with Welch's two-sample t-test. The effect is statistically significant when the threshold ranges from 0.5% to 3.8%. However, when the threshold exceeds 3.8%, the number of remaining wildfire events becomes too small (fewer than three), resulting in insufficient sample size for robust statistical inference.

Revisions: we will revise accordingly for the text in Section 2.3 (Lines 155–157), In Section 2.3 (Lines 166–167), and in Section 2.3 (Lines 181–182).

3.The CECS ET product is derived from internal water-balance calculations and is then used in the basin-scale closure analysis. Because ET, P, and runoff are therefore not fully independent within this framework, it would be helpful for the authors to clarify how assumptions within the CECS product may influence the observed P–ET–FNF residuals. A brief discussion of potential error propagation or circularity would strengthen confidence in the closure diagnostics.

Response:

We thank the reviewer for raising this important point regarding the potential dependence among precipitation, evapotranspiration, and runoff estimates within the CECS framework.

We clarify that the basin-scale closure analysis in this study uses independent flux estimates for each component of the water balance. Precipitation (P) is derived from the PRISM dataset and spatially downscaled within the CECS framework to 30-m resolution. Evapotranspiration (ET) is estimated independently using a remote-sensing and process-based modeling framework and is not calculated as a residual of precipitation or runoff.

Although the CECS modeling framework includes an internal water-balance structure capable of producing modeled runoff estimates, those modeled runoff outputs are not used in this study. Instead, basin outflow is represented by Full Natural Flow (FNF) observations obtained from the CDEC system. These FNF estimates are independently derived from hydrometric observations and reservoir accounting procedures.

Therefore, the basin-scale closure comparison in this study evaluates PRISM-derived precipitation and independently estimated ET against independently derived FNF observations. The three components used in the $P - ET -$ FNF comparison are not calculated from a single water-balance equation, and thus the analysis does not introduce circularity.

In the revised manuscript we will clarify this independence and briefly discuss how uncertainties in precipitation, ET estimation, and FNF reconstruction may propagate into the apparent residual $P - ET -$ FNF. As described in our response to the previous comment, the residual is interpreted diagnostically and may reflect a combination of basin storage change (ΔS) and uncertainties associated with each flux estimate.

Revisions:

Line 142-146, the original text will be revised into: “CECS provides annual evapotranspiration (ET) and precipitation (P) raster products from 1985 to 2024 at 30-m resolution derived from its internal water-balance calculations” will be revised into “The Center for Ecosystem Climate Solutions (CECS) provides annual evapotranspiration (ET) products at 30-m spatial resolution across California. These ET estimates are derived from a Landsat-based empirical model that combines remotely sensed vegetation indices with climate variables and eddy-covariance flux-tower observations. Specifically, annual ET is estimated from relationships between Landsatderived normalized difference vegetation index (NDVI) and measured ET at flux-tower sites distributed across California forests (Goulden & Bales, 2019; Guo., et al, 2023; Chung et al., 2024).

Precipitation (P) was obtained from the PRISM climate dataset, which provides gridded daily precipitation at approximately 800-m spatial resolution. Annual precipitation totals were aggregated from daily PRISM data and resampled to match the 30-m ET grid (Roche et al., 2020)".

Lines 471–473, the original text: "Persistent negative P–ET–FNF residuals reflect not only process-level shifts but also precipitation underestimation, gauge bias, and heterogeneous subsurface pathways." will be revised to "Persistent negative $P - ET - FNF$residuals reflect a combination of disturbance-driven hydrologic responses, transient storage variations, and observational uncertainties such as precipitation underestimation, gauge bias, and heterogeneous subsurface pathways."

Lines 147–152: Revise the sentence that begins "Additionally, we obtained monthly full natural flow…" to"Additionally, we obtained monthly full natural flow (FNF) estimates from the California Data Exchange Center (CDEC). CDEC defines FNF (also termed 'unimpaired runoff') as the natural water production of a basin absent upstream diversions, storage regulation, and net basin imports/exports; FNF is calculated by adjusting gauged flows at the reporting location to account for upstream operations. We use CDEC FNF as the best available long-term unimpaired outflow estimate for these basins, while recognizing that uncertainty in the naturalization procedure and in upstream operation/diversion records can contribute to apparent water-balance residual structure."

4.The manuscript primarily attributes closure deviations to precipitation underestimation and stream-gauge bias, which is plausible given the use of gridded 30 m precipitation data. However, this interpretation remains largely inferential. Comparison with one or more independent precipitation products (where available) could help further support this conclusion and strengthen attribution of closure imbalances.

Response:

We thank the reviewer for this helpful suggestion. We agree that attributing closure deviations to precipitation underestimation based solely on residual analysis is partly inferential, and that comparison with an independent precipitation dataset can strengthen interpretation.

To address this point, we compared basin-scale annual precipitation derived from PRISM with an independent gridded precipitation dataset (TerraClimate) over the study period. The comparison shows strong agreement in interannual variability across basins, but also reveals systematic differences in magnitude in some cases, with PRISM generally lower than TerraClimate in higher-precipitation conditions (Fig. S).

At the basin scale, time series comparison in the DAV basin further shows that differences between datasets vary through time, with both positive and negative deviations (PRISM − TerraClimate), indicating that precipitation uncertainty is non-negligible but not unidirectional (Fig. S).

These results support our revised interpretation that precipitation uncertainty may contribute to the apparent water-balance residuals in some basins, particularly in complex mountainous terrain, while also reinforcing that residual structure reflects a combination of factors, including storage variation and observational uncertainty in multiple components.

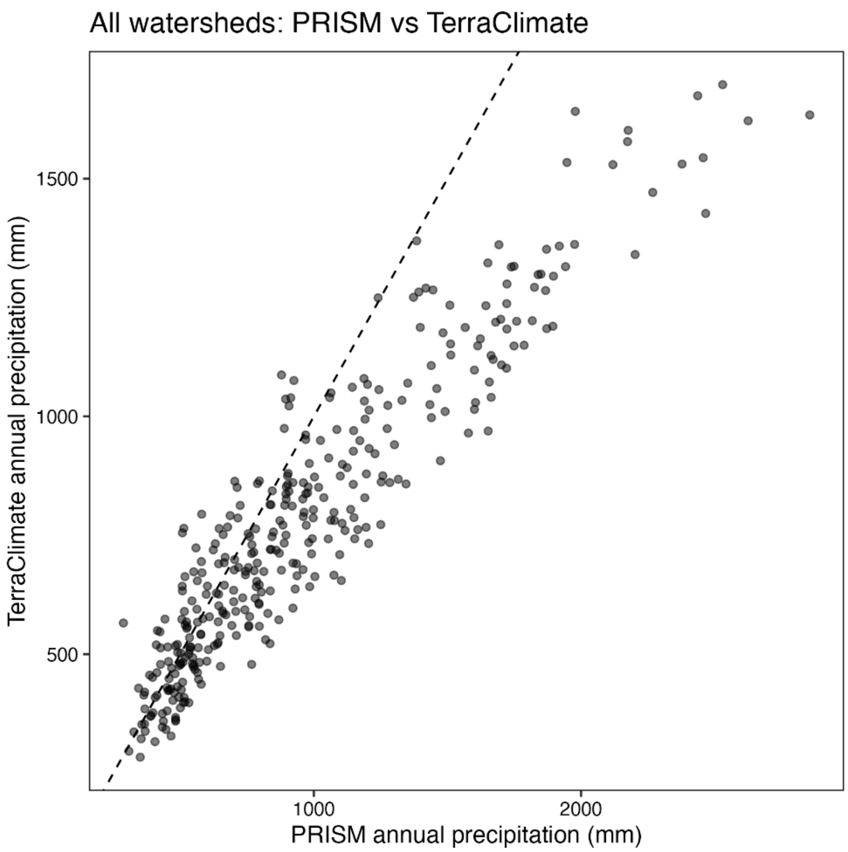
We will revise the manuscript accordingly to (1) explicitly describe this independent dataset comparison, and (2) avoid attributing closure deviations solely to precipitation underestimation.

Revisions:

In the Methods section after line 178, we will add "To evaluate potential precipitation uncertainty, basin-mean precipitation derived from PRISM was compared with an independent gridded precipitation dataset (TerraClimate) over the overlapping study period. (see Fig. S)." This addition clarifies that precipitation uncertainty was evaluated using an independent dataset rather than inferred solely from the closure residuals.

Discussion section describing the DAV basin case (Lines 422–430), the interpretation will be revised to avoid implying that precipitation underestimation is definitively responsible for the observed discrepancy. The original wording:"Instead, this discrepancy likely reflects precipitation underestimation." will be revised to: "Instead, this discrepancy may reflect precipitation uncertainty, particularly given the potential for interpolation error in complex mountainous terrain." In addition, the following sentence will be added to acknowledge the independent dataset comparison:"Comparison with an independent precipitation dataset shows broadly consistent basin-scale precipitation patterns but modest differences in some years (Fig. Sx), suggesting that precipitation uncertainty may contribute to the observed imbalance."

The Conclusion section (Lines 491–494) will be revised for consistency. The phrase: “closure errors from precipitation underestimation and stream-gauge bias” will be replaced with:“closure errors associated with precipitation uncertainty and stream-gauge bias.”



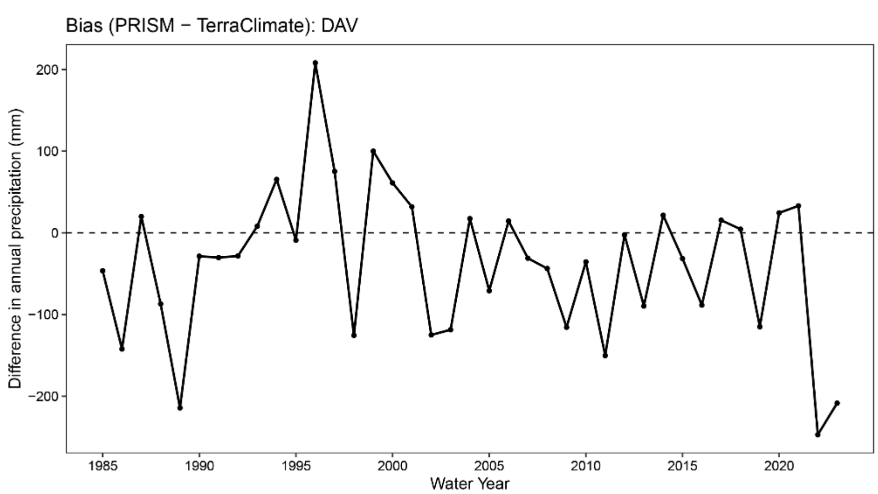


Figure S. Comparison of annual precipitation between PRISM and TerraClimate across all basins and Annual precipitation differences between PRISM and TerraClimate in the DAV basin (PRISM − TerraClimate).

5.Most fires affected relatively small portions of the basins, which likely limits detectability of hydrologic responses at downstream gauging stations. While this issue is acknowledged, additional discussion of potential scale mismatch between burned areas, ET aggregation, and FNF stations would help clarify the limits of inference, particularly for weaker runoff responses.

Response: We thank the reviewer for highlighting the important issue of spatial scale mismatch between burned areas, evapotranspiration (ET) aggregation, and basin-scale runoff observations. We agree that this scale mismatch can limit the detectability of wildfire-induced hydrologic responses at downstream gauging stations.

In our study, most fires affected relatively small fractions of the total watershed area, while full natural flow (FNF) observations integrate hydrologic responses across the entire basin. As a result, localized reductions in ET within burned areas may be diluted when aggregated at basin scale, particularly in large watersheds where burned area represents only a small proportion of total contributing area. This spatial averaging likely contributes to the weak or inconsistent runoff responses observed at some gauging stations.

To clarify this limitation, we have expanded the discussion to explicitly address the potential scale mismatch among burned area extent, ET aggregation, and basin-integrated FNF observations. These additions emphasize that small or spatially fragmented fires may produce localized hydrologic responses that are not detectable at downstream basin outlets.

Revisions:

In the Results section (Lines 269–276), we will revise the sentence describing the weak runoff response to wildfire to acknowledge that basin-scale aggregation may reduce detectability of hydrologic signals. The original sentence: “ET reductions were generally insufficient to produce measurable changes in FNF at gauging stations.” will be revised to: “ET reductions were generally insufficient to produce measurable changes in FNF at gauging stations, in part because most fires affected only small portions of the contributing watershed area and their hydrologic effects may be diluted when integrated across basin-scale runoff observations.”

In the Discussion section (Lines 389–397), we will expand the paragraph discussing watershed-scale hydrologic responses to explicitly describe the scale mismatch between burned areas and basin-integrated discharge observations. The paragraph beginning: “Most fires affected limited portions of each watershed, limiting their measurable hydrologic impact at larger scale.” will be revised to: “Most fires affected limited portions of each watershed, limiting their measurable hydrologic impact at the basin scale. Because full natural flow (FNF) observations integrate hydrologic responses across the entire watershed, localized ET reductions within burned areas may be diluted when aggregated to the basin outlet. Consequently, wildfire-induced changes in evapotranspiration may not produce detectable runoff signals unless a substantial fraction of the watershed is burned.”

In addition, we will add the following clarifying sentence at the end of the same paragraph (Lines 394–397): “This scale mismatch between burned-area extent, spatially aggregated ET estimates, and basin-integrated runoff observations introduces an important limitation for interpreting weak runoff responses, particularly when fires affect only small or spatially fragmented portions of the watershed.”

These revisions explicitly acknowledge the spatial-scale limitations associated with basin-integrated runoff observations and clarify the limits of inference for smaller fires.