Mind the Gap: Improving Water Markets with Field-Level Remote Sensing*

Katherine Wright† Andrew Ayres‡ Bryan Leonard§

PRELIMINARY DRAFT. PLEASE DO NOT CIRCULATE WITHOUT PERMISSION.

Abstract

Water markets are increasingly proposed as a climate-adaptation tool, but are hindered by the inability to accurately measure field-scale water use. Existing transfers often quantify savings by applying coarse averages and assuming fallowed land uses no water. We combine satellite evapotranspiration data with hand-coded contract records to evaluate one of the largest agricultural-to-urban water transfers in the United States. Causal estimates indicate that actual water savings are only 54% of reported savings, implying a cumulative shortfall of 624,916 acre-feet over 2005–2021. Roughly half the gap is residual water use on fallowed fields, driven by hydrologic spillover from irrigated neighbors. The other half is primarily due to strategic fallowing of fields with lower expected water use. Neither mechanism could be reliably detected without field-scale measurement.

*For helpful comments we thank Josh Abbott, Anna Boser, Tamma Carleton, Eric Edwards, Robert Heilmayr, Kelsey Jack, Kailin Kroetz, Morgan Levy, Buzz Thompson, and Will Rafey, as well as seminar participants at the Workshop on Advancing Sustainable Water Management at Stanford University, the Bren School of Environmental Science and Management at UC Santa Barbara, the Environmental and Resource Economics Series at UC San Diego, the Land, Environment, Economics and Policy Institute at the University of Exeter, the Agricultural and Resource Economics Department at UC Davis, the Economics Department at the Colorado School of Mines, the Department of Agricultural Economics & Economics at Montana State University, the Department of Applied Economics at Oregon State University, and the UC Riverside Water Dialogues. Talia Greco, Danielle Moon, and Mithran Mathiraj provided excellent research assistance. We have also benefited greatly from data sharing and discussions with staff at Metropolitan Water District.

†Hillsdale College kwright1@hillsdale.edu

‡University of Nevada, Reno andrew.ayres@unr.edu

§University of Wyoming bryan.leonard@uwyo.edu

Introduction

Many arid regions are already experiencing more severe surface water scarcity due to climate change (Rodell et al., 2018; Elliott et al., 2014; Gordon et al., 2024). The Western United States is a prime example. The Colorado River supplies water to 40 million people and irrigates more than five million acres, and consumptive demands have exceeded the river's flow in 16 of the years between 2000 and 2020 (Richter et al., 2024). In this setting, ensuring that water is allocated to its highest-valued use has emerged as a central policy challenge. Agriculture accounts for as much as 80% of freshwater use in the basin (Richter et al., 2020; Medellín-Azuara et al., 2024), even though urban willingness-to-pay often far exceeds the market value of water in agriculture (Brewer et al., 2008), and recent evidence suggests that the allocation even within agriculture is inefficient (Arellano-Gonzalez et al., 2021; Rafey, 2026).

Water markets are a leading candidate solution (Brewer et al., 2008; Ayres et al., 2021; Bruno and Jessoe, 2021; Rafey, 2023, 2026; Bruno and Jessoe, 2024), but their expansion has been hindered by transaction costs related to unclear property rights, administrative approvals, and third-party objections (Bretsen and Hill, 2008; Leonard et al., 2019; Hagerty, 2023). Many of these costs arise from the difficulty of estimating field-scale consumptive water use to determine how much of a farmer's water right can be transferred without reducing return flows and impairing other users. To minimize these costs, municipalities, NGOs, and the U.S. Bureau of Reclamation increasingly use "alternative transfer mechanisms" (ATMs)—contracts that pay farmers to change on-farm practices (i.e., fallowing or crop-switching), often for limited duration, with the understanding that water savings accrue to the buyer (Dilling et al., 2019). Water transfers of various forms span the Colorado River Basin, the broader Western United States, and other countries including Australia, Spain, and Chile (Grafton et al., 2012; Garrick et al., 2023; Rafey, 2023).

The verification problem that surface water transfers face is not unique. Programs that pay for changes in land use to deliver an environmental outcome—REDD+ forest carbon, agricultural conservation practices, and biodiversity offsets—often rely on two crucial assumptions to estimate environmental benefits: that the targeted activity (deforestation, irrigation, development) ceases entirely on enrolled land, and that enrolled land would have been used just as intensively as unenrolled land. Both assumptions are testable, but only when reliable measurement of the targeted activity is feasible at the contract scale. Recent advances in satellite measurement now make field-scale verification possible for water use (Volk et al., 2024). Despite this measurement revolution, the magnitude of the verification gap in real-world water transfer programs has not been documented, although related questions have been studied in technology adoption programs (Wong et al., 2026) and in conservation auctions, where landowners' private information about counterfactual land use reduces the realized conservation benefit of enrolled acres (Aspelund and Russo, 2026). Here, we fill this gap, in a setting characterized by arid, flood-irrigated agriculture without the possibility of groundwater substitution, where the verification problem is especially tractable.

We study the Palo Verde Irrigation District–Metropolitan Water District Forbearance and Fallowing Program, one of the largest and longest-running agricultural-to-urban alternative water

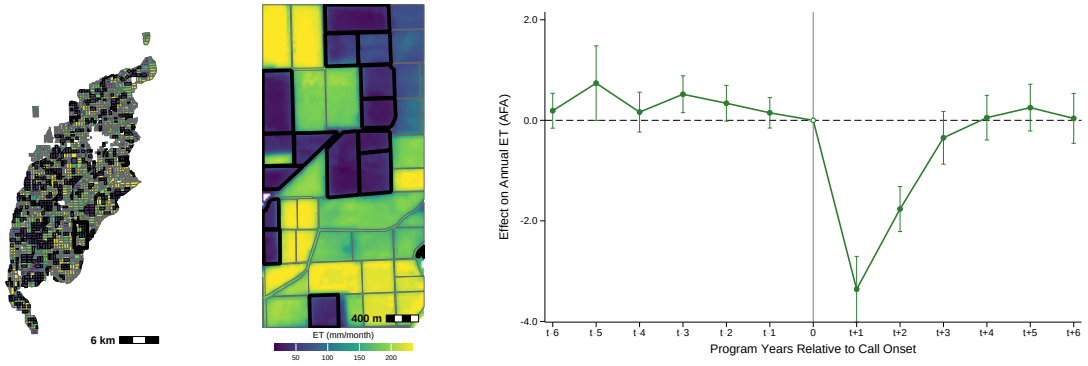
transfers to date, involving the largest municipal water wholesaler in the United States. We combine satellite evapotranspiration data from OpenET (Volk et al., 2024) with hand-digitized administrative records of fallowing calls issued by MWD between 2007 and 2021 to estimate the causal effect of a fallowing call on field-level water use. Identification rests on a dynamic difference-in-differences estimator (De Chaisemartin and d'Haultfoeuille, 2020) that accommodates the rotational, on/off nature of fallowing and absorbs owner-by-year shocks. We then decompose the gap between our estimated savings and those reported by the program into hydrologic spillover from irrigated neighbors (isolated from atmospheric measurement artifact using monthly prevailing wind directions) and strategic selection of which fields are offered (identified using AR(1) and LASSO predictions of counterfactual ET). The analysis reveals structural biases in common methods currently used by agencies such as the US Bureau of Reclamation to estimate the savings from water transfer programs.

Results

Water Savings from Fallowing Are Substantially Lower than Program Estimates

We estimate the causal effect of a fallowing call on field-level water use using a dynamic difference-in-difference framework that compares ET on called fields before versus during a call to ET on un-called fields over the same period. Panel (a) of Figure 1 depicts an illustrative example, with thicker borders indicating called fields, and darker shading indicating less water use. Panel (b) of Figure 1 reports the event-study estimates using the De Chaisemartin and d'Haultfoeuille (2020) estimator, which is the only modern DiD framework that accommodates the on/off rotational nature of fallowing assignments. Called fields exhibit similar trends to untreated parcels prior to onset and reduce their water use by an average of 2.53 acre-feet per acre (AFA) per year during a fallowing event ("spell") before returning to baseline.

Figure 1: Identifying and Measuring the Water-Savings Gap



(a) Field-Scale Measurement (b) Event Study

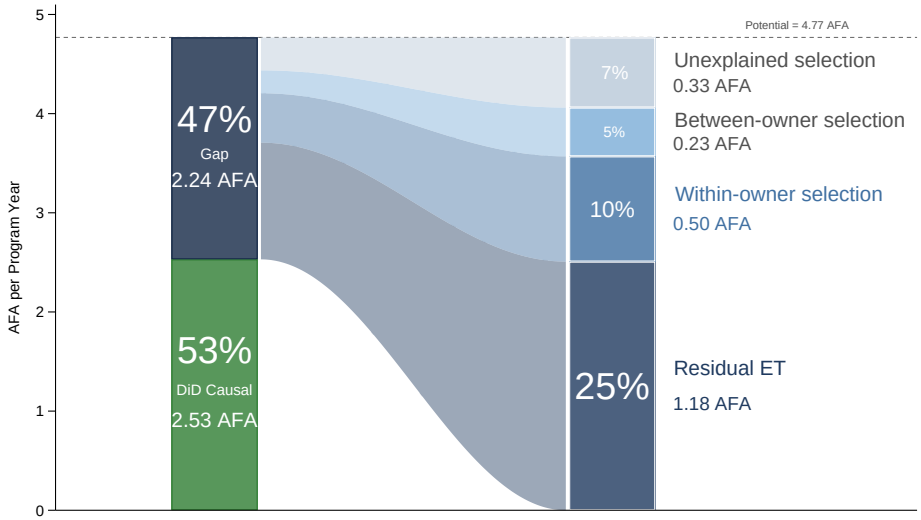
Notes: Panel (a) depicts field-scale water use from the OpenET Ensemble as well as "called" fields in August 2011, a representative example from our data. Panel (b) reports the dynamic event study estimates with six pre-period placebos and six post-onset effects. Connected green line with 95% CI bars, clustered at PLSS section.

We verify the robustness of our savings estimate in a variety of ways. The magnitude varies

slightly across the choice of ET model, with estimates for our preferred specification ranging from 1.84 to 3.01 AFA across the seven OpenET models (SI Table S3). The full set of 42 specifications, varying treatment coding, sample restrictions, and control sets, is summarized in SI Fig. S5. To check whether farmers offset fallowing on called fields by using more water on their other fields—which would inflate our savings estimates—we examined water-use patterns on *uncalled* fields of called farmers vs. uncalled farmers (SI Figure S8) and re-estimated the main savings effect under alternative specifications that address this concern (SI Table S4). Neither shows evidence of significant compensating water use, leaving within-farm reallocation of water as an unlikely driver of our results.

In contrast to the field-scale approach depicted in Figure 1, the method of estimating savings agreed to by MWD, USBR, and PVID for their annual verification reports (VR method) multiplies district-wide average water use per acre by the number of contracted fallow acres. This procedure makes two implicit assumptions: that average expected water use does not differ between called and uncalled fields and that water use goes to zero on fallowed parcels. Our empirical strategy is designed to relax both, and allows us to isolate the effect of each assumption on the overall differences between our causal estimates and the VR methodology. Figure 2 depicts the overall gap between the causal estimate and the average reported savings in MWD's verification reports, as well as our decomposition of the differences.

Figure 2: Decomposing the Savings Gap



Notes: Decomposition of the Potential savings rate (4.77 AFA, the ET baseline on uncalled ever-enrolled fields). Stage 1 splits Potential into the preferred DiD causal estimate (2.53 AFA, green, 53% of Potential) and a 2.24 AFA residual Gap (dark blue). Stage 2 decomposes the Gap into its four mechanism components: *Residual ET* on called fields (1.18 AFA), selection *Between owners* (0.23 AFA), selection to *Within owner* (0.50 AFA), and *Unexplained* selection (0.33 AFA).

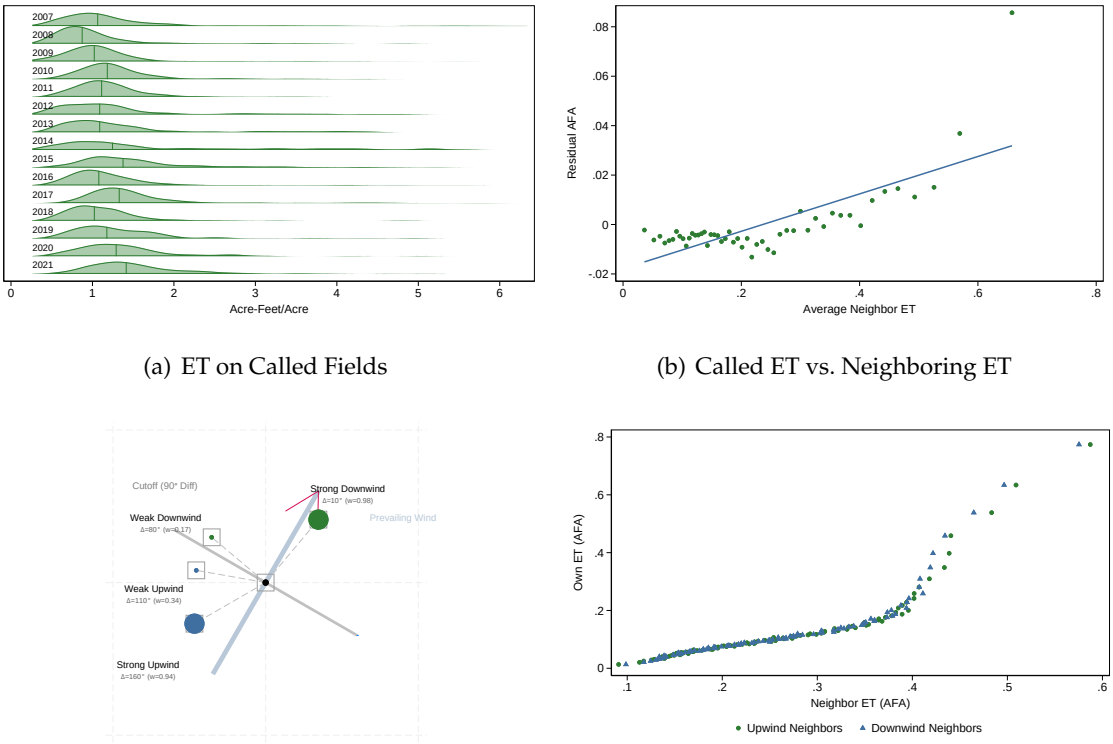
Average annual savings reported by MWD over 2005–2021 are approximated by district-wide average annual water use, at 4.67 AFA. Average water use on *enrolled* fields is slightly higher than the district-wide average, putting potential savings under perfect compliance and additionality at 4.77 AFA (SI Fig. S9). The realized causal savings of 2.53 AFA represent 53% of potential; the re-

maining 2.24 AFA is the savings gap. Next, we demonstrate that roughly half the gap is persistent water use on called fields (1.18 AFA), and roughly half is adverse selection of low water-use fields into the program (1.06 AFA).

Consumptive Use on Fallowed Fields Driven by Neighbors' Irrigation

A meaningful number of called fields have positive ET, with an overall mean of 1.18 AFA/year across fields and years (panel (a) of Figure 3). This residual accounts for roughly half the savings gap. A similar finding for fallowed land in PVID was previously noted by Wobus et al. (2024), who attributed it to errors in satellite classification of fallowed lands. Such measurement errors cannot explain the persistent use we observe, because our classification of fallowing is based on administrative records of fields that MWD verifies are in fact fallow.

Figure 3: Water Use on Fallowed Fields



(a) ET on Called Fields

(b) Called ET vs. Neighboring ET

(c) Upwind Cone Illustration

(d) Upwind vs. Downwind Neighbors

Notes: Panel (a) is a joy plot of annual ET across called field-years, one distribution per year. Panel (b) is a binscatter of within-field residual ET (after partialling out field and month fixed effects) against the field's average neighbor ET in the same month; green dots = 50 equal-count bins, blue fit line. Panels (c)–(d) illustrate our approach for ruling out atmospheric spillovers: (c) schematic of the prevailing-wind cone classification, where each neighbor is weighted by cos of its bearing relative to the monthly prevailing direction; (d) scatter of own ET against upwind-weighted (green dots) vs. downwind-weighted (blue triangles) neighbor ET on called field-months — the near-identical slopes bound atmospheric (wind-borne vapor) contamination at $\leq 6\%$ of the total spillover (econometrics and wind-test power calculation in SI).

We explore several potential drivers—distance to the Colorado River, soil drainage, cumulative length of fallowing spells, effective precipitation, and water use on neighboring fields (SI Fig. S10 and SI Table S5). Of these, neighbor ET is the only significant predictor. Panel (b) of Figure 3 plots within-field residual ET against the field's mean neighbor ET in the same month. The relationship between fallow ET and neighboring ET is statistically significant and large enough to fully explain fallow ET (Materials and Methods).

Next, we assess whether neighbor effects reflect actual spillover of applied water or atmospheric contamination of the satellite ET signature. Evaporative cooling on irrigated parcels could lower surface temperatures on adjacent fallowed parcels, creating a false ET reading. We rule out the latter channel using monthly prevailing wind directions constructed from hourly station data. Atmospheric vapor transport should originate disproportionately from upwind neighbors, and satellite measurement spillovers would reflect this. Panel (d) of Figure 3 plots own ET against cosine-weighted upwind versus downwind neighbor ET (weighting scheme depicted in panel (c)); the slopes are statistically indistinguishable ($p = 0.64$). At 80% power, the design could have detected an upwind–downwind contrast of 0.024 AFA, or 5.9% of the total neighbor-ET slope (SI Table S7). The atmospheric channel is therefore bounded well below the magnitude that would be required for it to be the dominant mechanism.

We also examine whether the residual reflects non-agricultural ET that is not associated with applied irrigation water. To do so, we adapt the methodology of Boser et al. (2024) to estimate agricultural ET as a sub-component of total ET using machine learning so that evapotranspiration that would have occurred in the absence of supplemental irrigation is not included in estimates of consumptive use. This approach was originally developed in California's Central Valley and relies on the identification of fallow fields to train a machine learning prediction of ET in the absence of irrigation and cropping. As depicted in SI Figure S14 and SI Table S11, applying the Boser et al. method to this setting produces estimates of non-agricultural ET that are implausibly large, resulting in agricultural ET estimates that are significantly lower than reported water use in PVID on a monthly and annual basis. When we re-estimate our main specification using these modeled data, the result is roughly 82.6% of the headline DiD effect (SI Table S10), albeit on a slightly different sample (Materials and Methods). Applying the agricultural share to the full-sample headline yields a smaller implied estimate of -2.090 AFA, and hence a larger gap from the VR reported savings . The upshot is that the residual water use component of the savings gap documented in SI Figure 3 is unlikely to be driven by natural ET; pre-adjusting the ET data by subtracting estimated natural ET actually *increases* the savings gap, rather than reducing it.

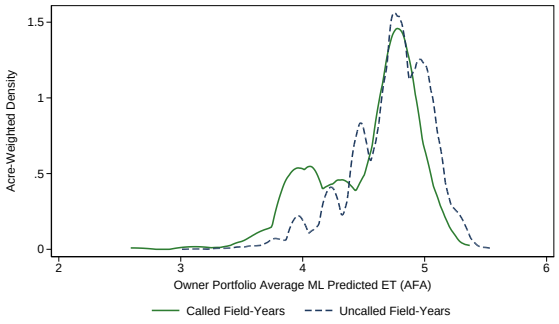
Strategic Farmer Selection of Fields into Fallowing Calls Reduces Potential Savings

The remaining 1.06 AFA of the savings gap reflects selection—farmers offer fields that would have used less water or been fallowed even without a call, so the contracted acreage yields savings that are systematically lower than district-wide average water use. This same type of strategic selection has recently been found to significantly erode expected benefits from the Conservation

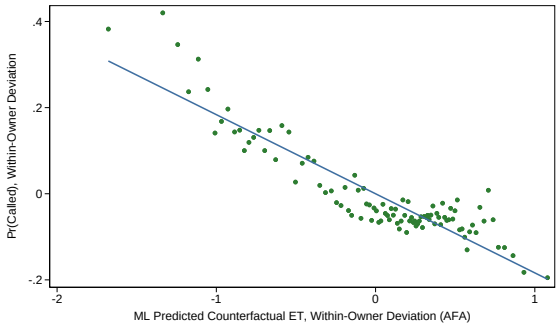
Reserve Program (Aspelund and Russo, 2026). We characterize the magnitude and structure of selection by predicting counterfactual water use on called fields with two different models. The first is an AR(1) model with field-by-month fixed effects. The second is a cross-validated LASSO with 2-year lags, neighbor-ET histories, and month-by-year fixed effects. We use these predictions to measure the extent of strategic selection of fields with lower-than-average predicted ET. First, we verify that fields with lower predicted ET are more likely to be selected for a call (SI Table S8).

Next, we distinguish between-owner from within-owner selection, focusing on the Lasso predictions. In panel (a) of Figure 4, we plot the distributions of acre-weighted portfolio-average predicted counterfactual ET separately at the owner-year level for called acres vs. acres left in production. By construction, this average does not vary within an owner-year, because the within-owner choice over which specific field to fallow is filtered out by the portfolio-averaging. Differences between the two densities are therefore driven entirely by compositional variation in which portfolios contribute to the called pool. The mean shift between the two densities indicates that called acres are drawn disproportionately from owner portfolios with somewhat lower predicted ET than the enrolled benchmark. Panel (b) shows the within-owner channel by residualizing on owner-by-year fixed effects and demonstrating that fields with lower predicted ET are more likely to be called than other fields associated with the same owner.

Figure 4: Selection into Fallowing



(a) Between-Owner Selection into Fallowing



(b) Within-Owner Selection into Fallowing

Component	AFA	Share
Between-Owner Selection	0.07	15%
Within-Owner Selection	0.37	85%
Total Predicted	**0.44**	**100%**

(c) Selection Decomposition, AR(1)

Component	AFA	Share
Between-Owner Selection	0.23	32%
Within-Owner Selection	0.50	68%
Total Predicted	**0.73**	**100%**

(d) Selection Decomposition, LASSO

Notes: Panel (a) plots the acre-weighted distribution of owner-year portfolio-average predicted counterfactual ET, separately for called vs. uncalled acres in a given year. Panel (b) is a within-owner-year binscatter of the call indicator (residualized on owner × year fixed effects) against predicted counterfactual ET (same residualization). Panels (a) and (b) both use the ML counterfactual prediction. Panels (c)–(d) decompose the selection gap into between-owner and within-owner components. Panel (c) uses an AR(1) model of own ET, yielding 0.44 AFA total predicted selection (15% between, 85% within). Panel (d) uses a cross-validated LASSO with 12- and 24-month lags, neighbor behavior, and month-by-year fixed effects, yielding 0.73 AFA (32% between, 68% within). The LASSO-predicted total absorbs ∼69% of the 1.06 AFA non-additionality component.

Panels (c) and (d) quantify the total observable selection effect, which is calculated as the difference between predicted ET on uncalled vs. called field-years among ever-enrolled fields. The total effect is 0.44 AFA under AR(1) (15% between, 85% within) and 0.73 AFA under LASSO (32% between, 68% within). The LASSO decomposition absorbs roughly 69% of the 1.06 AFA selection component of the savings gap. The remaining 0.33 AFA, which we label *unexplained*, captures selection on factors our prediction models cannot recover from spatial and historical covariates—private farmer information about expected ET, idiosyncratic operational decisions, or measurement noise in the prediction. The within-owner channel is the dominant component of observable selection under both prediction models, indicating that the dominant form of selection comes from the field-level decision within an enrolled portfolio.

Implications for the Lower Colorado River Basin

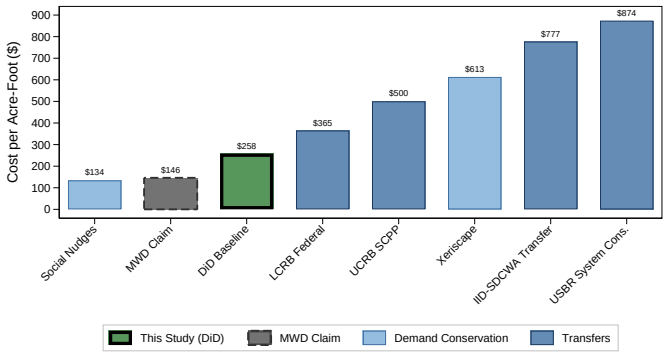
To translate per-acre savings into program-aggregate savings, we multiply our annual DiD estimates by the number of fallowed acres MWD reports in each program year. We then calculate the difference between VR reported savings and the savings implied by our causal estimate in each year. SI Figure S12 sums these annual differences over 2005–2021, with the preferred estimate indicated in green and the 41 alternative specifications in gray. The majority of estimates lie between 400,000 and 800,000 AF over the first 17 years of the program; our preferred estimate is 624,916 AF. Panel A of Figure 5 reports annualized versions of reported savings, our estimate of savings, and the associated gap, which is roughly 43.5% of reported savings on average. PVID as a whole typically consumes a reported 350,000 to 420,000 AF per year, so the cumulative shortfall is roughly 1.5 years of district-wide consumption.

The water MWD actually receives under the program is governed by Section 4.d of the 2003 Colorado River Water Delivery Agreement rather than by the Verification Reports. Section 4.d delivers to MWD the difference between a fixed 420,000 AF threshold and the combined consumptive use of PVID and the Yuma Project Reservation Division, with negative differences applied as a reduction to MWD's own priority-4 apportionment (Materials and Methods). Because the rule operates on measured consumption rather than on reported savings, the overstatement did not result in additional deliveries to MWD or draw additional water from Lake Mead. SI Figure S13 traces measured use, the counterfactuals implied by our estimates and by the Verification Reports, and the resulting deliveries across the program window.

Because Section 4.d caps the combined consumptive use of MWD, PVID, and Yuma at 420,000 AF, the incidence of the measurement gap we document falls within this group and leaves other users of the Lower Colorado River unaffected. Specifically, errors in the estimated savings are borne by MWD, which pays PVID farmers a fixed price per fallowed acre and receives whatever the priority block (comprised of PVID + Yuma) does not consume. The 624,916 AF gap is therefore water the district paid for and expected to receive, but did not. One way to frame the cost of the error is to ask what MWD would have to spend to replace it from another source. Valued at the price of a comparable transaction from the Imperial Irrigation District's transfer the San Diego County

Water Authority (which rose from $272 per acre-foot in 2005 to $680 in 2021 (nominal USD)) the shortfall is worth $306 million, or $490 per acre-foot on average (SI Table S9). Replacing the same volume at the most recent contracted transfer price of $777 per acre-foot would cost $486 million, and replacing it through residential xeriscaping, a demand-side measure MWD can pursue without a transfer partner, would cost $383 million at $613 per acre-foot (Brelsford and Abbott, 2021). The full set of replacement-cost benchmarks is reported in Materials and Methods.

Figure 5: Implications for Water Supply and Cost



Quantity	Value
Panel A: Conservation, 2005–2021 (AF/yr)	
MWD reported savings	84,560
DiD-measured conservation (ours)	47,801
Overstatement (reported − measured)	36,760
as % of reported	43.5%
Panel B: Effective cost per acre-foot ($210M program)	
Per AF reported (MWD)	$146
Per AF conserved (DiD)	$258

Notes: Upper panel: cost per acre-foot of the PVID fallowing program (green = preferred DiD estimate; neutral gray = MWD's reported cost) compared to alternative Southern California supply and demand-management options (light blue = demand-side programs; dark blue = agricultural-water transfers and system-conservation purchases). The IID–SDCWA bar is the 2023 contracted transfer price of $777 per acre-foot; the USBR bar is Reclamation's 2026 price of $874 per acre-foot for system-conservation water from IID (U.S. Bureau of Reclamation and Imperial Irrigation District, 2024). Lower table reports annual means. *Panel A* summarizes conservation over 2005–2021. *Panel B* reports the effective cost per acre-foot at the $210M program cost.

Panel B of Figure 5 reports the implications for the cost of water actually conserved by the program, and the chart in the same figure places these findings in the context of alternative ways of addressing water scarcity in the Lower Colorado River Basin. Our estimates imply a cost of $258 per acre-foot conserved vs. the $146 per acre-foot implied by MWD's reported savings, calculated by dividing the program's total cost by the two different estimates of water savings. This increase leaves the program's position among alternative supply and demand-management strategies broadly unchanged (see, e.g., Avila et al. (2025)). E.g., even though the program was roughly $100/AF more expensive than projected, it is still cheaper than most alternative water supply augmentation or demand management strategies.

Discussion

Our results indicate that reductions in consumptive water use due to the PVID-MWD Forbearance and Fallowing Program are substantially lower than the estimates from the verification methodology employed by MWD, USBR, and PVID. Crucially, the VR method applies district-wide average water use per acre and assumes 100% reductions in consumptive use. Using remote sensing to develop field-scale estimates of ET allows us to improve on this methodology in two key ways.

First, field-scale consumptive use estimates provide a more accurate benchmark for potential savings under the assumption of 100% reduction. Whereas USBR's approach quantifies consumptive use as a residual difference between diversions and return flows at the district level, remote sensing of ET directly targets the quantity of interest, which is water that is evaporated or transpired and leaves the watershed. In our setting, where effective precipitation is millimeters per year and soils are naturally dry, direct measures of ET can reliably track on-field consumptive use. Holding constant level differences in ET vs. other methods, field-scale estimation facilitates measurement of average water use on the subset of lands enrolled in the program, providing a more accurate measure of potential savings. This compositional difference between parcels could be re-scaled to be consistent with USBR's district-wide estimates by agencies uncomfortable with the average differences between ET and traditional measurement methods.

Second, field-scale water use estimates enable researchers to use causal inference techniques to assess the performance of the program itself. In this study, reductions in water use causally related to the program are just 54% of reported savings and 53% of ET-derived baseline use, as opposed to the assumed 100%. Half of this gap is likely explained by the strategic selection by farmers of which field to fallow in response to calls—called fields have lower predicted water use on average, suggesting that some water savings are not "additional" to the program. The other half is explained by persistent ET on fallowed lands. Much of this ET is driven by spillovers from neighboring irrigated parcels.

These findings relate to a long-standing debate in water policy and hydrology about how to conceptualize the potential water savings from changes in agricultural practices including both efficiency enhancements and fallowing (Frederiksen and Allen, 2011; Gleick et al., 2011; Frederiksen et al., 2012). Ultimately, the question is whether the consumptive water use that persists on fallowed fields should be deducted from savings estimates or not. Answering that question requires grappling with inherent difficulties associated with scaling changes in field-scale water use to district or basin-level water savings (Lankford et al., 2020). However, regardless of how one conceptualizes this number, our results indicate that half of the water savings gap is explained by behavior, not spillovers. This factor alone would account for approximately 300,000 AF of water over the life of the program.

Our results underscore how the design and evaluation of various environmental pay-for-practice programs—deforestation and development avoidance, afforestation, crop switching, and fallowing—can be improved with the use of remote sensing to develop contract-scale estimates of both behavioral compliance *and* ultimate environmental benefits. Doing so could help close

the gap between expected and realized benefits in these programs. Contract-scale benefit estimation allows managers to more accurately forecast program savings by accounting for baseline differences in land use across enrolled vs. unenrolled parcels, where they were previously blind to these differences. Perhaps more crucially, field-scale estimates enable the use of causal program evaluation to reliably estimate both behavioral compliance and environmental benefits ex post, facilitating program tweaks to address non-additionality, compliance gaps (for example through contingent contract terms), and potential spillovers.

Our study has several key limitations. First, our study area likely represents a "best-case scenario" for using total ET to measure consumptive agricultural use, because effective precipitation and natural ET are negligible in PVID. In other settings such as the Central Valley of California, techniques to separate natural ET from applied ET are more essential (Boser et al., 2024). Second, our setting benefits from a lack of groundwater pumping, making it straightforward to relate change in ET to changes in surface water use. In many other settings, it would prove more difficult to relate changes in ET to the supply of irrigation water, which may limit the utility of ET in facilitating transfers of water rights tied to specific sources.

Materials and Methods

Field-Level Water Use Estimation

We combine two data sets to develop field-scale estimates of water use across PVID. The first is monthly data on evapotranspiration from OpenET (Volk et al., 2024). These data are monthly raster files with measures of total evapotranspiration at a 30-meter resolution. We utilize the OpenET Ensemble mean as well as the six individual ET models that make up the mean. The constituent models include: ALEXI/DisALEXI v0.0.32, eeMETRIC v0.20.26, geeSEBAL v0.2.2, PT-JPL v0.2.1, SIMS v0.1.0, and SSEBop v0.2.6.

We obtained polygons of irrigated fields from the Metropolitan Water District (MWD). These polygons were created by PVID to track "water toll acres" that actually receive deliveries of Colorado River Water. We used the ArcGIS zonal statistics tool to calculate monthly mean ET for each field over 2000 to 2021. Monthly means are depicted in Panel (a) of Figure S1, and annual totals are depicted in Panel (b). Both panels also include MWD and USBR's corresponding estimates of total water in each time period. MWD's estimates are derived from annual Verification Reports on the fallowing program published by MWD. USBR estimates come from the USBR's annual "River Accounting" reports.[1] Because field-level estimates are not possible under these methods, district-wide averages provide the best comparison.

Following Volk et al. (2024), who compare OpenET estimates to ground-truth ET estimates from eddy covariance towers, we compare monthly total ET in PVID from each model to MWD's estimates of consumptive use. We estimate the following model to evaluate the alignment of each ET estimate with MWD's estimates:

$$MWD_{mt} = \beta_1 ET_{mt} + \varepsilon_{mt} \tag{1}$$

where MWD_{mt} is MWD's report estimate of consumptive use in month m in year t, ET_{mt} is total ET in month m in year t (summed across all fields in our sample), and ε_{mt} is an error term. Results are presented in Table S2. Regardless of model, the R-squared exceeds 0.89 and the uncentered slope is near 1.

Total ET vs. Agricultural Use

We take two additional measures to probe the relationship between raw total ET and consumptive agricultural use in PVID. First, we assess the extent to which precipitation may confound estimates of consumptive agricultural use. To do so, we download monthly data on total precipitation at a 400-meter resolution from PRISM. We calculate total precipitation and "effective precipitation"—defined as 0.65×precipitation (Wobus et al., 2024) for each parcel in each month of the sample. Using these data, Figure S2 illustrates that effective precipitation is a small percentage ($< 2\%$) of total ET (panels a and b) and that it varies inversely with the irrigation season (panel c). We

[1] https://www.usbr.gov/lc/region/g4000/wtracct.html

also decompose the variation in precipitation into its cross-sectional and time-series components (panel d). Nearly 80% of the variation in precipitation is monthly, with under 20% being parcel-specific. Less than 2% of the overall variation in precipitation is unexplained by parcel and month-of-sample fixed effects, suggesting it is unlikely to affect our results.

The agricultural ET decomposition follows Boser et al. (2024), who train a pixel-level machine-learning model to predict the agricultural component of evapotranspiration ($\widehat{\text{ET}}_{ag}$) and total predicted ET ($\widehat{\text{ET}}$) from soil quality and water-holding capacity (gNATSGO), elevation, slope, and aspect (USGS DEM), reference and potential evapotranspiration, and seasonal climate covariates. Their training set consists of pixels confirmed to be fallow in a given year, identified by intersecting the USDA Cropland Data Layer with the California Department of Water Resources crop maps; the non-agricultural residual is then computed as actual ET minus $\widehat{\text{ET}}_{ag}$ at each pixel-month.

We adapt the Boser et al. (2024) data-preparation pipeline to our setting and re-run the prediction step over our study period. Specifically, we use all of the same state-wide geophysical and satellite inputs, use our polygon layers to specify the fields in our study area. To avoid contamination of training data with the fallowing program we want to assess, we train the model in nearby Imperial Irrigation District (IID) and average over all available years. We apply the trained model to all PVID pixels for 2001–2021. Pixel-level predictions are aggregated to PVID parcels by spatial intersection and converted from millimeters per month to acre-feet per acre per month to match our master panel.

While this is a potentially useful method for isolating agricultural consumptive use, we caution that this approach was developed in the Central Valley of California and may not be appropriate for our setting for several reasons. First, there is far less scope for naturally occurring ET in Palo Verde due to limited precipitation and limited shallow groundwater. Second, as we demonstrate elsewhere in the paper, there is evidence of significant spillovers of water between fallow and non-fallow fields in this region where flood irrigation is common and soils are sandy. Hence, the ET signature on fallow fields that the Boser et al. approach attributes to natural factors in the Central Valley may actually be policy-relevant spillovers in our setting.

Figure S14 and Table S11 provide evidence that the Boser et al. method may not perform as well in PVID. Consistent with the limited scope for natural ET in this setting, district-wide consumptive use reported by MWD is much more aligned with Ensemble ET aggregates than with estimates of Agricultural ET produced with the Boser et al. method. This finding is consistent with the conjecture that the Boser et al. method is purging spillovers that are true agricultural water use, rather than natural ET from precipitation or groundwater. Nevertheless, we perform the Boser et al. calculation for each year of our sample and re-estimate our preferred specification (described below) on the transformed ET estimates.

Fallowing Calls

We obtained our data on which fields were selected for participation in fallowing "calls" each year from annual Verification Reports published by MWD. These reports contain information about

MWD, USBR, and PVID's monthly estimates of total water use and total fallowed acreage for each year of the program (referred to as the VR method), as well as two maps of fields that were selected for fallowing by farmers in response to that year's call (one in summer and one in winter). An example map is depicted in Figure S3. PVID created these maps by hand with a printed aerial map of the district and an overhead transparency—neither MWD nor PVID possessed spatial data describing fallowed fields. We were able to obtain verification maps for program years 2007–2021. We therefore exclude 2005 and 2006—the first two years of the program—from our sample.

We geo-referenced both maps from all available verification reports and overlaid these images with the irrigated fields shapefile obtained from MWD. We tasked three separate research assistants to independently code "called" fields from each image. We also had the research assistants create indicator variables for instances where the shaded area on the map represented a partial subset of a polygon in the shapefile, or when poor image quality made it difficult to determine whether a specific polygon was part of a call. From these underlying data, we create indicators for each field in each map to express whether one, two, or all three research assistants indicated that a field was i) called, ii) partial, and iii) of dubious quality. Thus, we have three separate definitions of treatment, three separate notions of which fields were only partially fallowed, and three separate notions of image quality—all of which we leverage to define treatment and our sample.

The final step in creating a month-year panel of fallowing calls is to interpolate from the images in the verification reports to the intervening months. Typically, each annual verification report contains two images—one in July and one in January/December. From these snapshots, we must infer how to assign treatment to parcels in February-July and August-November. This is made possible by the structure of the program: MWD issues a fallowing call near the end of each calendar year that takes effect the following August and runs for two full years. E.g., a fallowing call issued in December 2012 would correspond to required fallowing from August 2013 through July 2015. With this logic, we can construct "fallowing spells" and interpolate forward and backward from a given image. For each image, we determine candidate 24-month windows that it could be a member of, and then assign treatment within the associated window.

This method is imperfect because the two-year spell duration means that there are two potential spells that could cause a given field to appear as called in the monthly maps we observe. However, the relative frequency of the snapshots allows us to discipline potential over-coding by forcing our assignment to agree with previous and subsequent images. We do this for each version of treatment described above. Figure S4 depicts monthly called acreage implied by our approach, in comparison to reported fallowing by MWD in the Verification Reports. The figure depicts three different conventions for aggregating RA coding into treatment based on whether one, two, or three RAs coded a field as treated. The most inclusive rule that assigns treatment if *any* RA selected a field results in an over-estimate of called acreage in many months, while the restrictive version requiring unanimous agreement tends to result in an under-estimate. Across most months, we obtain close alignment with MWD's official records using the majority rule that treats a field as "called" if at least two RAs coded it as such. We therefore adopt this majority rule

as our preferred definition of treatment.

Other Data

We use several other supplemental data sets in our analysis. We estimate distance to the Colorado River as well as the mean of the soil drainage index of Schaetzl et al. (2009) for each field in the sample. We overlay fields with PLSS sections—square-mile grids that comprise the land survey system—to enable us to cluster at a coarser spatial unit than individual fields.

We also obtained assessor parcel data from Riverside County.[2] Irrigated field polygons were spatially joined to the Riverside County assessor parcel layer, with each field assigned to the parcel contributing the largest intersection area. We then grouped parcels on their cleaned assessor mailing address (street + city, upper-cased), which serves as a proxy for common ownership across Assessor Parcel Numbers. The resulting crosswalk covers 2,500 of the 2,711 irrigated fields in our ET panel, linked to 302 unique owners, of whom 175 own more than one field (127 are single-field owners). The 211 unmatched fields have ET measurements but no owner, so they enter the difference-in-differences estimation and drop out only of the owner-level selection analysis. This is an imperfect proxy for "farms" in at least three ways. First, it is a static snapshot of land ownership in 2025. Given that our panel begins in 2000, parcel ownership has likely changed over time. Second, if farmers lease acreage in or out, land ownership may not correspond to farm operations. Third, some farms may have distinct legal owners on paper (e.g., separate LLCs) but ultimately be jointly operated by the same entity or family. Still, these data form a useful proxy for farm ownership when unpacking the extent of selection into fallowing within vs. between landowners.

We also used USDA Cropland Data Layer (CDL) 30-m rasters (2007–2021) to produce per-field pixel count histograms by USDA crop code. Each field-year was assigned a dominant crop from the pixel counts for cotton, alfalfa, hay, and fallow. Wheat-dominant fields and CDL years 2008 and 2013 are dropped because nearly all PVID fields are misclassified in those cases. The resulting panel has 35,217 field-years across 2,709 fields and 13 years (2007–2021 excl. 2008, 2013), with 60.5% alfalfa, 24.9% fallow, 10.2% cotton, 2.0% hay, and 2.4% other. We then merge this crop panel onto the field panel used in the analysis. This field$\times$year crop panel is then used to assign each onset field its most recent non-fallow pre-call crop and to define the modal crop for never-called fields. An important limitation of these data is that they are not available until 2007, two years after the program began.

Finally, we obtain hourly wind speeds from the Iowa Environmental Mesonet. Each hourly observation is decomposed into meteorological vector components ($u = -\text{sknt} \cdot \sin\theta$, $v = -\text{sknt} \cdot \cos\theta$), averaged by calendar month, and recombined into a prevailing direction via $\text{atan2}(-\bar{v}, -\bar{u})$ converted to a compass bearing; a consistency index is computed as the magnitude of the monthly resultant vector divided by the mean scalar speed ($0 =$ chaotic, $1 =$ unidirectional). The resulting panel has 288 station-months (24 years $\times$ 12 months, 2000–2023), with a mean scalar speed of $\sim$7.0

[2] https://gisopendata-countyofriverside.opendata.arcgis.com/datasets/CountyofRiverside::parcels-crest/about

knots, mean consistency of ∼0.49, and a mean prevailing direction of ∼251° (WSW). This approach is depicted in Figure S11.

Summary statistics for core variables used in the analysis are reported for the enrolled vs. unenrolled group in Table S1.

Difference-in-Difference Estimation

We estimate the causal impact of a fallowing "call" using a difference-in-difference framework that estimates changes in ET on called fields before vs. during a call compared to changes in uncalled fields over the same time period. Recent advances have demonstrated that classic two-way fixed effects regressions can result in biased difference-in-difference (DiD) estimates when treatment is staggered (different units are treated at different times) or when treatment effects change over time. A variety of estimators have been proposed to address this problem. We employ the estimator proposed by De Chaisemartin and d'Haultfoeuille (2020), which constructs a weighted average of all valid comparisons of switchers to non-switchers in the data.

We prefer the De Chaisemartin and d'Haultfoeuille (2020) approach to the prevailing alternatives for several reasons. First and foremost, to our knowledge, this is the only modern DiD estimator that is able to deliver causal estimates when treatment can switch on and off, which it does frequently in our setting of rotational fallowing. Second, the double-robust implementation of the De Chaisemartin and d'Haultfoeuille estimator constructs a propensity-score weighted average of comparisons that is explicitly designed to address potential selection into treatment—a major concern in our setting. The identifying assumption behind the De Chaisemartin and d'Haultfoeuille is that there are parallel trends in untreated outcomes between the treated and untreated groups. While the potential untreated outcome for the group that ultimately receives treatment cannot be observed, De Chaisemartin and d'Haultfoeuille argue that placebo tests to rule out differential pre-trends provide evidence in support of the identifying assumption of the estimator. After discussing our implementation of the estimator, we present the results of these tests.

Ultimately, we are interested in comparing *annual* water savings to MWD's estimates. However, because fallowing calls start in August and end in July two years later, treatment status can vary within a calendar year. We therefore re-aggregate the monthly data into "program years" that run August–July. We implement the dynamic version of the De Chaisemartin and d'Haultfoeuille (2020) estimator, which allows treatment effects to depend on past treatment status, and also accommodates owner-by-year fixed effects.

We apply the estimator to our setting as follows. Let i index PVID fields, t index program years (Aug–Jul), and $o(i)$ denote the assessor-inferred owner of field i. Let Y_{it} be annual ET in acre-feet per acre, and let $D_{it} \in \{0, 1\}$ be the staggered call indicator. For switching fields, define the onset year $F_i = \min\{t : D_{it} = 1\}$ (with $F_i = \infty$ for never-called fields). The De Chaisemartin and d'Haultfoeuille dynamic estimator reports, for each post-onset horizon $\ell = 1, \ldots, 6$, the cohort-

weighted average of within-owner switcher-vs.-stayer comparisons,

$$\widehat{\delta}_\ell \;=\; \sum_{g} w_{g,\ell} \underbrace{\Big\{ \mathbb{E}\big[\, Y_{i,g+\ell} - Y_{i,g-1} \;\big|\; F_i = g,\; o(i) \big] \;-\; \mathbb{E}\big[\, Y_{i,g+\ell} - Y_{i,g-1} \;\big|\; F_i > g+\ell,\; o(i) \big] \Big\}}_{\text{DID}_{g,\ell} \text{ within owner cell}}, \quad (2)$$

where the outer expectation averages over owners o and the weights $w_{g,\ell}$ are proportional to the number of switchers in cohort g observed at horizon ℓ.

The inclusion of owner-by-year fixed effects ensures that every comparison in (2) is executed within an owner $\times$ program-year cell, so that $\widehat{\delta}_\ell$ differences out owner-specific shocks of arbitrary shape. The placebo leads $\widehat{\pi}_k$ for $k = 1, \ldots, 6$ replace $Y_{i,g+\ell} - Y_{i,g-1}$ in (2) with $Y_{i,g-k} - Y_{i,g-k-1}$, allowing us to test for parallel trends by comparing pre-onset outcome changes between switchers and not-yet-treated fields in the same owner cell. Standard errors are clustered at the PLSS section level. Our estimates and subsequent calculations in the main text focus on the average total effect

$$\widehat{\delta}^{\text{ATT}} \;=\; \frac{1}{6} \sum_{\ell=1}^{6} \widehat{\delta}_\ell \quad (3)$$

The preferred specification emphasized in the main text utilizes the majority convention for coding treatment (a field is "called" if at least two RAs coded it as such) and a more restrictive sample quality criterion that drops fields that *any* RA flagged as only partially called or of dubious image quality. We perform this drop at the call-spell level, so if a parcel is ever flagged during a set of months where it is coded as called, we omit it from the entirety of that spell. Table S3 reports the results of estimating Equation 2 on all seven ET models, with the average total effect, individual placebo coefficients, and the p-val associated with the null hypothesis that the placebo estimates are jointly equal to zero. Across four of seven models, we fail to reject the null of no pre-trends at the 10% level, and we fail to reject the null at the 5% level across all seven models.

The sample that identifies our headline estimate is narrower than full ET panel that covers 2,711 fields across 18 program years, or 48,798 field-year observations. We first drop fields whose RA coding places a call across only part of a program year, because annual treatment status is ambiguous for those fields at the frequency at which we estimate, leaving 1,539 fields. We then impose the spell-level quality screen described above, dropping any spell that even one RA flagged as partially fallowed or of dubious image quality, which leaves 1,170 fields and 12,018 field-year observations. Of these, 562 are called at some point during the window and 106 switch treatment status within it. Since the De Chaisemartin and d'Haultfoeuille estimator forms comparisons between fields that switch and fields that do not switch in the same period, these 106 switchers supply the identifying variation, while the remaining fields serve as the comparison group against which their changes are measured.

Robustness

We run 42 alternative specifications to illustrate the stability of our headline estimate and present the results in Figure S5. The specifications span seven OpenET models, three RA call-coding thresholds (any, majority, unanimous), and two control sets (none, +precipitation), holding identification related choices about the estimating sample fixed at their preferred values. Specifically, relaxing the data quality flag threshold destroys parallel trends across nearly every ET model (SI Table S12), and the SUTVA-relevant sample-restriction question is examined directly in SI Table S4 alongside a monthly visualization (SI Figure S8). Results are broadly similar across the 42 specifications, with the preferred specification sitting near the middle of the distribution; full per-cell estimates and placebo diagnostics are reported in SI Tables S13–S15.

Although the significance of the placebo tests varies across specifications, many estimates display an apparent decline in water use on treated parcels prior to the onset of the call. With period $t-1$ differences normalized to zero, this manifests in positive placebo values for periods before $t-1$. We posit that this apparent relative dip in period $(t-1)$ is mechanical in nature. Moving from production to complete fallowing entails a major shift in agricultural operations, and unless this occurs on the final day of a given period, there is likely to be some dip in ET in the period just before fallowing is set to begin. E.g., if a farmer is set to begin fallowing on August 1, unless they wait and clear their field literally overnight on July 31, there will be some decline in ET associated with the foregone production in preparation for fallowing.

We provide evidence on this posited mechanical dip in ET prior to the onset of a call in Figure S6. To do so, we subset the data to about-to-be called fields and never-called parcels and then compare average *monthly* water use across each group in the six months leading up the onset of a fallowing spell. Average monthly water use is closely aligned and not statistically different across groups in January through May, with slight differences emerging in June and a major dip observed for the about-to-be-called group in June and July, consistent with harvesting and preparing fields for fallowing in August.

To understand why this dip emerges as early as May rather than solely in July, we repeat the analysis from Figure S6 separately for alfalfa vs. cotton and present the results in Figure S7. The alfalfa dip does not appear until July, whereas the trend for cotton emerges in May. This reflects differences in the timing of planting across these two crops—farmers are able to grow alfalfa and get a cutting in prior to August, but they must forgo cotton entirely for the season in anticipation of an August call.

To assess whether within-farm reallocation of water onto uncalled fields biases our headline estimate—an identification violation in which control fields are partially treated—we conduct two complementary checks. First, in Figure S8, we plot mean monthly ET around August call onset for three mutually exclusive groups of field-months: fields currently in a clean call window; uncalled fields whose owner had another field in a call window that month; and uncalled fields whose owner had no concurrent call. For the two uncalled groups we drop field-months that occur after any earlier call on the same field, so that post-call carryover does not contaminate the contrast.

Second, we re-estimated the De Chaisemartin and d'Haultfoeuille (2020) dynamic treatment effect on three alternative samples: (i) dropping never-called fields entirely; (ii) dropping uncalled fields owned by ever-called farmers and removing the non-parametric owner-trend absorption (which otherwise mechanically collapses this specification to (i)); and (iii) removing the owner-trend absorption alone (Table S4). Across both checks we find that, if anything, water use also declines on uncalled fields. The alternative point estimates all fall within 22% of the headline. They are modestly larger in magnitude on average, indicating that the headline may slightly understate per-field savings. However, the central finding that observed conservation falls well below the program's reported savings is unchanged.

Comparison to MWD Estimates and Decomposition of Differences

The difference between VR estimates of average water savings and our estimates is driven by several factors. Here, we decompose and characterize each in turn. First, there may be overall level differences in average water use estimates. Even if we employed the VR method of predicting fallowing savings based on district-wide average use, this could cause our estimates to diverge from theirs. Second, the VR method applies a district-wide average, implicitly assuming that average water use is the same across fields enrolled in the program and unenrolled fields. Even if called fields reduce their water use by 100%, compositional differences in enrolled vs. unenrolled fields would lead to difference between the VR estimate and a causal estimate of water savings.

Third, the VR method assumes that water use declines by 100% when a field is called to fallow. Even though MWD does carefully enforce fallowing calls and include images to this effect in their Verification Reports, water savings on called fields may not be 100% of baseline water use—e.g., there may still be a "savings gap." This savings gap can be decomposed into i) residual water use on called fields (water use does not go to zero) and ii) "non-additionality," or the strategic selection of fields that would have had lower water use even in the absence of a call (some fields that were fallowed would have been fallowed in the absence of a call, resulting in zero net savings relative to a counterfactual no-call state).

To be precise, the differences between the VR approach and our approach can be written as:

$$\text{Difference} = S_{\text{MWD}} - S_{\text{Actual}} = (\Omega_{\text{Fallow}} + \Omega_{\text{Add}}) - (\Delta_{\text{Meas}} + \Delta_{\text{Select}})$$

$$S_{\text{Actual}} = S_{\text{MWD}} + \underbrace{\Delta_{\text{Meas}} + \Delta_{\text{Select}}}_{\text{Measurement Differences}} \underbrace{- \Omega_{\text{Fallow}} - \Omega_{\text{Add}}}_{\text{Savings Gap}}$$

Where the components are defined as:

- $\Delta_{\text{Meas}} = ET_{\text{avg}} - S_{\text{MWD}}$ (Measurement adjustment)
- $\Delta_{\text{Select}} = ET_{\text{enrolled}} - ET_{\text{avg}}$ (Enrollment selection bonus)
- $\Omega_{\text{Fallow}} = ET_{\text{residual}}$ (Residual water use)

- $\Omega_{\text{Add}} = (ET_{\text{enrolled}} - \Omega_{\text{Fallow}}) - S_{\text{Actual}}$ (Additionality gap)

Each of these components can be calculated directly from the data to better understand why the causal estimate differs from the VR estimate by roughly 50%. As a practical matter, Δ_{Meas} and Δ_{Select} are quite small. The VR average annual savings estimate is 4.67 AF per fallowed acre, whereas we estimate district-wide average water use to be 4.63 AFA, implying that $\Delta_{Meas} = -0.04$. As depicted in Figure S9, average water use on enrolled parcels is slightly but systematically higher than on unenrolled parcels. Average annual water use on enrolled parcels is 4.77 AFA, implying that $\Delta_{Select} = 0.14$, also quite small.

After accounting for measurement differences and selection into the program, potential water savings from ET-based measurement are 4.77 AFA per year. The causal estimate of 2.53 implies a 2.24 AFA savings gap which can be further decomposed into Ω_{Fallow} and Ω_{Add}. Moreover, Ω_{Add} can be further decomposed into a measurable between-farmer component, a measurable within-farmer component, and an unexplained residual component. The next section describes our approach to quantifying each component of the savings gap.

Breaking Down the Savings Gap

Our approach to unpacking the savings gap is to first focus on directly measurable residual water use on called fields. After documenting non-trivial water use on called fields, we explore several potential explanations. We then turn our attention to the remaining portion of the gap and assess the extent to which modeling predicted water use on about-to-be-called fields can explain the remainder of the gap that we have attributed to non-additionality. The residual from these two exercises is non-additionality driven by factors that our model cannot predict (such as private farmer information). For both exercises, we focus our attention on enrolled parcels only.

Residual ET

Panel (a) of Figure 3 depicts the distribution of average water use across called fields in each year for which we have call data. A significant mass of called fields have non-zero ET, with an overall mean of 1.18 AFA across fields and years. Hence, roughly half of the savings gap can be explained by residual ET on called fields. Contrary to the VR assumption that water use declines by 100% in response to a call, water use remains roughly 25% of baseline even after a field is called.

Figure S10 presents several variables that could explain residual water use: distance to the Colorado River and the drainage index of Schaetzl et al. (2009), as well as the cumulative length of a given fallowing spell. None of these variables has significant power in explaining the variation in ET on called fields. We also explore whether spillovers from irrigation on neighboring parcels seep onto fallowed lands. We estimate average ET across each fallowed parcel's neighbors on a monthly basis and provide a binned scatter plot in panel (b) of Figure 3.

Table S5 reports the results of several regressions of residual ET on each potential explanatory variable depicted in Figure S10, as well as water use on neighboring fields (depicted in panel b of

Figure 3). There is a strong positive relationship only between neighbor ET and own ET, with a slope of about 0.34, suggesting that surface spillovers and seepage are a major driver of residual ET. Average annual neighbor water use for called fields is 4.43 AFA. Hence, the linear prediction for ET on a fallow field would be 0.34×4.43 = 1.5, implying that the magnitude of the estimated spillovers could fully explain residual ET on called fields.

One possible concern is that these estimated neighbor effects could reflect measurement error in ET itself, rather than ground-level spillovers. ET models work by using variation in surface temperature to infer evaporation and transpiration associated with evaporative cooling. It may be that there are low-level atmospheric spillovers associated with evaporative cooling, such that the air around a fallow field is cooler when its neighbors are evaporating more, leading to a false ET signature.

We assess the plausibility of this alternative mechanism by exploiting variation in wind direction. The intuition is that evaporative/atmospheric spillovers should be stronger from upwind neighbors than from downwind neighbors. If neighbor effects are not correlated with wind direction, it suggests the effects are indeed ground-level spillovers. We calculate prevailing winds for each month in the sample by obtaining hourly wind speeds and directions and constructing a vector average for each month. We then construct a field's average upwind vs. downwind ET by taking the cosine of the difference between the angle between two parcel's centroids and the angle of the prevailing wind for a given month. The assigned weight is then the maximum of the cosine and zero. This means that a parcel directly downwind of its neighbor would receive a weight of one, whereas a parcel at exactly 90 degrees would obtain a weigh of zero. Panel (c) of Figure 3 illustrates.

Panel (d) of Figure 3 demonstrates that neighbor spillovers are not correlated with wind direction. This can be seen visually in the figure, which plots a fallow parcel's own ET against its upwind vs. downwind neighbors. We confirm this by regressing fallow ET on total neighbor ET and upwind vs. downwind neighbor ET. Table S6 presents regression results which confirm that upwind and downwind neighbors do not have a differential effect on called parcels' ET. The p-value on the hypothesis test that upwind and downwind ET matter equally ranges from 0.643 to 0.845. Moreover, the sum of upwind and downwind effects is roughly equal to the total effect.

A non-rejection of the upwind=downwind null is consistent with either the absence of an atmospheric channel or with the test being underpowered, so we quantify what the design could have detected. Taking the cluster-robust standard error of the upwind−downwind contrast from column (3) of SI Table S6, the minimum detectable effect (MDE) at conventional thresholds is $\text{MDE} = (z_{1-\alpha/2} + z_{1-\beta})\,\widehat{\text{SE}}(\hat{\beta}_{\text{up}} - \hat{\beta}_{\text{down}})$. At 80% power and $\alpha = 0.05$ (two-sided), this yields an MDE of 0.024 AFA, or roughly 5.9% of the total neighbor-ET slope; the observed contrast is 0.004 ($p = 0.64$). The non-rejection therefore bounds any atmospheric contribution to the neighbor spillover well below the magnitude that would be required for it to be the dominant mechanism. The full set of MDEs across (α, power) combinations is reported in SI Table S7.

Non-Additionality

To isolate the mechanisms driving non-additionality in the water fallowing program, we estimate the extent to which farmers strategically select specific fields for enrollment based on heterogeneous water-use expectations. The total observed non-additionality gap—the difference between the total savings gap (2.24) and observed residual ET on called fields (1.18)—is 1.06.

We decompose this 1.06 AFA gap into three distinct factors: (1) observable between-farmer selection, (2) observable within-farmer selection, and (3) private, unobservable transient information. We achieve this by horse-racing a baseline autoregressive model against a cross-validated machine learning approach, and algebraically partitioning the resulting selection effects.

We first estimate a counterfactual ET for all field-months using an AR(1) specification on historically uncalled fields. This model captures baseline field-specific seasonality and mean reversion:

$$ET_{imt} = \alpha_{im} + \beta(ET_{i,t-12} - \alpha_{im}) + \varepsilon_{imt} \tag{4}$$

where α_{im} represents field-by-month fixed effects. This specification allows every field to exhibit a unique seasonal water-use curve, adjusting for a 12-month lag to avoid seasonal confounding.

To ensure our estimate of structural selection is not downwardly biased by the functional form assumptions of the AR(1) model, we construct an upper bound on observable predictability using a cross-validated LASSO model. This approach includes the field's seasonal baseline as an unpenalized regressor, extended 24-month lags to capture multi-year crop rotations, spatial neighbor histories (average 12-month lagged ET of adjacent fields), and month-by-year fixed effects to capture district-wide weather or macroeconomic shocks. The machine learning model is tuned via 5-fold cross-validation to select the optimal penalty parameter, preventing overfitting while maximizing out-of-sample predictive power. The regression coefficients reported in Table S8 confirm that predicted counterfactual ET is a statistically significant predictor of the call decision. Fields with lower predicted water use are more likely to be selected for a call, both between and within owners, across both the AR(1) and LASSO models. Overall predictive accuracy of the two models is broadly comparable—the LASSO does not dramatically improve out-of-sample fit relative to the AR(1)—but the additional dimensions it captures (long-horizon cycles, spatial structure, aggregate shocks) are precisely those that correlate with the call decision. The decomposition below makes this concrete.

For both models, we aggregate the predicted counterfactual ET to the program-year level to quantify *total predicted non-additionality* as the difference between average enrolled water use and either AR(1) or LASSO-predicted water use on called fields each year. Let $\hat{Y}_{ift}$ denote the predicted counterfactual ET for field i in owner f's portfolio in program-year t; let $C_{ift} \in \{0, 1\}$ indicate whether the field is called for fallowing in that year. The total observable selection effect is the difference between the predicted ET on uncalled and called field-years among ever-enrolled fields:

$$\Delta_{\text{total}} \;=\; \mathbb{E}\Big[\hat{Y}_{ift} \mid C_{ift} = 0\Big] \;-\; \mathbb{E}\Big[\hat{Y}_{ift} \mid C_{ift} = 1\Big]. \tag{5}$$

We can then decompose the total predicted non-additionality exactly into a between- and a within-owner component. Let $\bar{\hat{Y}}_{ft}$ denote the owner-year portfolio average of $\hat{Y}$ across owner f's enrolled fields in year t. Adding and subtracting the owner-year portfolio mean conditional on a call partitions Δ_{total} exactly into two interpretable margins:

$$\Delta_{\text{between}} = \mathbb{E}\left[\hat{Y}_{ift} \mid C = 0\right] - \mathbb{E}\left[\bar{\hat{Y}}_{ft} \mid C = 1\right], \tag{6}$$

$$\Delta_{\text{within}} = \mathbb{E}\left[\bar{\hat{Y}}_{ft} \mid C = 1\right] - \mathbb{E}\left[\hat{Y}_{ift} \mid C = 1\right], \tag{7}$$

$$\Delta_{\text{total}} = \Delta_{\text{between}} + \Delta_{\text{within}}. \tag{8}$$

Δ_{between} compares the enrolled-baseline mean of predicted ET (uncalled field-years) to the average owner-portfolio mean across called field-years. Each called field-year contributes equally to that average, so owners enter the called-side expectation in proportion to the number of called fields they have. Because the program calls a uniform fraction of every enrolled owner's land in a given year rather than activating some owners and not others, this margin does not identify year-to-year owner-level participation. It captures the compositional effect of differential enrollment intensity—owners who enrolled a larger absolute share of their land contribute more called field-years. Hence, the between-owner component is driven by any systematic relationship between enrollment intensity and portfolio ET.

Δ_{within} is the average gap, within an owner-year, between that owner's full portfolio mean and the predicted ET of the specific fields the owner called. It captures the micro-level strategic sorting of land within a farm.

The residual non-additionality not absorbed by a given prediction model is

$$\Delta_{\text{unobs}} = \text{Gap} - \Delta_{\text{total}},$$

where Gap $= 1.06$ AFA is the share of the total savings gap remaining after subtracting residual ET on called fields. Under the LASSO prediction, Δ_{unobs} serves as a bound on the contribution of farmers' private information (beyond what we condition on) to selection.

Applying the two sets of predictions to equations 5–6 reveals three distinct tiers of adverse selection. The baseline AR(1) model explains 0.44 AFA (panel (c) of Fig. 4), or roughly 42% of the 1.06 AFA non-additionality component of the savings gap. This indicates that farmers systematically rely on basic historical performance when making fallowing decisions, selecting fields with historically lower baselines. This effect is driven primarily by within-owner field sorting (0.37 AFA) rather than between-owner portfolio selection (0.07 AFA).

The LASSO model explains 0.73 AFA (panel (d) of Fig. 4), or roughly 69% of the 1.06 AFA non-additionality component of the savings gap. The improvement is heavily concentrated in the between-owner margin, which more than triples from 0.07 AFA to 0.23 AFA, while the within-owner margin rises more modestly from 0.37 AFA to 0.50 AFA. The growth in the between-margin does not reflect a year-to-year participation decision (every enrolled owner is called for the same

fraction of their enrolled acreage), but rather the LASSO's improved ability to capture owner-level structural variation in ET—through multi-year lags, neighbor histories, and aggregate month-by-year shocks—that the AR(1) cannot. Mechanically, this owner-level predictable variation enters the between-margin because owners enrolled different absolute shares of their land in the program. Owners with larger enrolled portfolios contribute more called field-years, so any correlation between enrollment intensity and portfolio ET manifests as between-owner selection.

Subtracting the LASSO prediction from the true causal gap leaves a residual unobservable non-additionality of 0.33 AFA, or roughly 31% of the gap. Because this remainder cannot be predicted even by a flexible machine learning algorithm armed with extensive spatial and historical covariates, it serves as an upper bound—conditional on the covariate space we consider—on the contribution of farmers' private information to selection. Approximately one-third of the program's non-additionality therefore appears to be driven by factors not recoverable from our historical or spatial covariates and that program administrators relying on the same data would be unable to anticipate.

Aggregate Implications

The estimates in the Results Section imply a sustained gap between realized water savings from fallowing in PVID and the savings MWD and USBR have reported on the same acres. MWD and PVID both divert water from the Colorado River, which raises questions about where within the Lower Colorado River Basin the difference is being realized. Here, we quantify the overall magnitude of that gap and discuss the implications for MWD. All estimates in this section combine the preferred dynamic DiD per-acre savings rate ($\widehat{\delta}^{\text{ATT}} = -2.53$ AFA) with MWD's own contracted-acre-year ledger from the Verification Reports over the program window (program years 2005–2021; 17 years).

We estimate the per-acre effect on program years 2007–2021, where maps exist, and apply it to MWD's reported contracted acre-years across the full program window, which extends back to 2005. The two extrapolated years contribute 40,875 AF, or 6.5%, of the cumulative shortfall reported below. Restricting the aggregate to years with maps yields 584,041 AF, which is 46.9% of reported savings over that window rather than 43.5% over the full one, so extending the calculation to the first two program years is conservative.

Aggregate shortfall. Across this window, the Verification Reports credit the program with **1,437,527 AF** of saved water. Applying our DiD rate to the same acre-year ledger yields **812,611 AF** of physically defensible savings, leaving a cumulative shortfall of

$$\Delta\text{AF} \;=\; \underbrace{1{,}437{,}527}_{\text{MWD reported}} \;-\; \underbrace{812{,}611}_{\text{DiD implied}} \;=\; 624{,}916\ \text{AF}, \tag{9}$$

or **43.5%** of the program's claimed yield. On an annual basis the shortfall averages **36,760 AF per program year**, equivalent to roughly **2.2%** of MWD's $\sim$1.7 MAF total annual deliveries. The

cumulative gap traced out in Figure S12 is exactly this 624,916 AF spike by construction.

Priority accounting and water deliveries While the Verification Reports provide a detailed accounting of MWD/USBR/PVID's initial *estimate* of savings from the program in each year, these numbers do not necessarily translate directly into increased deliveries to MWD. The relationship between consumptive use in PVID and deliveries to MWD are governed by Section 4.d of the 2003 Colorado River Water Delivery Agreement. This agreement specifies that MWD receives the difference between a fixed annual threshold of 420,000 AF and the combined consumptive use of PVID (priorities 1 and 3b) together with the Yuma Project Reservation Division and Yuma Island (priority 2). Writing U_t for that combined use in calendar year t, the quantity credited to MWD is

$$D_t \;=\; 420{,}000 \;-\; U_t, \tag{10}$$

which is delivered to MWD when $U_t < 420{,}000$ and is instead applied as a reduction to MWD's own priority-4 apportionment when $U_t > 420{,}000$. We construct U_t from the consumptive-use estimates reported in the annual USBR Decree Accounting Reports. Our reconstruction of (10) reproduces the deliveries to MWD reported by USBR to the acre-foot, so that in 2016, for example, $420{,}000 - 379{,}111 = 40{,}889$ AF. Over 2006–2020 U_t averages 399,203 AF and D_t averages 20,797 AF per year, or 270,367 AF cumulatively, which is 46% of the 588,109 AF that our DiD estimates attribute to fallowing over the same window. In 2008 and again in 2013–2015, combined use exceeded the threshold and D_t was negative.

SI Figure S13 plots lines for the 420,000 benchmark and for U_t, and the shaded differences between these two lines represent D_t in each year. In addition to this pure replication of the accounting rules, Figure S13 also plots implied counterfactual water use if fallowing had not occurred according to both our causal estimate and the reported savings from the Verification Reports. The rule benchmarks use against a fixed threshold rather than against the water PVID would have consumed absent the program (approximated by the two counterfactuals in the figure), so the official delivery and the conservation we measure differ each year by exactly $420{,}000 - C_t$, where $C_t = U_t + S_t$ is counterfactual use and S_t is our estimated savings. For example, in 2010, counterfactual use was 377,791 AF, with a reduction of 65,653 AF that we attribute to fallowing. Because the rule uses a baseline of 420,000, it credited MWD with 107,862 AF, so roughly 42,000 AF would have reached MWD even with no fallowing at all. In 2013, counterfactual use was 505,957 AF and the rule recorded a 69,528 AF reduction to MWD's apportionment (compared to 16,429 AF we attribute to fallowing) so genuine conservation occurred but none of it was delivered to MWD under the rule because D_t was less than zero.

It is also worth emphasizing that U_t in (10) is consumptive use as measured by USBR in the Decree Accounting Reports. The savings figures published in MWD's Verification Reports do not enter the calculation. Deliveries to MWD are therefore determined by what PVID and Yuma actually consumed rather than by what the program reports having saved, so the overstatement we document cannot by itself have moved water to MWD or drawn additional water from Lake

Mead, and this holds regardless of how the 420,000 AF block would be allocated absent the program. The overstatement bears on what MWD paid rather than on what it received. (10) also implies that in years when U_t falls below the threshold the block's combined draw is 420,000 AF irrespective of how much of it PVID consumes, so fallowing in those years reallocates use within the block rather than adding to it. Verification Report savings do enter the calculation of Intentionally Created Surplus (ICS) credits, which MWD may bank in Lake Mead and withdraw in later years. Because MWD generates credits across many programs and has exceeded the cap for the number of credits they may claim in most years, it is essentially not possible to directly connect savings estimation errors to ICS credit overages.

Financial bound. The first panel of Figure 5 also provides some additional context for these estimates by comparing the VR estimated cost per acre-foot of the program, actual cost per acre-foot implied by our estimates, and cost estimates for alternative supply augmentation or demand management strategies facing municipal providers such as MWD. Estimates of the cost of social nudges come from Bernedo et al. (2014); Price et al. (2014), whereas xeriscaping estimates come from Brelsford and Abbott (2021).

The first way to value the shortfall is to price each program year's shortfall at the price of conserved agricultural water contracted in the region that year. This approach measures what the undelivered water was worth as it accrued. We use the Imperial Irrigation District's transfer agreement with the San Diego County Water Authority as a comparison price. This agreement's price-per-acre-foot schedule was fixed by contract through 2015 and escalated with the GDP implicit price deflator thereafter (Imperial Irrigation District and San Diego County Water Authority, 2009); the resulting series runs from \$272/AF in 2005 to \$680/AF in 2021 and is reported in full in SI Table S9. On this basis the 624,916 AF shortfall is worth **\$306 million**, an average of \$490/AF.

Another approach is to measure the cost of replacing this water *today* by pricing the entire volume at a single current rate. We use five current reference prices to provide a range of estimates under this approach, spanning what MWD already pays, what it could contract for, and what manufactured supply would cost. MWD's *own* program cost (program-cost / claimed-AF $\approx$ \$146/AF) yields a **\$91 million** mismatch between dollars spent and physical water delivered. MWD's Tier 1 untreated rate (\$766/AF) yields **\$479 million**. The most recent contracted IID–SDCWA transfer price (\$776.97/AF for 2023, U.S. Bureau of Reclamation and Imperial Irrigation District, 2024) yields **\$486 million**. Carlsbad desalination, as a marginal manufactured-supply alternative (\$3,000/AF), yields **\$1.87 billion**, which should be thought of as upper bound on opportunity cost, rather than a realized expenditure. Residential xeriscaping, a demand-side measure MWD can pursue without a transfer partner, costs \$613/AF (Brelsford and Abbott, 2021) and yields **\$383 million**.

We separately report Reclamation's price for system-conservation water from IID, \$873.99/AF for 2026 (U.S. Bureau of Reclamation and Imperial Irrigation District, 2024), in Figure 5. That price is escalated from the 2023 transfer price at a fixed four percent per year in lieu of the deflator

adjustment, and it is paid by the federal government rather than by an urban buyer. It is the closest available market comparison to the program we study, because it purchases irrigation from a neighboring district under the same priority system.

References

Arellano-Gonzalez, J., A. AghaKouchak, M. C. Levy, Y. Qin, J. Burney, S. J. Davis, and F. C. Moore (2021). The adaptive benefits of agricultural water markets in California. *Environmental Research Letters 16*(4), 044036.

Aspelund, K. M. and A. Russo (2026). Additionality and asymmetric information in environmental markets: Evidence from conservation auctions. *American Economic Review 116*(8).

Avila, P., M. Nemati, A. Dinar, D. Crespo, Z. Frankel, and N. Halberg (2025, October). Public spending and water scarcity: An empirical analysis of USBR investments in the Colorado River Basin. *JAWRA Journal of the American Water Resources Association 61*(5), e70042.

Ayres, A., K. Meng, and A. Plantinga (2021, October). Do environmental markets improve on open access? Evidence from California groundwater rights. *Journal of Political Economy 129*(10), 2817–2860.

Bernedo, M., P. J. Ferraro, and M. Price (2014). The persistent impacts of norm-based messaging and their implications for water conservation. *Journal of Consumer Policy 37*(3), 437–452.

Boser, A., K. Caylor, A. Larsen, M. Pascolini-Campbell, J. T. Reager, and T. Carleton (2024). Field-scale crop water consumption estimates reveal potential water savings in California agriculture. *Nature Communications 15*(1), 2366.

Brelsford, C. and J. K. Abbott (2021). How smart are 'Water Smart Landscapes'? *Journal of Environmental Economics and Management 106*, 102402.

Bretsen, S. N. and P. J. Hill (2008). Water markets as a tragedy of the anticommons. *Wm. & Mary Envtl. L. & Pol'y Rev. 33*, 723–783.

Brewer, J., R. Glennon, A. Ker, and G. Libecap (2008). 2006 presidential address water markets in the west: prices, trading, and contractual forms. *Economic Inquiry 46*(2), 91–112.

Bruno, E. M. and K. Jessoe (2021). Missing markets: Evidence on agricultural groundwater demand from volumetric pricing. *Journal of Public Economics 196*, 104374.

Bruno, E. M. and K. Jessoe (2024). Designing water markets for climate change adaptation. *Nature Climate Change 14*(4), 331–339.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review 110*(9), 2964–2996.

Dilling, L., J. Berggren, J. Henderson, and D. Kenney (2019). Savior of rural landscapes or Solomon's choice? Colorado's experiment with alternative transfer methods for water (ATMs). *Water Security 6*, 100027.

Elliott, J., D. Deryng, C. Müller, K. Frieler, M. Konzmann, D. Gerten, M. Glotter, M. Flörke, Y. Wada, N. Best, et al. (2014). Constraints and potentials of future irrigation water availability on agricultural production under climate change. *Proceedings of the National Academy of Sciences 111*(9), 3239–3244.

Frederiksen, H. D. and R. G. Allen (2011). A common basis for analysis, evaluation and comparison of offstream water uses. *Water International 36*(3), 266–282.

Frederiksen, H. D., R. G. Allen, C. M. Burt, and C. Perry (2012). Responses to gleick et al., which was itself a response to frederiksen and allen. *Water International 37*(2), 183–197.

Garrick, D. E., S. Balasubramanya, M. Beresford, A. Wutich, G. G. Gilson, I. Jorgensen, N. Brozović, M. Cox, X. Dai, S. Erfurth, R. Rimsaitė, J. Svensson, J. Talbot-Jones, H. Unnikrishnan, C. Wight, S. Villamayor-Tomas, and K. Vazquez Mendoza (2023). A systems perspective on water markets: Barriers, bright spots, and building blocks for the next generation. *Environmental Research Letters 18*(3), 031001.

Gleick, P. H., J. Christian-Smith, and H. Cooley (2011). Water-use efficiency and productivity: rethinking the basin approach. *Water International 36*(7), 784–798.

Gordon, B. L., G. F. Boisrame, R. W. Carroll, N. K. Ajami, B. Leonard, C. Albano, N. Mizukami, M. A. Andrade, E. Koebele, M. H. Taylor, et al. (2024). The essential role of local context in shaping risk and risk reduction strategies for snowmelt-dependent irrigated agriculture. *Earth's Future 12*(6), e2024EF004577.

Grafton, R. Q., G. D. Libecap, E. C. Edwards, R. J. O'Brien, and C. Landry (2012). Comparative assessment of water markets: Insights from the Murray–Darling Basin of Australia and the Western USA. *Water Policy 14*(2), 175–193.

Hagerty, N. (2023). What holds back water markets? Transaction costs and the gains from trade. Working paper, Montana State University.

Imperial Irrigation District and San Diego County Water Authority (2009). Fifth amendment to agreement between imperial irrigation district and san diego county water authority for transfer of conserved water. Amendment to transfer agreement. Dated December 21, 2009, amending the Agreement of April 29, 1998. Section 5.2(e) states the price per acre-foot for Agreement Years 1–13 (calendar years 2003–2015); Section 5.2(f) provides for annual escalation by the GDP Implicit Price Deflator for Agreement Years 14–32. Reproduced as Exhibit A to U.S. Bureau of Reclamation and Imperial Irrigation District (2024).

Lankford, B., A. Closas, J. Dalton, E. L. Gunn, T. Hess, J. W. Knox, S. Van der Kooij, J. Lautze, D. Molden, S. Orr, et al. (2020). A scale-based framework to understand the promises, pitfalls and paradoxes of irrigation efficiency to meet major water challenges. *Global Environmental Change 65*, 102182.

Leonard, B., C. Costello, and G. Libecap (2019, Winter). Expanding water markets in the western United States: Barriers and lessons from other natural resource markets. *Review of Environmental Economics and Policy 13*(1), 43–61.

Medellín-Azuara, J., A. Escriva-Bou, A. C. Gaudin, K. A. Schwabe, and D. A. Sumner (2024). Cultivating climate resilience in California agriculture: Adaptations to an increasingly volatile water future. *Proceedings of the National Academy of Sciences 121*(32), e2310079121.

Price, J. I., J. M. Chermak, and J. Felardo (2014). Low-flow appliances and household water demand: An evaluation of demand-side management policy in Albuquerque, New Mexico. *Journal of Environmental Management 133*, 37–44.

Rafey, W. (2023, February). Droughts, deluges, and (river) diversions: Valuing market-based water reallocation. *American Economic Review 113*(2), 430–471.

Rafey, W. (2026). Measuring water misallocation in California. NBER Working Paper 35176, National Bureau of Economic Research.

Richter, B. D., D. Bartak, P. Caldwell, K. F. Davis, P. Debaere, A. Y. Hoekstra, T. Li, L. Marston, R. McManamay, M. M. Mekonnen, et al. (2020). Water scarcity and fish imperilment driven by beef production. *Nature Sustainability 3*(4), 319–328.

Richter, B. D., G. Lamsal, L. Marston, S. Dhakal, L. S. Sangha, R. R. Rushforth, D. Wei, B. L. Ruddell, K. F. Davis, A. Hernandez-Cruz, et al. (2024). New water accounting reveals why the Colorado River no longer reaches the sea. *Communications Earth & Environment 5*(1), 134.

Rodell, M., J. S. Famiglietti, D. N. Wiese, J. Reager, H. K. Beaudoing, F. W. Landerer, and M.-H. Lo (2018). Emerging trends in global freshwater availability. *Nature 557*(7707), 651–659.

Schaetzl, R. J., F. J. Krist, Jr., K. Stanley, and C. M. Hupy (2009). The natural soil drainage index: An ordinal estimate of long-term soil wetness. *Physical Geography 30*(5), 383–409.

U.S. Bureau of Reclamation and Imperial Irrigation District (2024). System Conservation Implementation Agreement (SCIA) for Calendar Years 2024 through 2026 between the United States Bureau of Reclamation and the Imperial Irrigation District to implement the Lower Colorado River Basin System Conservation and Efficiency Program. Agreement between Reclamation and IID. SCIA No. 24-XX-30-W0825. Section 5.7 states the 2023 IID–SDCWA transfer price of $776.97 per acre-foot; Sections 5.7.1–5.7.3 state Reclamation's prices of $808.05 (2024), $840.37 (2025), and $873.99 (2026) per acre-foot for System Conservation Water.

Volk, J. M., J. L. Huntington, F. S. Melton, R. Allen, M. Anderson, J. B. Fisher, A. Kilic, A. Ruhoff, G. B. Senay, B. Minor, et al. (2024). Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water resource and land management applications. *Nature Water 2*(2), 193–205.

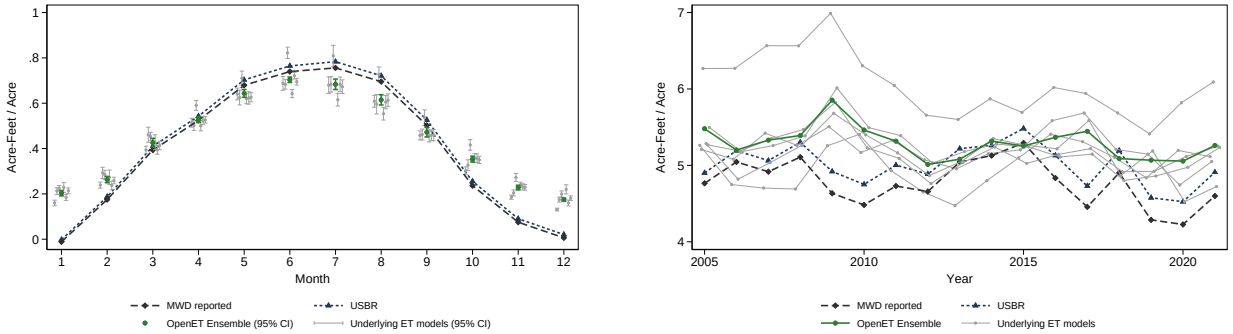
Wobus, C., C. Nash, P. Culp, M. Kelly, and K. Kennedy (2024). Simplified agricultural water use accounting in the Colorado River Basin using OpenET. *Environmental Research Letters 20*(1), 014020.

Wong, C. A., W. S. Abramson Sklarin, M. S. Mauter, and D. B. Lobell (2026, January). Measuring impacts of California agri-environmental programs using field-scale satellite data. *Environmental Research Letters 21*(1), 014006.

Supplementary Material

Figures

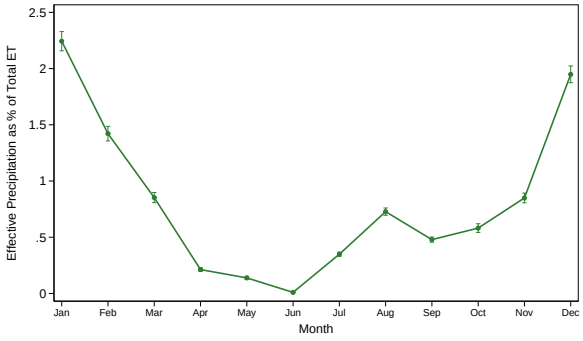
Figure S1: District-Wide Consumptive Use: ET Models vs. MWD Reported



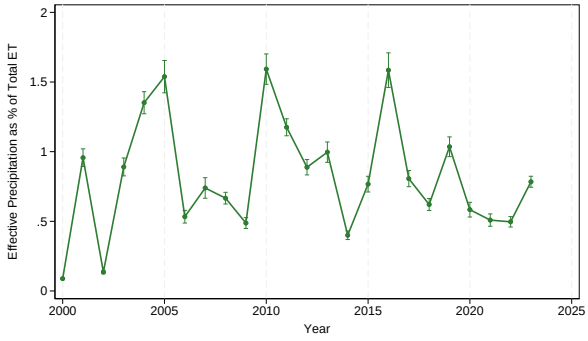
(a) Monthly AF Consumed (District-Wide) (b) Annual AF Consumed (District-Wide)

Notes: District-wide consumptive-use aggregates compare the seven OpenET models (`ENSEMBLE`, `PTJPL`, `SIMS`, `SSEBOP`, `geeSEBAL`, `EEMETRIC`, `DISALEXI`) against MWD's reported consumptive-use aggregated from the annual USBR Decree Accounting Reports and MWD Verification Reports. Panel (a) plots monthly total AF consumed across all PVID fields; panel (b) plots the same aggregated to calendar year. The ensemble and per-model totals track the MWD series closely in the aggregate, motivating the use of field-level ET as the outcome for the DiD. Companion forces-through-origin regressions are reported in SI Table S2.

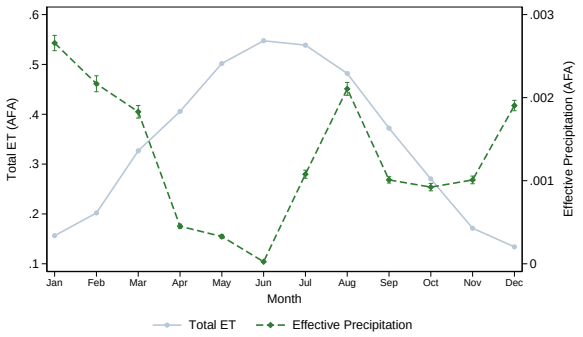
Figure S2: Effective Precipitation on Called Fields



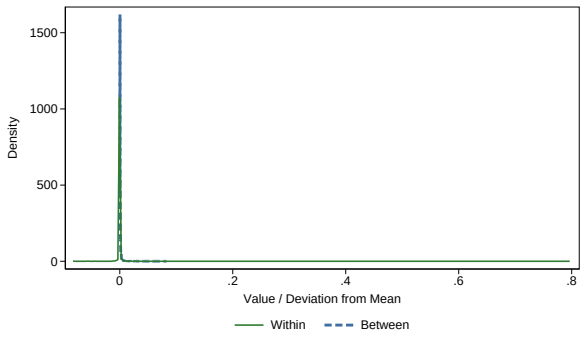
(a) Monthly Precipitation as % of ET



(b) Annual Precipitation as % of ET



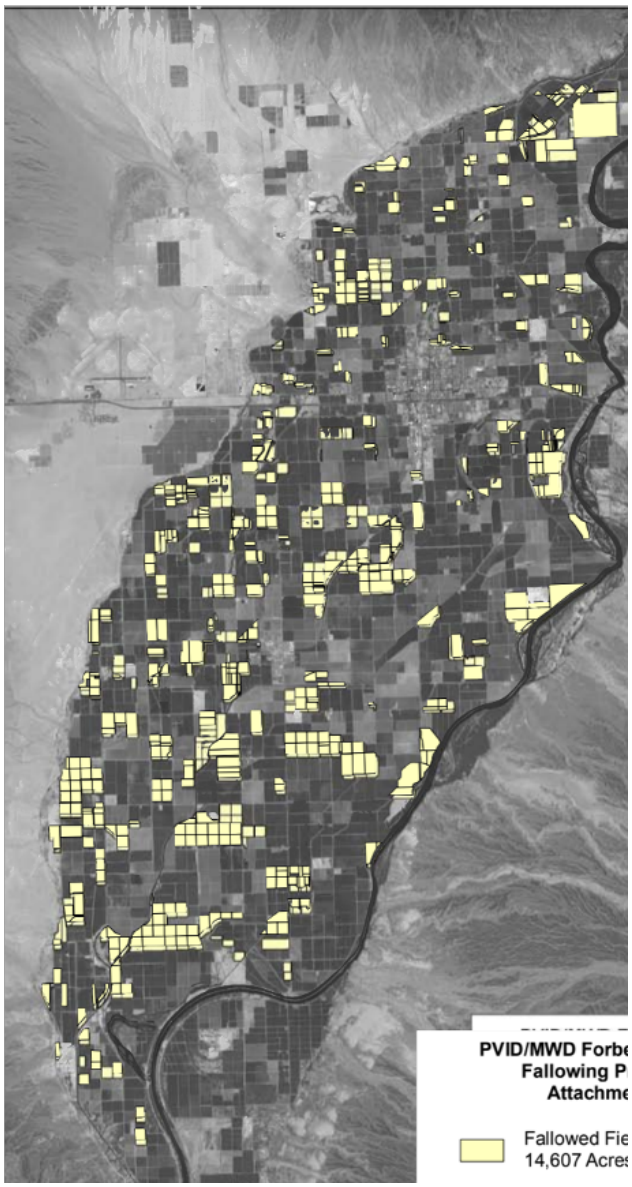
(c) Monthly Precipitation vs. ET



(d) Precipitation Decomposition of Residual ET

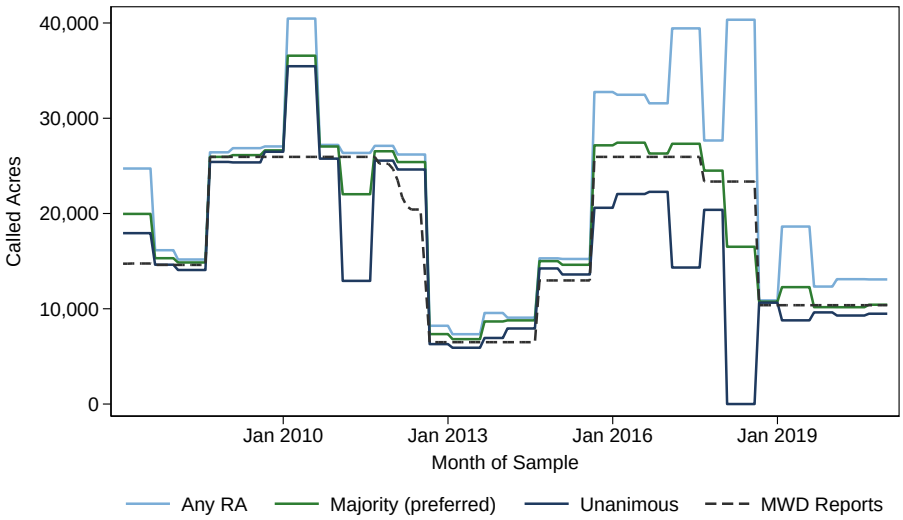
Notes: The PVID service area is hyper-arid (long-run precipitation ∼9 cm/yr from PRISM), so effective precipitation is a trivial contributor to called-field ET. Panel (a) plots monthly PRISM precipitation as a share of monthly ET on called fields, pooled across parcel-months; panel (b) plots the same aggregated to prog-year. Even in the wettest months, rainfall explains less than 10% of observed ET; annual shares are an order of magnitude smaller. Panel (c) is a bivariate scatter of monthly precipitation against ET on called fields, with the best-fit line; the flat slope confirms precipitation does not covary materially with month-to-month ET variation in this sample. Panel (d) decomposes total residual ET on called fields into the portion explained by concurrent precipitation (<5% AFA) and the unexplained remainder, which is the focus of the hydrologic-spillover analysis in the main text (Fig. 3).

Figure S3: Example MWD Verification-Report Map



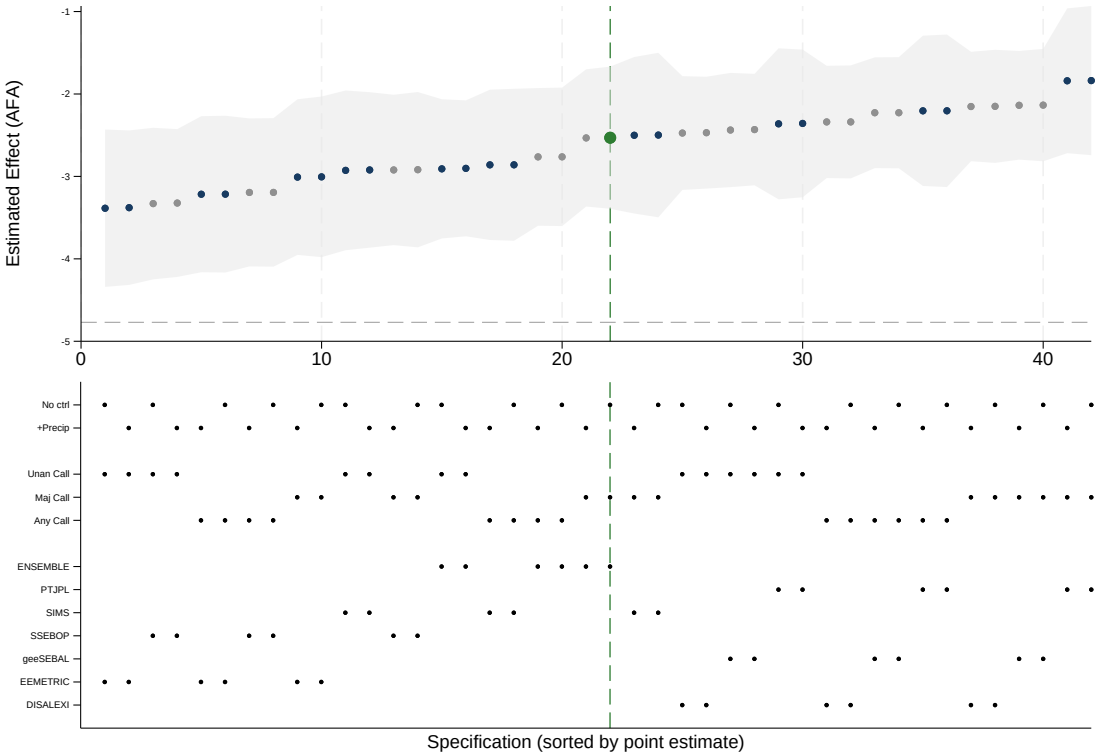
Notes: Representative page from an MWD Verification Report.

Figure S4: Called Acreage by RA Agreement Threshold vs. MWD Reported



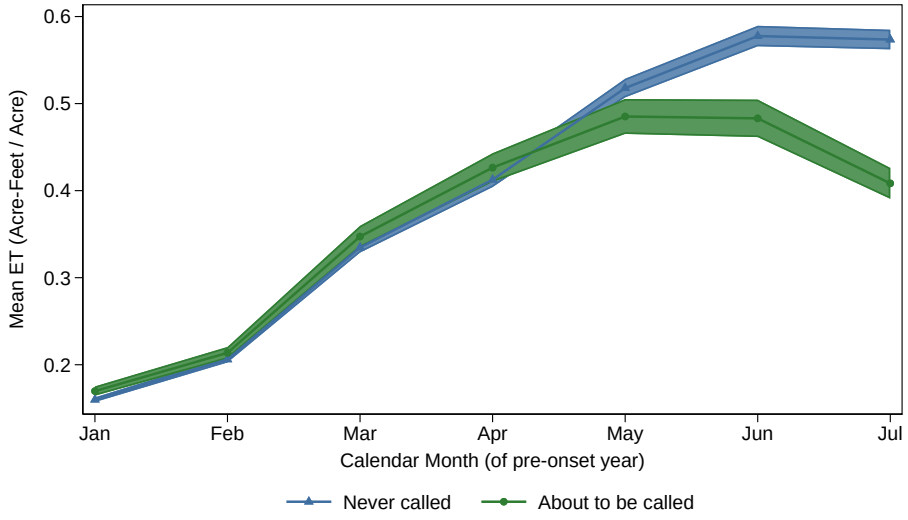
Notes: Monthly called acreage under the three RA agreement thresholds — any (light blue), majority (green; the preferred spec), unanimous (dark blue) — overlaid on the MWD-reported Fallowed Land acre-month series (dark gray dashed). The three RA series bracket the MWD series through the program window, with the majority ($\geq$2 RAs) threshold tracking MWD most closely.

Figure S5: Specification Curve



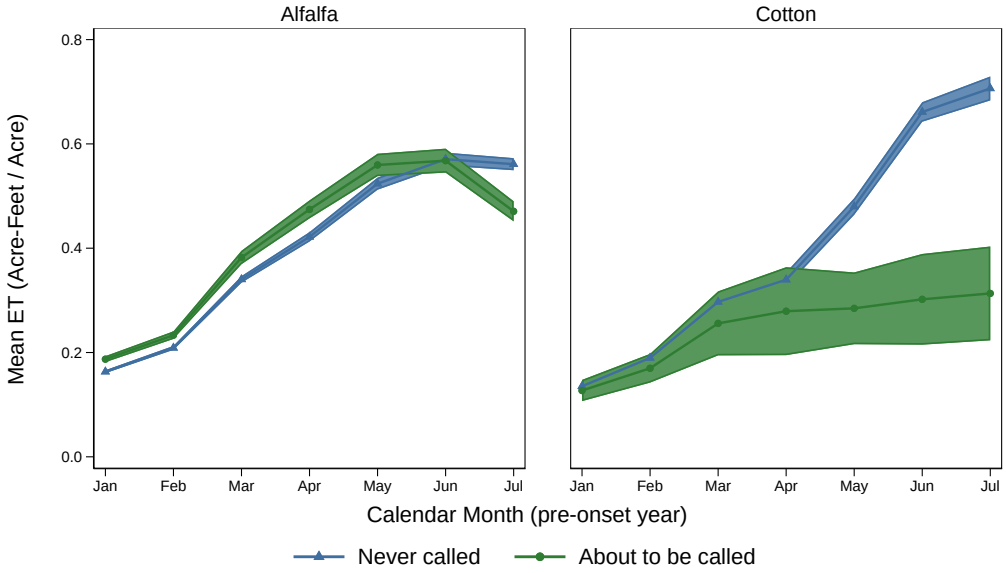
Notes: Coefficient + 95% CI for each of 42 specifications obtained by crossing seven OpenET models, three RA call thresholds (any, majority, unanimous), and two control sets (none / +precipitation). Identification-relevant choices are held fixed at the preferred values: sample = All, spatial-flag threshold $=\geq 1$ RA (the most conservative; see SI Table S12). Specifications sorted ascending by coefficient. Top-panel dot coloring distinguishes specifications that pass the joint-nullity placebo test at $p > 0.10$ (blue) from those that do not (grey); the preferred specification (ENSEMBLE, call = 2, no controls) is the green dot. The vertical green dashed line marks the preferred-spec rank and aligns with the indicator grid below. Horizontal dashed reference line at -4.77 AFA marks the maximum potential savings, equal to average water use on enrolled parcels.

Figure S6: Pre-Onset Monthly ET on About-to-Be-Called vs. Never-Called Fields



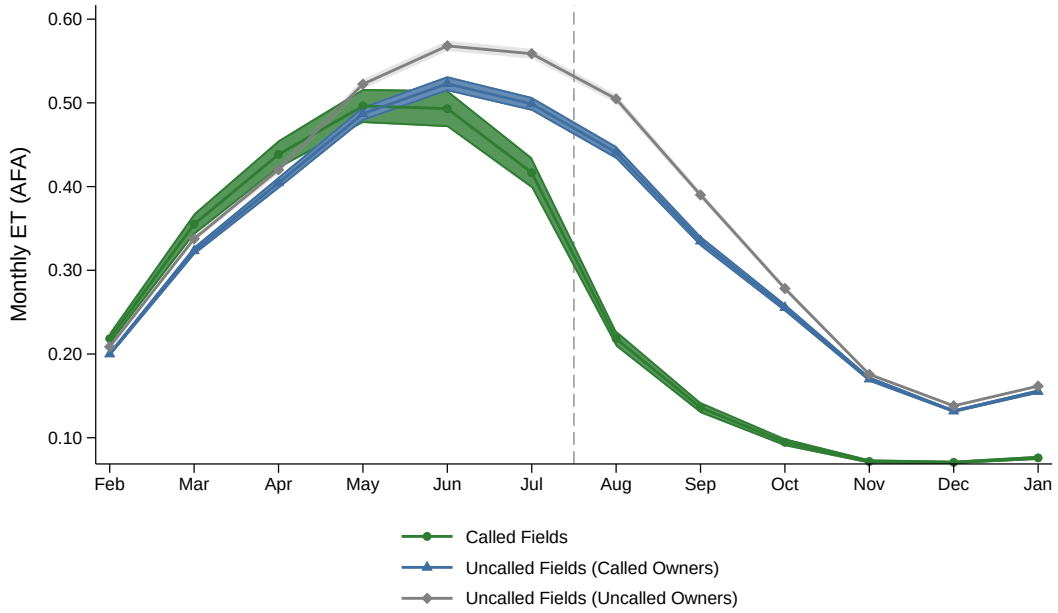
Notes: Monthly ET for fields in the six months *prior* to a call onset (green) versus fields that are never called (blue). 95% CIs on each point. The two series track one another January–May, then the about-to-be-called line drops sharply in June and July. This is the Jun/Jul fallowing-prep phenomenon described in the main text.

Figure S7: Pre-Onset Monthly ET Drop by Crop Type



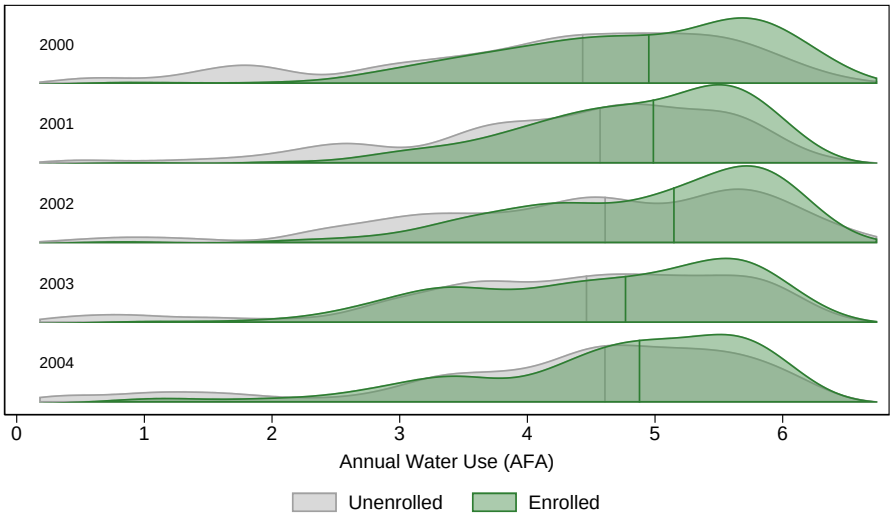
Notes: Monthly difference in ET (about-to-be-called vs. never-called) by CDL crop type. Crop classifications are from the USDA Cropland Data Layer matched to PVID field boundaries at the pre-call cropping decision (the CDL layer for the year before call onset). The June–July dip documented in SI Figure S6 is concentrated in the crops most amenable to mid-season termination — alfalfa/forage and annual row crops—but emerges during the planting season in April for permanent crops like cotton.

Figure S8: Monthly ET around Call Onset: Called vs. Enrolled-Uncalled vs. Unenrolled



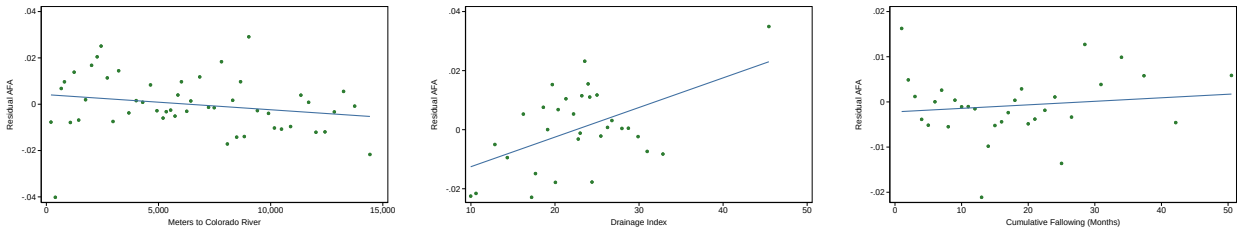
Notes: Monthly per-acre ET on three mutually exclusive groups of fields during a 12-month window around a August call onset. Called (green) are in-window observations of fields whose Aug-onset spell passes the spell-level quality screen. Uncalled fields associated with farmers who have some called fields are depicted in blue. Uncalled fields associated with farmers who have zero called field are depicted in grey.

Figure S9: Enrollment Selection: Pre-Program ET Distributions



Notes: Joy plot of pre-program annual ET distributions on fields that are ever enrolled in the fallowing program (green) versus never enrolled (gray). Each row is one calendar year; the horizontal axis is AFA per field-year.

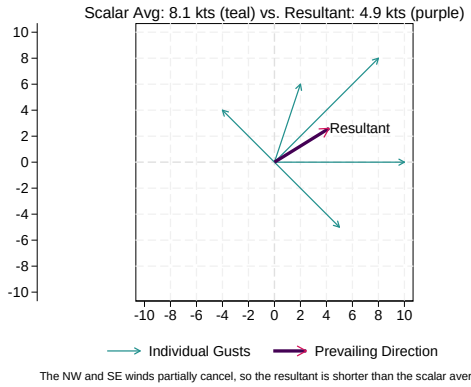
Figure S10: Residual-Water Mechanism Binscatters: Additional Covariates



(a) Distance to Colorado River (b) Drainage Index (c) Cumulative Call Spell Length

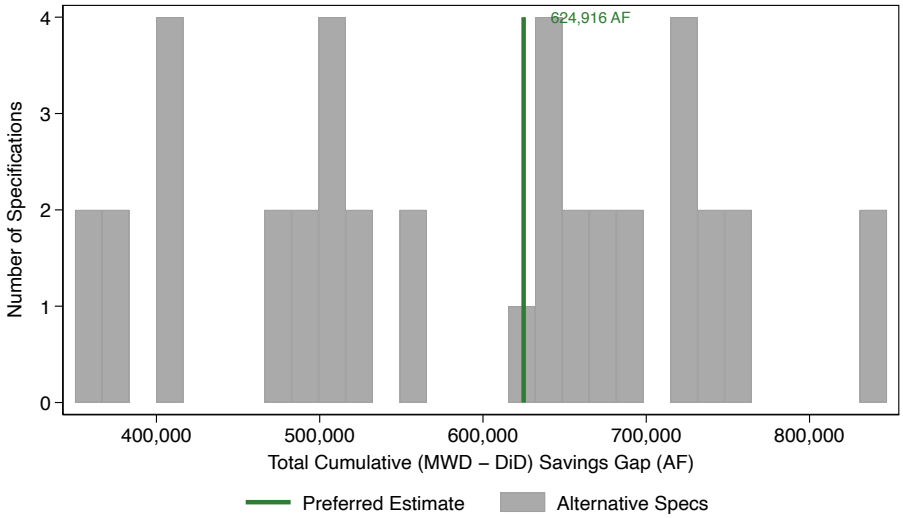
Notes: Additional binscatters of within-field residual ET (after partialling out field and month-of-sample fixed effects) against three field-level drivers that the main-text Fig. 3 does not report. Panel (a) plots residual ET against Euclidean distance from the field's centroid to the Colorado River mainstem. Panel (b) plots residual ET against a within-field Drainage Index (Schaetzl et al., 2009), which captures similar hydrologic structure. Panel (c) plots residual ET against cumulative months called for a given fallowing spell.

Figure S11: Prevailing Wind Direction at PVID



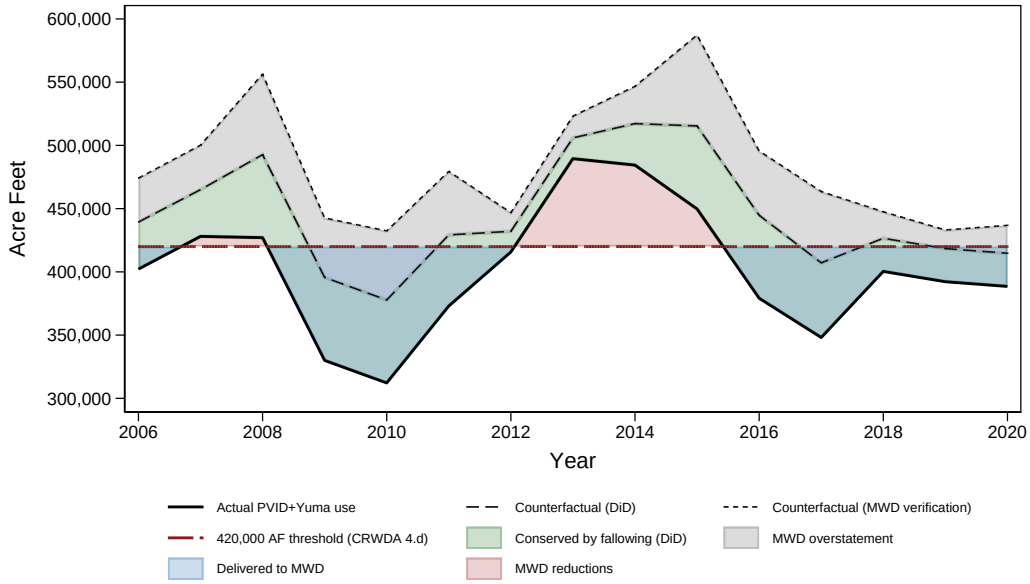
Notes: Wind rose illustrating how we aggregate hourly wind data from Iowa Environmental Mesonet into monthly prevailing wind directions using vector algebra, accounting for both wind direction and wind speed for each unique month-of-sample.

Figure S12: Cumulative Water-Savings Gap



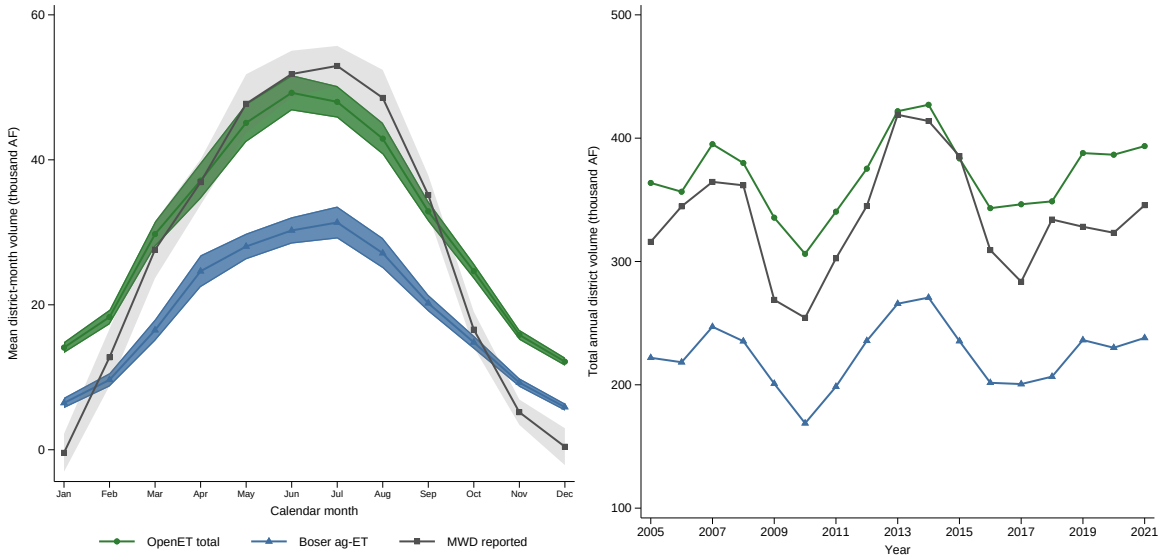
Notes: This figure integrates annual implied differences over the program window: each gray bar is one of the 41 robustness specifications' total cumulative (MWD − DiD) savings gap summed over program years 2005–2021, and the vertical green spike marks the preferred specification's value of 624,916 AF, or 43.5% of MWD's reported total.

Figure S13: Savings Estimates vs. Actual Deliveries to MWD



Notes: Combined consumptive use of PVID (priorities 1 and 3b) and the Yuma Project Reservation Division with Yuma Island (priority 2), 2006–2020, comparted to the 420,000 AF threshold that governs deliveries to MWD under Section 4.d of the 2003 Colorado River Water Delivery Agreement (maroon dash-dot line). The solid black curve is actual combined use (U_t), taken from the annual USBR Decree Accounting Reports. The long-dashed curve adds our DiD-estimated savings to actual recorded use and so traces the counterfactual combined use that would have occurred without fallowing; the green band between the two is the water our estimates attribute to the program. The short-dashed curve adds MWD's reported savings instead, and the gray band between the two counterfactuals is the overstatement documented in Figure S12. Where actual use falls below the threshold, the combined area of the blue band and the green band is the residual delivered to MWD; where it rises above, the red band is the corresponding reduction to MWD's own priority-4 apportionment, which occurs in 2008 and in 2013–2015.

Figure S14: Comparison to Boser et al. (2024) Agricultural ET Adjustment



Notes: Left panel: mean district-wide ET (thousand AF) by calendar month with 95% CI ribbons; right panel: total annual district ET by calendar year, 2005–2021. Boser et al. (2024) (blue), OpenET Ensemble (green), MWD reported consumptive use (grey). ET volumes are computed as AFA per field × field acres summed over all PVID fields. MWD reported is from the annual USBR Decree Accounting Reports and MWD Verification Reports. Companion regression slopes in SI Table S11.

Tables

Table S1: Field-Level Balance and Summary Statistics

	Never-Called		Ever-Called		Difference	
	Mean	SD	Mean	SD	Estimate	(SE)
Panel A: Field characteristics (time-invariant)						
Acres	25.8	22.1	35.4	23.3	9.6***	(1.4)
Distance to Colorado River (m)	6462	5217	6018	4015	-444	(613)
Drainage index	22.269	8.628	23.048	5.701	0.779	(0.660)
Panel B: Baseline ET, uncalled field-prog-years (AFA per year)						
Own field ET	4.36	1.37	4.76	1.05	0.40***	(0.09)
Mean neighbor ET	4.24	1.12	4.43	0.86	0.19**	(0.08)
Panel C: Crop shares, 2007–2010 (CDL)						
Alfalfa	0.664	0.339	0.500	0.348	-0.165***	(0.023)
Cotton	0.116	0.236	0.059	0.167	-0.057***	(0.015)
Hay	0.031	0.113	0.021	0.100	-0.010	(0.007)
Fallow	0.152	0.249	0.412	0.345	0.260***	(0.020)
Other	0.036	0.123	0.009	0.054	-0.028***	(0.007)
Fields	726		1,985		2,711	
Panel D: Treatment intensity (ever-called fields, $N = 1,985$)						
Cumulative fallowing months	mean 42.8, median 36					
Longest spell (program-years)	mean 2.5, median 2					
Panel E: Panel structure (prog-year analysis sample)						
Field × prog-year observations	48,798					
Unique fields	2,711					
Program years	18					
Unique owners	302					

Notes: Comparison of treated (ever-called) and control (never-called) fields across field characteristics, pre-program ET outcomes, and early-sample crop shares. Panel A reports time-invariant covariates (acreage; Euclidean distance to the Colorado River; drainage index). Panel B reports baseline ET by ever-enrolled status. For ever-enrolled fields this equals the paper's "Potential" reference of 4.77 AFA. Standard errors are clustered at the PLSS section level. Panel C reports field-level shares from the USDA Cropland Data Layer over the earliest available CDL years (2007–2010). Difference columns report the regression coefficient from $y_i = \alpha + \beta\,\text{Ever-Called}_i + \varepsilon_i$ with cluster-robust standard errors at the PLSS section level; significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$. Panel D summarizes treatment intensity among the ever-called fields (cumulative months in any call; longest contiguous call spell measured in program-years).

Table S2: Regressions of MWD-Reported Consumptive Use on Each OpenET Model

	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
Slope	0.992***	1.021***	1.001***	1.007***	0.998***	0.865***	1.037***
	(0.0153)	(0.0213)	(0.0163)	(0.0143)	(0.0187)	(0.0130)	(0.0149)
adj. R^2	0.930	0.895	0.924	0.937	0.916	0.932	0.944
RMSE	9161.7	11265.6	9571.0	8712.5	10069.3	9037.4	8190.0
N	204	204	204	204	204	204	204

Notes: Slope is the elasticity of MWD AF on OpenET AFA (no intercept); adjusted R^2 and RMSE are of the same regression. All seven models are close to a 1:1 correspondence with MWD's reported totals at the district-monthly aggregation, with slope estimates ranging from 0.865 (`EEMETRIC`) to 1.037 (`DISALEXI`). Companion aggregate time-series in SI Figure S1.

Table S3: Full Regression Results (Preferred Specification)

	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
Treatment	-2.530***	-1.837***	-2.498***	-2.918***	-2.135***	-3.005***	-2.150***
	(0.440)	(0.462)	(0.509)	(0.481)	(0.348)	(0.497)	(0.350)
Placebo 1 (t−1)	0.149	0.128	0.143	0.146	0.134	0.150	0.178
	(0.155)	(0.166)	(0.182)	(0.165)	(0.127)	(0.180)	(0.131)
Placebo 2 (t−2)	0.340*	0.311	0.442*	0.356*	0.304**	0.333	0.302**
	(0.181)	(0.193)	(0.227)	(0.194)	(0.143)	(0.204)	(0.140)
Placebo 3 (t−3)	0.518***	0.442**	0.573**	0.654***	0.455***	0.356	0.455***
	(0.187)	(0.211)	(0.229)	(0.198)	(0.147)	(0.257)	(0.151)
Placebo 4 (t−4)	0.163	0.231	0.088	0.158	0.170	0.327	0.264
	(0.202)	(0.246)	(0.229)	(0.222)	(0.154)	(0.210)	(0.162)
Placebo 5 (t−5)	0.739*	0.862**	0.693*	0.820*	0.608**	0.881**	0.663**
	(0.379)	(0.436)	(0.393)	(0.428)	(0.295)	(0.442)	(0.303)
Placebo 6 (t−6)	0.191	0.235	0.242	0.225	0.135	0.258	0.059
	(0.177)	(0.290)	(0.248)	(0.183)	(0.117)	(0.209)	(0.110)
p-value (joint nullity of placebos)	0.132	0.335	0.163	0.055	0.057	0.363	0.070
DiD / MWD reported	54.1%	39.3%	53.5%	62.4%	45.7%	64.3%	46.0%
MWD reported baseline (AFA)				4.67			
Field × prog-year observations				12,018			
Switcher fields				106			
Clusters (PLSS sections)				195			

Notes: All columns use the same preferred spec (call=2, flag=1, all-sample, no controls, `did_multiplegt_dyn` with `effects(6) placebo(6)`), varying only the ET outcome. Standard errors clustered at the PLSS section are in parentheses below each coefficient. Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$. The row "p-value (joint nullity of placebos)" reports the Wald test of the six pre-period placebos jointly equal to zero; all models reject only weakly or fail to reject, consistent with no pre-trend violation. The preferred ENSEMBLE column yields −2.53 AFA (54.1% of MWD's reported baseline of 4.67 AFA).

Table S4: SUTVA Sensitivity: Preferred Specification vs. Drops Pure Controls

	(1) Preferred (samp = All)	(2) Drop pure control	(3) Purified ctrl, no FE	(4) No owner trend
Treatment (Av_tot_eff)	-2.530***	-3.086***	-3.046***	-2.905***
	(0.440)	(0.464)	(0.418)	(0.398)
Placebo 1 (t−1)	0.149	0.050	0.115	0.097
	(0.155)	(0.159)	(0.131)	(0.129)
Placebo 2 (t−2)	0.340*	0.194	0.249	0.235
	(0.181)	(0.209)	(0.187)	(0.178)
Placebo 3 (t−3)	0.518***	0.292	0.476***	0.467***
	(0.187)	(0.206)	(0.175)	(0.160)
Placebo 4 (t−4)	0.163	-0.416**	0.374**	0.320*
	(0.202)	(0.206)	(0.176)	(0.169)
Placebo 5 (t−5)	0.739*	0.023	1.076***	0.887***
	(0.379)	(0.339)	(0.263)	(0.261)
Placebo 6 (t−6)	0.191	-0.363	0.707**	0.558*
	(0.177)	(0.239)	(0.313)	(0.307)
p-value (joint nullity of placebos)	0.132	0.048	<0.001	0.006
Field × prog-year obs	12,018	5,447	9,569	12,018
Unique fields	1,170	803	1,033	1,170
Clusters (PLSS sections)	195	177	190	195

Notes: Column (1) reports the preferred specification, which uses the full sample and absorbs non-parametric trends in landowner identity. Column (2) drops never-called fields, so identification comes only from variation in call timing across fields that are ever called. Column (3) implements a retains ever-called fields together with fields whose owner never has any field called (i.e., it drops uncalled fields of called owners); it also omits the owner-trend absorption, which would otherwise mechanically collapse this sample to column (2) by absorbing never-called owners entirely. Column (4) retains the full sample but drops the owner-trend absorption. Six placebo coefficients reported below the treatment estimate; joint-nullity p-value reported as "p-value (joint nullity of placebos)." Standard errors clustered at PLSS section in parentheses. Significance: $^{*}p< 0.10$, $^{**}p< 0.05$, $^{***}p< 0.01$. Companion monthly visualization in SI Figure S8.

Table S5: Predictors of Residual ET on Called Fields

	Cross-sectional gradient		Within-field over time	
Meters to Colorado River	-0.000			
	(0.000)			
Drainage Index		0.000		
		(0.000)		
Cumulative Fallowing (Months)			-0.000	
			(0.000)	
Average Neighbor ET (AFA)				0.344***
				(0.033)
Constant	0.111***	0.099***	0.108***	0.026***
	(0.005)	(0.009)	(0.003)	(0.008)
Field FE	No	No	Yes	Yes
Month-of-sample FE	Yes	Yes	Yes	Yes
Observations	17,299	17,299	17,299	17,299
R^2	0.217	0.216	0.491	0.557

Notes: OLS regressions of monthly per-acre ET (AFA, ENSEMBLE) on parcel characteristics hypothesized to explain positive water use on called/fallow fields. The sample is restricted to field-months in which the field is called and not flagged or partially fallowed. Columns (1) and (2) examine cross-sectional gradients in time-invariant covariates and absorb month-of-sample fixed effects only; columns (3) and (4) identify within-field variation over time and absorb both field and month-of-sample fixed effects. *Note:* standard errors clustered at the PLSS section level in parentheses; significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table S6: Regressions of Own ET on Neighbor ET by Wind Orientation (Called Field-Months)

	(1)	(2)	(3)	(4)
All Neighbors' ET Mean		0.404***		0.283***
		(0.0107)		(0.0217)
Upwind Neighbors' ET Mean			0.187***	0.0641***
			(0.00648)	(0.0100)
Downwind Neighbors' ET Mean			0.183***	0.0624***
			(0.00627)	(0.0109)
N	95866	95866	95866	95866
adj. R^2	0.386	0.477	0.471	0.479
p-val (Upwind = Downwind)			0.643	0.845

Notes: The estimating sample is called field-months. All four columns absorb field, calendar-month, and year fixed effects, with standard errors clustered at the field. Column (1) is a fixed-effects-only baseline carrying no neighbor regressors, and it establishes the reference adjusted R^2 against which the remaining columns are read. Column (2) adds the field's mean neighbor ET (all neighbors equally weighted); column (3) splits the neighbor mean into upwind and downwind components using each month's prevailing wind; column (4) includes both the all-neighbor mean and the upwind/downwind split as a joint test. Upwind and downwind slopes are statistically indistinguishable (p=0.64 in column 3, 0.85 in column 4), consistent with the residual-ET spillover being mediated by sub-surface hydrology rather than atmospheric vapor transport.

Table S7: Minimum Detectable Effects (MDE) for the Upwind-vs.-Downwind Difference

Quantity	Value
Total neighbor effect	0.404 (0.011)
Upwind neighbor effect	0.187 (0.006)
Downwind neighbor effect	0.183 (0.006)
Upwind − Downwind	0.004 (0.009)
p-value (H_0: equal)	0.64
MDE (power=0.80, α=0.05)	0.024 (5.9% of total)
MDE (power=0.90, α=0.05)	0.028 (6.8% of total)
MDE (power=0.80, α=0.10)	0.021 (5.2% of total)

Notes: Computed from the observed SE on the upwind−downwind contrast in column (3) of SI Table S6. At conventional thresholds (80% power, α=0.05 two-sided), the smallest atmospheric contribution to the neighbor effect the design could have detected is 0.024 AFA, or 5.9% of the total neighbor-ET coefficient. Given the observed upwind−downwind difference of 0.004 (p=0.64), the atmospheric channel is bounded well below the full neighbor spillover.

Table S8: Within-Owner Selection Regressions

	AR(1)	AR(1)+Fld FE	Lagged	Both	ML (LASSO)
Predicted CF (AR1)	-0.054***	-0.247***		0.014*	
	(0.012)	(0.020)		(0.008)	
Lagged ET			-0.067***	-0.071***	
			(0.006)	(0.005)	
Predicted CF (ML)					-0.114***
					(0.013)
_cons	0.314***	1.231***	0.369***	0.320***	0.595***
	(0.058)	(0.096)	(0.026)	(0.047)	(0.063)
Field FE	No	Yes	No	No	No
Owner × program-year FE	Yes	Yes	Yes	Yes	Yes
N	18244	18235	14401	14401	18244
adj. R^2	0.143	0.258	0.249	0.250	0.186

Notes: Dependent variable: field-prog-year call indicator (`called_2`). All regressions condition on owner × year fixed effects so that coefficients reflect *within-owner* selection among simultaneously eligible fields. Columns vary the counterfactual-ET prediction strategy: (1) AR(1); (2) AR(1) with field fixed effects; (3) simple lagged ET; (4) AR(1) + lagged; (5) cross-validated LASSO (`ml_cf_pred`) including 12- and 24-month lags, neighbor behavior, and month × year fixed effects. Standard errors clustered at the owner level; significance $^*p< 0.10$, $^{**}p< 0.05$, $^{***}p< 0.01$. The LASSO coefficient of -0.114 (column 5) is the largest in magnitude and absorbs the selection signal left in the simpler specifications.

Table S9: Valuation of the Reported-versus-Measured Shortfall

Program year	Shortfall (AF)	Transfer price ($/AF)	Value ($m)
2005	7,215	272	2.0
2006	33,660	282	9.5
2007	34,721	292	10.1
2008	35,169	299	10.5
2009	63,743	330	21.0
2010	47,065	382	18.0
2011	54,662	429	23.4
2012	50,227	472	23.7
2013	14,749	520	7.7
2014	17,251	572	9.9
2015	29,486	612	18.0
2016	71,924	627	45.1
2017	50,888	634	32.3
2018	56,374	644	36.3
2019	20,976	658	13.8
2020	14,723	670	9.9
2021	22,087	680	15.0
Total	624,916	490	306
Single-price benchmarks, same volume			
2023 QSA transfer	624,916	777	486
2026 SCIA federal	624,916	874	546

Notes: Column 2 is the annual difference between MWD's reported savings and our preferred DiD estimate, summing to the 624,916 AF reported in Figure S12. Column 3 is the price per acre-foot paid by the San Diego County Water Authority to the Imperial Irrigation District for conserved water under their transfer agreement. That schedule is stated by contract for calendar years 2003–2015 and escalates thereafter with the GDP implicit price deflator (Imperial Irrigation District and San Diego County Water Authority, 2009); because the program year runs August–July, we blend the two calendar prices it spans in a five-to-seven month weighting. Applying the unweighted calendar-year price instead gives a total of $297 million rather than $306 million. Our escalated series implies $774.43/AF for 2023 vs. the $776.97 stated independently in U.S. Bureau of Reclamation and Imperial Irrigation District (2024), a difference of 0.33%. The two single-price rows below the total apply one current rate to the entire volume. The 2023 QSA price is what an urban buyer would contract to pay, so it bounds MWD's own replacement cost. The 2026 SCIA price is paid by Reclamation rather than by an urban buyer, and it is reported for comparison with Figure 5 rather than as a price available to MWD.

Table S10: Results Using Boser et al. (2024) Adjusted ET

	(1) OpenET total (Boser subsample)	(2) Boser ag-only (Boser subsample)	(3) Full-sample OpenET headline	(4) Boser-corrected (headline × ag share)
DiD coef. (AFA)	-3.288***	-2.716***	-2.530***	-2.090***
	(0.159)	(0.141)	(0.440)	(0.363)
% of OpenET total	100%	82.6%	100%	82.6%
Unique fields	1,148	1,148	1,170	1,170
Field × prog-year obs	11,623	11,623	12,018	12,018

Notes: Columns (1) and (2) re-estimate the preferred dCdH specification on the 1,148-field Boser-covered subsample using OpenET total ET (col. 1) and Boser ag-only ET (col. 2). The ratio of column 2 to column 1 yields an implied Boser ag share of 82.6%. Columns (3) and (4) apply that ag share to the full-sample OpenET headline of −2.530 AFA, yielding a Boser-corrected headline of −2.090 AFA (ag-only causal effect). See SI Figure S14 for the levels comparison and SI Table S11 for the scaling regression. Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table S11: Comparison of Boser et al. (2024) to MWD-Reported Benchmark

	(1) Boser ag-ET	(2) OpenET total ET
Panel A: MWD reported = β× ET aggregate (no constant, robust SE)		
β	1.603***	0.993***
	(0.025)	(0.015)
p-value, H_0: $\beta = 1$	<0.001	0.639
R^2	0.942	0.931
N (year-months)	204	204
Panel B: ET aggregate as % of MWD reported (district-month, 2005–2021)		
Mean (% of MWD)	66.9%	110.4%

Notes: Panel A regresses MWD reported on each ET aggregate without a constant. Column (1): Boser ag-only. Column (2): OpenET Ensemble. Panel B reports mean district-monthly volumes (thousand AF) for the LHS benchmark and each RHS ET aggregate, plus the implied share (RHS / LHS). Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table S12: Pre-Trend p-Value by Spatial-Flag Threshold and ET Model

Flag threshold	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
≥1 RA (any; PREFERRED)	**0.132**	**0.335**	**0.163**	0.055	0.057	**0.363**	0.070
≥2 RAs (majority)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
≥3 RAs (unanimous)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Joint-nullity p-value of the six pre-period placebo coefficients, varying the RA data-flag threshold (rows) and ET model (columns). All other dimensions held fixed at preferred values (call = 2, sample = All, no controls). Bold cells pass the parallel-trends test at $p > 0.10$.

Appendix specification tables: variation across ET model and call coding

Each of the three tables below covers one RA call-coding threshold (the digitization aggregation rule), holding sample, spatial-flag threshold, and estimator fixed at their preferred values. Within each table, columns report the seven OpenET models and panels vary the two control specifications (no controls vs. + precipitation), so each table summarizes fourteen specifications and the bundle as a whole summarizes forty-two.

Table S13: Specification Robustness Across ET Models — Call Coding ≥1 RA (Any)

	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
Panel A: No controls							
Treatment	-2.763***	-2.204***	-2.859***	-3.194***	-2.227***	-3.215***	-2.339***
	(0.429)	(0.472)	(0.470)	(0.460)	(0.344)	(0.485)	(0.350)
p-value (joint nullity of placebos)	*0.064*	*0.187*	*0.105*	*0.038*	*0.019*	*0.197*	*0.012*
Panel B: + precipitation							
Treatment	-2.763***	-2.204***	-2.859***	-3.194***	-2.227***	-3.216***	-2.339***
	(0.426)	(0.465)	(0.465)	(0.458)	(0.343)	(0.483)	(0.347)
p-value (joint nullity of placebos)	*0.060*	*0.184*	*0.101*	*0.036*	*0.018*	*0.194*	*0.011*
N (switcher-year cells, dCdH)				2,925			

Notes: dCdH average treatment-on-treated coefficients for the post-call ET effect under the ≥1 RA (any) call-coding rule — the most permissive coding; treats any RA-flagged month as a call. Sample includes all PVID fields (ever-called and never-called); spatial-flag threshold = ≥1 RA (drops any field flagged as a fuzzy image by any RA). Columns are OpenET ensemble and per-model outputs. Panel A omits time-varying controls; Panel B adds monthly field-level precipitation. The italic row in each panel reports the joint-nullity *p*-value of the six pre-period placebo coefficients. Standard errors clustered at the PLSS section in parentheses. Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table S14: Specification Robustness Across ET Models — Call Coding ≥2 RAs (Majority; Preferred)

	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
Panel A: No controls							
Treatment	-2.530***	-1.837***	-2.498***	-2.918***	-2.135***	-3.005***	-2.150***
	(0.440)	(0.462)	(0.509)	(0.481)	(0.348)	(0.497)	(0.350)
p-value (joint nullity of placebos)	*0.132*	*0.335*	*0.163*	*0.055*	*0.057*	*0.363*	*0.070*
Panel B: + precipitation							
Treatment	-2.533***	-1.840***	-2.501***	-2.921***	-2.137***	-3.008***	-2.152***
	(0.425)	(0.448)	(0.484)	(0.465)	(0.336)	(0.482)	(0.339)
p-value (joint nullity of placebos)	*0.089*	*0.213*	*0.103*	*0.038*	*0.041*	*0.329*	*0.059*
N (switcher-year cells, dCdH)				1,929			

Notes: dCdH average treatment-on-treated coefficients for the post-call ET effect under the ≥2 RAs (majority; the preferred coding) — requires at least two of the three RAs to agree. Sample includes all PVID fields (ever-called and never-called); spatial-flag threshold = ≥1 RA (drops any field flagged as a fuzzy image by any RA). Columns are OpenET ensemble and per-model outputs. Panel A omits time-varying controls; Panel B adds monthly field-level precipitation. The italic row in each panel reports the joint-nullity *p*-value of the six pre-period placebo coefficients. Standard errors clustered at the PLSS section in parentheses. Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table S15: Specification Robustness Across ET Models — Call Coding ≥3 RAs (Unanimous)

	ENSEMBLE	PTJPL	SIMS	SSEBOP	geeSEBAL	EEMETRIC	DISALEXI
Panel A: No controls							
Treatment	-2.908***	-2.362***	-2.927***	-3.330***	-2.437***	-3.386***	-2.474***
	(0.431)	(0.467)	(0.494)	(0.469)	(0.353)	(0.487)	(0.352)
p-value (joint nullity of placebos)	*0.142*	*0.301*	*0.140*	*0.045*	*0.060*	*0.125*	*0.065*
Panel B: + precipitation							
Treatment	-2.902***	-2.357***	-2.921***	-3.323***	-2.432***	-3.379***	-2.469***
	(0.421)	(0.457)	(0.481)	(0.457)	(0.345)	(0.478)	(0.346)
p-value (joint nullity of placebos)	*0.121*	*0.278*	*0.111*	*0.035*	*0.051*	*0.128*	*0.060*
N (switcher-year cells, dCdH)				1,629			

Notes: dCdH average treatment-on-treated coefficients for the post-call ET effect under the ≥3 RAs (unanimous) — the most conservative coding; requires all three RAs to agree. Sample includes all PVID fields (ever-called and never-called); spatial-flag threshold = ≥1 RA (drops any field flagged as a fuzzy image by any RA). Columns are OpenET ensemble and per-model outputs. Panel A omits time-varying controls; Panel B adds monthly field-level precipitation. The italic row in each panel reports the joint-nullity *p*-value of the six pre-period placebo coefficients. Standard errors clustered at the PLSS section in parentheses. Significance: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.