

Do Flexible Water Allocations Reduce Groundwater Extraction? Evidence from Multi-Year Flex Accounts in Kansas*

Micah V. Cameron-Harp

Department of Agricultural Economics, Kansas State University

mcameronharp@ksu.edu

September 11, 2026

Abstract

This paper examines the impact of Kansas’s Multi-Year Flex Account (MYFA) program, which allows irrigators to bank unused water allocations across a five-year period rather than facing rigid annual limits. Using administrative data on groundwater withdrawals for 13,000 water right groups from 2001 to 2021, my preferred specification applies the de Chaisemartin and D’Haultfœuille (2024) dynamic difference-in-differences estimator, which accommodates reversible treatment and restricts comparisons to switchers and stayers within the same Groundwater Management District. This specification yields an average treatment effect of approximately 12.5 percent reduction in annual groundwater withdrawals, with negative effects in every estimated event-time period, including the post-expiration years. Later cohorts that enrolled during drought conditions have wide confidence intervals reflecting smaller samples and

*Paper prepared for the American Journal of Agricultural Economics. I thank Gabe Sampson, Karina Schoengold, Felipe Avileis, Ellen Bruno, and Nathan Hendricks for their feedback on early versions of this work. All errors are my own.

shorter post-enrollment windows, so the cross-cohort heterogeneity is statistically inconclusive. The calendar-year aggregation, however, identifies the largest treatment effects in years with mixed-to-favorable precipitation, consistent with a banking mechanism whereby MYFAs enable conservation primarily when favorable weather allows irrigators to accumulate unused allocations. These findings suggest that relaxing temporal rigidity in water allocations may actually encourage conservation by reducing the “use it or lose it” incentive inherent in annual allocation regimes.

Keywords: Groundwater management, water allocation, flexible permits, irrigation, difference-in-differences, Ogallala Aquifer

JEL Codes: Q15, Q25, Q28, Q58

1 Introduction

Systems of water rights are expensive to build. Converting an open-access flow into a set of enforceable claims requires hydrologic measurement, adjudication of competing and often historical uses, and a permanent administrative apparatus to record and police diversions, and the resulting entitlements are durable property interests that holders defend vigorously (Colby, 1990; McCann and Easter, 2004). Any change in the quantity, timing, or transferability of an existing right must clear procedural requirements designed to protect senior claimants from impairment, and the resulting transaction costs are large enough to deter reallocation even where the gains from reallocating water are substantial (Womble and Hanemann, 2020). Regulators are consequently hesitant to modify rights once granted, which is a serious constraint when a right structure written for an era of resource development becomes a poor instrument for conservation. My study context, the Kansas portion of the High Plains Aquifer, is the quintessential example of this problem. Kansas allocated groundwater rights decades ago under what critics have termed a “policy of planned depletion,” entitling irrigators to a fixed quantity each year that is forfeited if unused, and the Ogallala Aquifer beneath them has declined accordingly: Kansas levels fell by over 1.5 feet in a single recent year, continuing a decades-long trend (Kansas Geological Survey, 2024). Against this backdrop of institutional inertia, the Multi-Year Flex Account (MYFA) program that Kansas authorized in 2012 is a rare case in which the terms of existing water rights were in fact modified in the service of improving efficiency. In permitting water to be moved across time, a MYFA gives irrigators a capacity for intertemporal arbitrage similar to a water bank. However, unlike a water bank, MYFAs do not allow users to sell unused allocation. As such, the MYFA policy allows me to isolate the impact of this intertemporal arbitrage feature on withdrawals.

California’s 1991 Drought Water Bank purchased and resold several hundred thousand acre-feet during a severe drought, and the pattern of participation and prices indicates that the exchange moved water from lower- to higher-valued uses at a time when the gap between them was widest (Howitt, 1994). Across the western United States, the volume of water transferred through markets has grown steadily since the 1990s, and the persistent premium that urban and environmental buy-

ers pay over agricultural sellers reveals substantial gains that reallocation captures (Brewer et al., 2008). In Australia’s Murray–Darling Basin, allocation and entitlement trading became a routine instrument for managing scarcity over the course of the Millennium Drought (Wheeler et al., 2014), and short-term trade on the Watermove exchange yielded measurable efficiency improvements in the allocation of a fixed seasonal supply (Brooks and Harris, 2008). Closest to the setting studied here, Ayres et al. (2021) exploit a spatial discontinuity in California groundwater rights and find that the market created by adjudication generated large net benefits capitalized into land values, with gains attributable in part to trading that moved water toward higher-valued uses and with groundwater levels rising rather than falling. Banking across time follows the same logic as trading across users: Arellano-Gonzalez and Moore (2020) show that access to groundwater banking in California is precisely what makes the intertemporal arbitrage of water valuable enough to support long-lived perennial investments under drought risk. A MYFA reproduces the temporal half of this mechanism while withholding the rest. Carryover is bounded rather than open-ended, since unused allocation may be shifted only within a five-year term, after which the account expires and the right reverts to annual administration; water therefore cannot be stored indefinitely against an arbitrarily distant drought.

In this paper, I examine how the capacity for intertemporal arbitrage provided by MYFAs affects groundwater withdrawals for irrigation. To motivate my analyses and illustrate why I expect MYFA enrollment to affect withdrawals, I first develop a model of irrigation pumping under a multi-period allocation. Within the mode, a risk-neutral irrigator faces stochastic growing-season precipitation that substitutes for applied water. Under an annual allocation, or a “hard cap”, the two periods are decoupled and unused allocation is worthless. In contrast, under a multi-year allocation, or a “soft cap”, water left in the account carries a shadow value into the following period. That option value gives the irrigator a reason to preposition for drought, holding water back in favorable years to guard against an adverse realization later. I then use my model to characterize the properties of the precipitation distribution under which the conservation force dominates the offsetting incentive to draw the account down in dry years, so that the multi-year allocation reduces

aggregate use and not merely the timing of withdrawals. I demonstrate the option value is largest, and total extraction falls, when precipitation shocks are variable, left-tailed, and positively autocorrelated. I then use three decades of empirical data on withdrawals by irrigators in Kansas to test whether MYFA participation altered how much water enrolled irrigators extract. I find enrollment reduces annual withdrawals by approximately 12.5 percent on average. Interestingly, the dynamic estimates indicate the effect is negative in every estimated event-time period and the reduction in withdrawals persists into the years following expiration of the MYFA. Consistent with the option-value mechanism, reductions are concentrated in years when precipitation was mixed-to-favorable, the conditions under which an irrigator has unused allocation worth banking in the first place.

The empirical analysis exploits the staggered enrollment of water right groups in the MYFA program between 2012 and 2021, using administrative records from the Kansas Division of Water Resources on annual water use by more than 13,000 irrigation water right groups from 2001 to 2021. A methodological challenge specific to this setting is that MYFA participation is reversible: an account is in force for approximately five years, after which the water right reverts to annual allocation. I address this by adopting the de Chaisemartin and D'Haultfœuille (2024) dynamic difference-in-differences estimator as the preferred specification, since it accommodates treatment status that switches on and off at the unit level and identifies event-time effects both within and after the flex period. I further restrict comparisons to switchers and stayers within the same Groundwater Management District, which weakens the parallel trends assumption from requiring comparability across all Kansas water right groups to requiring comparability within each Groundwater Management District (GMD). I complement this analysis with the Callaway and Sant'Anna (2021) group-time estimator, which supports natural aggregation by adoption cohort and by calendar year and thereby enables a test of whether the effect varies with the conditions under which irrigators enroll.

This paper makes three contributions. First, to my knowledge, I provide the first estimates of how the capacity for intertemporal arbitrage affects the quantity of groundwater extracted. The literature on water banks and markets summarized above evaluates such institutions on the efficiency

with which a given quantity of water is allocated rather than on whether the total quantity withdrawn changes (Howitt, 1994; Brewer et al., 2008; Ayres et al., 2021; Wheeler et al., 2014). The same emphasis appears in the design of tradable groundwater permits, where the object of interest is the spatial pattern of extraction relative to the streams it depletes (Kuwayama and Brozović, 2019). The closest antecedent to the present analysis is Young et al. (2021), who simulate hard- and soft-cap allocations in a coupled hydrologic-economic model but do not estimate behavioral responses from observed data. Because a MYFA grants temporal flexibility without granting tradability, the Kansas program supplies the variation needed to identify the extraction consequences of intertemporal arbitrage alone, holding the reallocation margin fixed.

The second contribution is to the literature on flexibility as a response to resource depletion. Flexibility is conventionally defended on cost grounds, since allowing regulated entities to choose when and how to comply lowers the cost of achieving a given target (Goulder and Parry, 2008; Holland et al., 2009). The model developed here shows that the suspicion is not general: once an allocation spans multiple periods, saved water acquires an option value that annual allocation suppresses, and that option value is itself a conservation incentive built into the flexible instrument rather than an unintended consequence of it. This complements what has been learned from Kansas's Local Enhanced Management Areas, the leading alternative approach to slowing depletion in the region. LEMAs impose binding, locally negotiated reductions, and the evidence indicates that they work: the Sheridan 6 LEMA reduced pumping substantially while maintaining farm profitability (Golden and Leatherman, 2015), irrigators met a multi-year allocation largely by reducing water applied to the same crops rather than by fallowing (Drysdale and Hendricks, 2018; Deines et al., 2019), and a decade of monitoring shows the resulting extension of the aquifer's usable life (Orduña Alegría et al., 2024). The MYFA results suggest that a voluntary instrument that relaxes rather than tightens the per-period limit can move withdrawals in the same direction.

The third contribution is a caution that follows from the estimates for enrollment cohorts and calendar years. While I find the conservation effect is strongest in years and cohorts with favorable precipitation, my estimates for the effects during times of drought are too imprecise to rule out

a smaller or opposite response. The cohort estimates for irrigators who enrolled in dry years are larger in absolute magnitude but carry confidence intervals roughly five times wider than the best-identified cohort. This leaves a non-zero risk that flexibility permits heavier extraction precisely when the aquifer and the systems connected to it are least able to absorb it. This is precisely the risk that Young et al. (2021) find when modeling soft-cap allocations—they find soft caps raise farm profitability relative to equivalent hard caps but produce higher rates of streamflow depletion in dry periods when users draw down accumulated balances.

The remainder of the paper proceeds as follows. Section 2 describes the institutional background of groundwater management in Kansas and the MYFA program. Section 3 describes my two-period stochastic model that contrasts hard cap and soft cap allocation policies and identifies the conditions under which the soft cap reduces expected extraction. Section 4 describes the data employed in my empirical analyses and the empirical properties of the precipitation distribution that the conceptual model invokes. Section 5 outlines the empirical strategy and translates the model’s predictions into testable hypotheses. Section 6 presents the results and robustness checks. Section 7 concludes.

2 Background on Kansas Water Rights and the MYFA Program

2.1 Groundwater Management in Kansas

Kansas manages groundwater through the Kansas Water Appropriation Act, which establishes a prior appropriation system for water rights. Under this system, water rights holders receive authorization to divert a specified quantity of water annually for beneficial use from the Kansas Department of Agriculture’s Division of Water Resources (DWR). Groundwater Management Districts (GMDs), local entities governed by water users, gained authority to propose Local Enhanced Management Areas (LEMAs) with stricter extraction limits. The Multi-Year Flex Account pro-

gram and the first Local Enhanced Management Area (LEMA) both date to this legislative shift: MYFAs were authorized in 2012 and the Sheridan 6 LEMA took effect in 2013. The LEMA experience subsequently demonstrated that collective reductions in pumping could stabilize water levels while maintaining farm profitability (Golden and Leatherman, 2015), and it is the leading alternative instrument against which the MYFA results below can be judged. In addition, most LEMAs incorporate elements of MYFAs by providing affected irrigators with a multi-year allocation instead of the traditional annual allocation. As such, this study of MYFAs can also be viewed as a means of understanding the components of LEMA policies that facilitate conservation.

2.2 Multi-Year Flex Accounts

The Kansas Legislature authorized Multi-Year Flex Accounts in 2012 through Senate Bill 272, with subsequent amendments expanding eligibility in 2021 through House Bill 2172. The statutory framework recognizes that “an opportunity exists to improve water management by enabling multi-year flexibility in the use of water authorized to be diverted under a groundwater water right, provided that such flexibility neither impairs existing water rights, nor increases the total amount of water diverted” (K.S.A. 82a-736).

Under a MYFA, an existing water right is suspended for a five-year term and replaced with a “term permit” that establishes an aggregate allocation. The aggregate quantity equals either (a) five times the water right’s average annual use during the base period 2000–2009, or (b) five times a net irrigation requirement calculated from crop water needs. Critically, no annual extraction limit applies during the flex period. Enrollment in a MYFA is voluntary and require application to the DWR. Water rights within certain Intensive Groundwater Use Control Areas are ineligible, and rights already subject to LEMA restrictions face additional coordination requirements. Upon approval, the MYFA takes effect and runs for five years, after which the right holder may renew or revert to annual allocation.

The program’s design creates several theoretical predictions about extraction behavior. First, by eliminating annual forfeiture of unused allocation, MYFAs remove the “use it or lose it” in-

centive that may cause irrigators to pump water they do not immediately need. Second, temporal flexibility enables irrigators to concentrate extraction in years when water has highest marginal value (e.g., drought years) and conserve in years when precipitation provides adequate moisture. Third, the multi-year accounting horizon may encourage longer-term planning and water-saving investments. However, if irrigators simply shift extraction across years without changing total use, or if flexibility enables higher peak-year extraction, total depletion could remain unchanged or even increase. The conceptual model developed in Section 3 formalizes these competing forces and identifies the conditions under which net conservation is the predicted outcome.

3 Conceptual Model

The MYFA institutional design corresponds to a “soft cap”, a multi-period aggregate quota with full intertemporal flexibility, while the standard Kansas water right is a “hard cap”, per-period quota applied independently each year. I develop a two-period stochastic model to contrast the two policies, derive the irrigator’s optimal extraction rule under each, and identify the conditions under which the MYFA-style soft cap reduces expected total extraction.¹ The model serves two purposes for the empirical analysis that follows. First, it generates a testable prediction that average annual withdrawals fall under the MYFA as a result of the option value unused allocation now has in future periods. Second, it generates a testable prediction for the cohort-specific and calendar-year treatment effects, that their magnitude will be determined by the contemporaneous drought conditions.

3.1 Setup

Consider a risk-neutral irrigator who chooses water use $q_t \geq 0$ and $q_{t+1} \geq 0$ in two consecutive periods to maximize expected profits. His per-period net profit is $\pi(q + \varepsilon)$, with $\pi'(\cdot) > 0$ and

¹The model is closest in spirit to the hydrologic-economic framework of Young et al. (2021), who simulate hard- and soft-cap allocation policies in a calibrated agronomic model. The analytical model developed here abstracts from the hydrologic externality and instead isolates the option-value mechanism that distinguishes soft caps from hard caps in any setting with stochastic resource availability.

$\pi''(\cdot) < 0$, where ε is an additive quantity shock representing growing-season precipitation that substitutes for irrigation.² I assume the irrigator discounts the future at $\delta \in (0, 1]$. I define the unconstrained profit-maximizing effective quantity as q^* , such that his unconstrained *chosen* use is $q^* - \varepsilon$. The shocks $(\varepsilon_t, \varepsilon_{t+1})$ are drawn from a joint distribution F with marginal distributions F_t and F_{t+1} and correlation $\rho = \text{Corr}(\varepsilon_t, \varepsilon_{t+1})$. Finally, I assume the current-period shock is realized before the current-period choice of withdrawals for irrigation, meaning the first-period shock (ε_t) is known when q_t is chosen but the second-period shock (ε_{t+1}) is not.

3.2 Hard Cap (Annual Allocation)

Under the standard Kansas water right, the regulator imposes a per-period quota $\bar{Q}$ that applies independently each year: $q_t \leq \bar{Q}$ for $t \in \{0, 1\}$. In each period the irrigator maximizes profits according to the following three-region rule:

$$q_s^H(\varepsilon_s) = \begin{cases} 0, & \varepsilon_s \geq q^*, \\ q^* - \varepsilon_s, & q^* - \bar{Q} \leq \varepsilon_s < q^*, \\ \bar{Q}, & \varepsilon_s < q^* - \bar{Q}. \end{cases} \quad (1)$$

Three features of the solution to the hard cap problem in equation (1) contrast with the solution to the soft cap policy below. First, in the interior region, the irrigator perfectly absorbs the realized shock by adjusting withdrawals one-for-one with realized precipitation, such that $q_s^H + \varepsilon_s = q^*$. Second, the cap only binds under sufficiently adverse shocks in the left tail of F_s . Favorable shocks never cause binding constraints because the irrigator simply reduces q_s , meaning welfare losses from the policy arise solely in low-rainfall years. Third, because the periods are decoupled, the shock correlation ρ is irrelevant to behavior.

²Modeling the shock as additive in the quantity argument captures the agronomic substitutability of natural and applied water at the margin: a millimeter of rainfall and a millimeter of irrigation supply the crop equivalently.

3.3 Soft Cap (Multi-Year Flex Account)

Under a MYFA, the regulator imposes a single aggregate quota covering both periods: $q_t + q_{t+1} \leq \tilde{Q}$. The irrigator solves a two-stage stochastic program by backward induction.

Stage 2: period $t + 1$. After observing ε_{t+1} with remaining budget $R = \tilde{Q} - q_t$, the period $t + 1$ value function is

$$V(R, \varepsilon_{t+1}) = \max_{0 \leq q_{t+1} \leq R} \pi(q_{t+1} + \varepsilon_{t+1}). \quad (2)$$

Note, the solution shares the same three-region structure as equation (1) but with the remaining allocation after period t use, R , replacing $\tilde{Q}$. By the envelope theorem, the marginal value of any remaining budget is

$$\mu(\varepsilon_{t+1}, R) = \frac{\partial V}{\partial R} = \max[0, \pi'(R + \varepsilon_{t+1})], \quad (3)$$

where μ is the LaGrange multiplier representing the option value of carrying one additional unit of unused allocation into the next period. Notice the marginal value of unused allocation in equation (1) is positive whenever the period $t + 1$ budget constraint binds and zero otherwise.

Stage 1: period t . The irrigator observes ε_t and chooses q_t to maximize $\pi(q_t + \varepsilon_t) + \delta \mathbb{E}_{\varepsilon_{t+1}|\varepsilon_t}[V(\tilde{Q} - q_t, \varepsilon_{t+1})]$. The interior first-order condition is

$$\pi'(q_t^S + \varepsilon_t) = \delta \mathbb{E}_{\varepsilon_{t+1}|\varepsilon_t}[\mu(\varepsilon_{t+1}, \tilde{Q} - q_t^S)]. \quad (4)$$

The irrigator equates the marginal benefit of pumping today, evaluated at the shock-adjusted effective quantity $q_t^S + \varepsilon_t$, to the discounted expected shadow value of preserving budget for the next period. Three properties of the soft-cap optimum drive the empirical predictions developed below and my subsequent analyses.

(i) *Muted response to current shocks.* Implicit differentiation of (4) yields $dq_t^S/d\varepsilon_t \in (-1, 0)$. Chosen use falls when current-period rainfall is favorable, but less than one-for-one—in contrast to the exact one-for-one offset under the hard cap interior. The option value of saved budget restrains

the irrigator from drawing down the aggregate quota by the full amount the current shock would otherwise permit.

(ii) *Budget pre-positioning against adverse left tails.* Because the irrigator can choose $q_t^S < \bar{Q}$, he can leave a remaining budget $R^S = \bar{Q} - q_t^S$ that exceeds the per-period hard-cap allowance $\bar{Q}$. This shrinks the region of F_{t+1} in which the period- $t + 1$ constraint binds, directly reducing expected period- $t + 1$ welfare loss.

(iii) *Use of shock correlation.* The conditional expectation on the right hand side of (4) means q_t^S is a function of ρ . Positive correlation between current and future precipitation shocks amplifies precautionary saving: an adverse ε_t raises the probability of an adverse ε_{t+1} , raising $\mathbb{E}[\mu \mid \varepsilon_t]$ and reducing q_t^S . Negative correlation reverses the sign of the response.

3.4 Model Implications for Empirical Analyses

To illustrate the effect of MYFA enrollment on total withdrawals, I set $\tilde{Q} = 2\bar{Q}$ and define $\bar{Q}^S$ and $\bar{Q}^H$ as the total withdrawals over both periods for the soft cap and hard cap policies, respectively. As such, the two policies allow identical total extraction in expectation under symmetric, non-binding conditions. Whether the soft cap reduces expected aggregate extraction ($\bar{Q}^S$) relative to the hard-cap aggregate ($\bar{Q}^H$) depends on the balance between two opposing forces. The option-value channel reduces q_t below $\bar{Q}$ in favorable years through precautionary saving, while the budget-flexibility channel allows extraction above $\bar{Q}$ in adverse years. The conservation channel dominates ($\bar{Q}^S < \bar{Q}^H$) when (a) the discount factor δ is high (b) the distribution F_{t+1} is sufficiently variable, left-skewed, or has a low mean, so that adverse second-period shocks are common enough to make $\mathbb{E}[\mu]$ large; (c) the irrigator's budget slack R^S is not exhausted in expectation, so banked allocation tends to expire unused; and (d) shocks across periods are positively correlated, reinforcing precautionary saving. Conversely, when F_{t+1} has a thin left tail and δ is low, the soft cap may merely redistribute extraction across periods without reducing total withdrawals.

Note, while the above result concerning total withdrawals is derived under an assumption of a risk-neutrality, relaxing risk neutrality strengthens the conservation claim. With a strictly concave

utility function the irrigator additionally values the consumption-smoothing benefit of moving extraction toward states of greater marginal value, which raises the option value of preserved budget and amplifies the substitution gain against the left tail of F_{t+1} .³

Based on this theoretical model, I test two implications in the following empirical analyses:

Prediction 1: average reduction in total extraction. In Section 4.5, I provide data characterizing the precipitation patterns in Kansas observed during the growing season. The observed precipitation distribution in Kansas conforms with conditions (b) and (d) above. As such, I predict that average annual groundwater extraction falls under the MYFA relative to annual allocation. The main results presented in Section 6.1 test whether withdrawals decrease following enrollment on average.

Prediction 2: drought at enrollment attenuates the conservation effect. The conservation channel operates through $\mathbb{E}[\mu]$, so the magnitude of precautionary saving in t depends on how favorable the current-period shock is relative to the future shock distribution. Consider cohorts who enroll in MYFAs during drought years and experience an adverse current-period ε_t . Based on the hard-cap rule in equation (1), these irrigators would extract close to the cap imposed by their annual allocation. For those in a MYFA, there is little scope to reduce q_t below the agronomic requirement. Without accumulating any banked allocation, I expect to see little difference between the post-enrollment trajectory of MYFA enrollments and what withdrawals would have been if they had remained using a hard cap, annual allocation. Conversely, cohorts enrolling in MYFAs during years with adequate rainfall can immediately allocate $q_t < \bar{Q}$ and bank the residual, generating the largest expected reduction in $\bar{Q}^S$. The model therefore predicts a systematic ordering of cohort-specific treatment effects by precipitation conditions during the enrollment year. The cohort-aggregated estimates in Section 6.2 test this prediction.

³Formally, under a strictly concave utility $u(\cdot)$ over realized profits, the second-period value function in equation (3) becomes $V(R, \varepsilon_{t+1}) = \max_{0 \leq q \leq R} u(\pi(q + \varepsilon_{t+1}))$ and the shadow value $\mu = u'(\pi(\cdot)) \cdot \pi'(\cdot)$ is larger in absolute terms in adverse-shock states (where u' is larger) than in the risk-neutral case.

4 Data and Descriptive Statistics

4.1 Water Use Data

I draw data on groundwater withdrawals for irrigation, irrigated acreage, and crop choice from the Water Information Management and Analysis System (WIMAS). WIMAS is a joint product of the Kansas Department of Agriculture, Division of Water Resources (DWR) and the Kansas Geological Survey (KGS), and is comprised of annual water records submitted by water right holders. I aggregate individual water right data into water right groups as in Cameron-Harp and Hendricks (2025), where a water right group represents a unique network of overlapping water rights, points of diversion, and authorized places of use. This aggregation reflects the level at which authorized acre-feet and irrigated acreage decisions are made.

The analysis sample covers ten years prior to the MYFA policy beginning through the most recently available water use data in WIMAS, the years 2001–2024. The dataset includes 13,328 unique water right groups with at least one year of observed water use. Following standard practice, I exclude with incomplete panel data due to rights termination or other administrative changes. To address extreme values that may reflect data errors, I drop observations above the 95th percentile of annual water use. I also exclude outside of Groundwater Management Districts, as these areas face different regulatory environments and represent a small share of total irrigation. In addition, I exclude water right groups subject to LEMA restrictions, as LEMAs often involve a similar multi-year restriction. These restrictions results in the canonical estimation panel of 12,948 water right groups and 239,947 water right group-year observations.

4.2 Treatment Definition

I define treatment as the year in which a water right group first enrolls in a MYFA. Treatment status is obtained from DWR administrative records identifying term permits issued under the MYFA program. Among water right groups in the analysis sample, adoption begins in 2012 and continues through 2024. Enrollment is heavily front-loaded: 608 of the 753 ever-enrolled water right groups

(80.7 percent) entered in 2012, the program's first year, with later cohorts numbering between 8 and 29 water right groups each (Table 1, Panel B). Water right groups that enroll in MYFAs in multiple non-consecutive periods (e.g., enrolled, expired, then re-enrolled) are excluded to maintain a clean treatment definition. In table 1, I include information on the distribution of MYFA participants by enrollment year and Groundwater Management District. Several features of the MYFA enrollment data warrant discussion.

Table 1: Distribution of MYFA Enrollment Across Groundwater Management Districts

<i>Panel A: water right group counts and identifying-variation shares by GMD</i>						
GMD	Total water right groups	Treated water right groups	% treated	Share of treated water right groups (%)	Share of water right group-yrs (%)	Share of $D = 1$ obs (%)
1	955	23	2.4	3.1	6.7	3.2
2	1,348	131	9.7	17.4	10.2	17.1
3	4,921	161	3.3	21.4	38.9	21.7
4	2,258	56	2.5	7.4	15.0	6.0
5	3,466	382	11.0	50.7	29.2	52.0
Pooled	12,948	753	5.8	100.0	100.0	100.0

<i>Panel B: Year of first MYFA enrollment, by GMD (number of water right groups)</i>						
Start year	GMD 1	GMD 2	GMD 3	GMD 4	GMD 5	Total
2012	20	100	115	47	326	608
2013	0	4	14	1	4	23
2014	1	6	9	3	2	21
2015	0	1	3	2	5	11
2016	0	0	0	0	8	8
2017	1	5	2	3	18	29
2018	1	8	9	0	11	29
2019	0	6	0	0	7	13
2020	0	1	9	0	1	11
Total	23	131	161	56	382	753

Notes: Counts computed from the canonical processed panel (deduplicated, $N = 239,947$ water right group-years across 12,948 water right groups in 2001–2021, non-LEMA, GMD-located). “Treated water right groups” are water right groups with at least one year of active MYFA enrollment in the panel. “ $D = 1$ obs” are water right group-year observations during an active MYFA term. In Panel B, “Year of first MYFA enrollment” is the minimum `water_right_group_myfa_start` for each treated water right group. Source: `Kansas_MYFA/code/gmd_distribution.R`, run on the canonical panel `cleaned_IRR_water_right_group_2000_2021_nolema_feb16.rds`.

First, the temporal distribution of enrollments is concentrated in 2012: 608 of 753 treated water right groups (80.7 percent) entered the program in its first year, immediately following the severe 2012 drought. Within that initial cohort, GMD 5 alone supplies 326 of the 608 enrollments (53.6 percent). Later cohorts are small and scattered across years, with single-year cohort sizes of

8–29 water right groups and several GMD-by-year cells containing zero enrollments. The cohort-aggregated estimates in Section 6.2 are best identified for the 2013 cohort precisely because that cohort has a long post-enrollment window; the 2012 cohort is larger but its event-time coefficients are estimated from comparisons in which most of the eventual MYFA participants in the sample have already switched into treatment.

The credibility of this restriction depends on how the identifying variation is distributed across the five Kansas GMDs. In table 1 I report how MYFA participants are distributed across the Groundwater Management Districts. If a single GMD accounts for most of the treated water right groups and most of the active-treatment observations, the within-GMD restriction is anchored largely in that GMD and is a less powerful safeguard than the language “within-GMD comparisons” might otherwise suggest. Table 1 reports the relevant counts.

Second, GMD 5 contains 50.7 percent of treated water right groups and 52.0 percent of the active-treatment ($D = 1$) observations. This is significant because my preferred empirical specification compares MYFA enrollees to non-participating irrigators within the same Groundwater Management District. The within-GMD restriction in the preferred specification effectively pools two qualitatively different exercises: a within-GMD-5 comparison that supplies roughly half of the identifying weight, and a within-GMD-2/3/4 comparison that supplies most of the remainder. The 23 treated water right groups in GMD 1 contribute too little to identify a GMD-specific effect at conventional precision, and the 56 treated water right groups in GMD 4 are also a thin layer.

4.3 Control Variables

Weather conditions during the growing season represent the primary time-varying determinant of irrigation water demand. To address differences in precipitation and evapotranspiration between water right groups, I obtain gridded weather data from the PRISM Climate Group and calculate growing season precipitation and evapotranspiration based on water right group’s authorized place of use. I use the latitude and temperature data to calculate evapotranspiration according to the Penman-Monteith equation as in . Taken together, these variables control for the direct effect of

growing conditions on irrigation demand.

4.4 Descriptive Statistics

Table 2 presents summary statistics for the analysis sample. The mean annual water use is approximately 167 acre-feet per water right group, with substantial variation (standard deviation of 140 acre-feet). Irrigated acreage averages 161 acres, slightly more than a typical 125 acre center-pivot. Mean growing-season precipitation is 78 millimeters, though this varies considerably across the semi-arid western Kansas study region.

The panel is highly balanced: 81 percent of water right groups appear in all 21 sample years. This high retention rate reflects the stability of irrigation water rights and reduces concerns about selective attrition biasing the estimates.

Table 2: Summary Statistics

Variable	Mean	Std. Dev.
Annual water use (acre-feet)	167	140
Irrigated acres	161	160
Authorized quantity (acre-feet)	376	320
Authorized acres	248	210
Growing season precipitation (mm)	78	26
Number of water right groups	13,328	
Panel years	2001–2021	
Percent with complete panel	81.5%	

4.5 Properties of the Precipitation Distribution

The conceptual model in Section 3.4 identifies four conditions under which the option-value channel dominates the budget-flexibility channel and a soft cap reduces expected total extraction relative to a hard cap: (a) a high discount factor δ , (b) a sufficiently variable, left-tailed, or low-mean distribution F_{t+1} of growing-season precipitation shocks, (c) expected unused budget at the end of the flex window, and (d) positive autocorrelation in shocks across years. The model is formally

two-period; conditions (a) and (c) are not directly recoverable from data, but conditions (b) and (d) are. Table 3 documents that the May–September PRISM precipitation distribution in the analysis sample satisfies the variability and autocorrelation conditions the model invokes.

Table 3: Properties of May–September Precipitation, by Groundwater Management District

GMD	Mean (mm)	SD (mm)	CV	p10 (mm)	Skewness
1	69.9	19.0	0.27	43.1	−0.15
2	105.0	32.3	0.31	67.3	+0.45
3	67.2	21.3	0.32	42.5	+0.51
4	69.6	18.8	0.27	43.0	−0.07
5	90.3	24.2	0.27	57.4	+0.02
Pooled	78.4	26.7	0.34	46.7	+0.63

The distribution is empirically variable ($CV = 0.34$ pooled), with a left tail that places ten percent of the mass at or below 47 mm against a mean of 78 mm. Although the standardized skewness coefficient is positive (driven by occasional very wet years such as 2015–2016), the relevant question for the conservation channel is whether adverse-shock years are common enough to make the expected option value of saved budget large; the gap between p10 and the median (47 vs. 76 mm) confirms that adverse years occur with non-trivial probability. Year-to-year persistence is also positive: the within-water right group autocorrelation of growing-season precipitation has a mean of 0.24 and is positive in 83% of water right groups.⁴

⁴The two-period FOC in Equation (4) generalizes to a T -period setting in the natural way: at each stage $s = 1, \dots, T - 1$ the irrigator equates the marginal benefit of pumping today to a discounted expected shadow value of remaining budget, and the chain of shadow values $\mu_s = \partial V_s / \partial R_s$ is computed by backward induction from period T . The conditions for the conservation channel to dominate carry over with F_{t+1} replaced by the conditional joint distribution of $(\epsilon_{s+1}, \dots, \epsilon_T) \mid \epsilon_s$. The empirical setting in this paper corresponds to $T = 5$, and the moments documented in Table 3 bear on the joint distribution rather than on a single-period F_{t+1} . The two-period model is retained in Section 3 for analytical transparency; the qualitative predictions are unchanged in the multi-period extension because the option-value mechanism operates at every stage.

5 Empirical Analysis

5.1 Identification Strategy

I employ a difference-in-differences approach to estimate the causal effect of MYFA adoption on groundwater withdrawals. In essence, my approach compares irrigation withdrawals for participants with irrigators who never enroll in a MYFA or who enroll later, using the never-treated or not-yet-treated to approximate participants' counterfactual behavior. There are two distinct challenges with using this approach in the MYFA context. First, there is the standard selection problem and the potential for violations of the parallel trends assumption. Water right groups that choose to enroll may differ systematically from non-enrollees in ways that affect their groundwater extraction behavior over time. The second challenge is specific to the MYFA institutional design: treatment is reversible. A MYFA is a fixed-term contract that runs for approximately five years and then expires, after which the water right reverts to its original annual allocation unless the irrigator enrolls in a new MYFA.

Staggered adoption across years 2012–2021 and plausibly heterogeneous, dynamic treatment effects rule out traditional two-way fixed effects estimators (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021). To address the aforementioned identification challenges and test the predictions from my theoretical model, I make use of two complementary DiD estimators. My preferred specification is the de Chaisemartin and D'Haultfœuille (2024) estimator, as it permits treatment to be reversible. For my preferred specification, I address the possibility of selection into treatment by including growing season precipitation as a control variable and specifying that comparisons compare irrigators within the same Groundwater Management District. The second estimator I use, that of Callaway and Sant'Anna (2021), requires that treatment is an absorbing status. As such, employing the Callaway and Sant'Anna (2021) estimator involves some misclassification that combines the effects of irrigators opting out of re-enrolling in a MYFA with the impact of those who choose to re-enroll. However, I argue that the advantages of clean aggregations by adoption cohort and calendar year provided by the Callaway and

Sant’Anna (2021) approach make it the natural tool for testing the predictions concerning weather at the time of enrollment from the theoretical model. I address selection into treatment within the Callaway and Sant’Anna (2021) approach by including pre-development aquifer characteristics in the inverse probability weighting component. As such, both estimators rely on conditional parallel trends assumptions.

I investigate whether MYFA participants differ significantly in their irrigation behavior over time using the placebo and pre-treatment event-study coefficients provided by both specifications diagnostics. If the respective conditional parallel trends assumptions hold, these coefficients should be statistically indistinguishable from zero. I illustrate the implications of any deviations from such in Section 6.1 and provide a formal sensitivity analysis in Appendix D to bound the average treatment effect estimates reported in the next section. In addition, I estimate the effects of participation after restricting the control pool to only not-yet-treated water right groups. The goal of this robustness check is to address systematic differences between MYFA participants and irrigators who choose not to enroll. The results from this robustness check and others are discussed further in Section 6.3.

5.2 Preferred Specification: de Chaisemartin and D’Haultfœuille Dynamic Estimator

The de Chaisemartin and D’Haultfœuille (2024) estimator (implemented in the `DIDmultiplegt` DYN R package) is designed for settings in which units may move in and out of treatment over time. Treatment is coded as a time-varying binary indicator $D_{i,t}$, equal to one when water right group i holds an active MYFA in year t (i.e., t lies between the MYFA start and end year recorded in the administrative data) and zero otherwise, including both years before enrollment and years after expiration. Identification exploits *switcher–stayer* comparisons: at each period t , units whose treatment status changes between $t - 1$ and t (switchers) are compared to units whose status does not change (stayers), and the resulting first-differenced comparisons are aggregated across all switching events to form dynamic event-study estimates. The preferred de Chaisemartin and D’Haultfœuille (2024)

specification requires conditional parallel trends *within each GMD*: absent a change in treatment status, water right groups that switch at period t and water right groups in the same GMD that do not switch would have experienced the same average change in log water use between $t - 1$ and t , conditional on growing-season precipitation. This is a weaker assumption than conventional pooled parallel trends, as it permits GMDs to follow distinct aggregate trajectories.

The results from the de Chaisemartin and D’Haultfœuille (2024) estimator are reported as a sequence of “Effects”: Effect ℓ is the average treatment effect ℓ periods after a treatment-status switch, relative to the period immediately before the switch. Because MYFA terms are approximately five years, Effects 1–5 fall within the dominant cohort’s flex period and Effects 6+ identify *post-expiration* dynamics. The estimator also reports “Placebo” estimates in the pre-enrollment periods, which serve as a diagnostic for the parallel trends assumption. As mentioned above, I utilize the nonparametric restriction of this estimator to ensure comparisons are made within Groundwater Management Districts the comparison set. The five Kansas GMDs differ materially in hydrology, aquifer thickness, crop mix, and historical water-use trajectories, so within-GMD comparisons remove a substantial source of potential confounding by unobserved regional trends. In the preferred specification, I impose this restriction and include growing-season precipitation (May–September rainfall from PRISM) as the sole time-varying control. Finally, I cluster the standard errors at the water right group level.

5.3 Complementary Specification: Callaway and Sant’Anna Group-Time Estimator

I also estimate the Callaway and Sant’Anna (2021) group-time estimator, which defines cohort-by-year average treatment effects,

$$ATT(g, t) = E[Y_{i,t}(g) - Y_{i,t}(0) \mid G_i = g], \quad (5)$$

where g denotes the year of first MYFA enrollment and t denotes the calendar year. Each $ATT(g, t)$ is identified from a 2×2 DiD comparing cohort g to a comparison group of irrigators who never participate or who are not-yet-treated by time t . The ability to flexibly aggregate $ATT(g, t)$'s makes it the natural estimator for testing Prediction 2 of the conceptual model (Section 3.4): cohorts that enter the MYFA during drought years should display attenuated treatment effects because the adverse current-period shock leaves little scope to reduce extraction below agronomic needs and prevents accumulation of banked allocation in early flex years. I report the $ATT(g, t)$'s aggregated by event-time $e = t - g$ as a robustness check for the main results concerning the average treatment effect and dynamic effects, and I use the results aggregated by cohort year g and calendar-year t to test whether weather at the time of enrollment influences the observed effects.

I use the doubly-robust version of the Sant'Anna and Zhao (2020) estimator, which combines outcome regression adjustment with inverse probability weighting and is consistent if either the outcome or the propensity-score model is correctly specified. For controls, I include the average growing-season precipitation for water right groups over all pre-enrollment periods and pre-development measures of aquifer conductivity. I use the multiplier bootstrap to construct the standard errors with 1,000 replications. Just as for my preferred specification, I cluster the standard errors at the water right group level.

6 Results and Discussion

I organize the results in three parts. Section 6.1 presents the average and dynamic treatment effects. The preferred de Chaisemartin and D'Haultfœuille (2024) event-study estimates reported there constitute the main causal evidence for a reduction in groundwater extraction following MYFA enrollment, and the Callaway and Sant'Anna (2021) group-time estimator provides an independent corroborating check whose full event study is reported in Appendix A. Section 6.2 exploits the group-time aggregations of the Callaway and Sant'Anna (2021) estimator to depict the treatment effect by adoption cohort and by calendar year, the central exercise for probing the

weather-at-enrollment mechanism. Section 6.3 collects the robustness checks: the distribution of identifying variation across groundwater management districts, selection into enrollment and the choice of comparison group, an alternative that conditions on pre-treatment water use in place of the within-GMD restriction, formal sensitivity bounds on the residual pre-trend, and the parallel estimates for irrigation intensity.

6.1 Average and Dynamic Treatment Effects

Figure 1 and Table 4 report the de Chaisemartin and D’Haultfœuille (2024) dynamic event-study estimates under the preferred specification: reversible treatment coded from MYFA start and end dates, within-GMD switcher–stayer comparisons, and growing-season precipitation as a time-varying control. Effect ℓ is the average treatment effect ℓ periods after a change in treatment status; Placebo ℓ is the analogous pre-switch estimate.

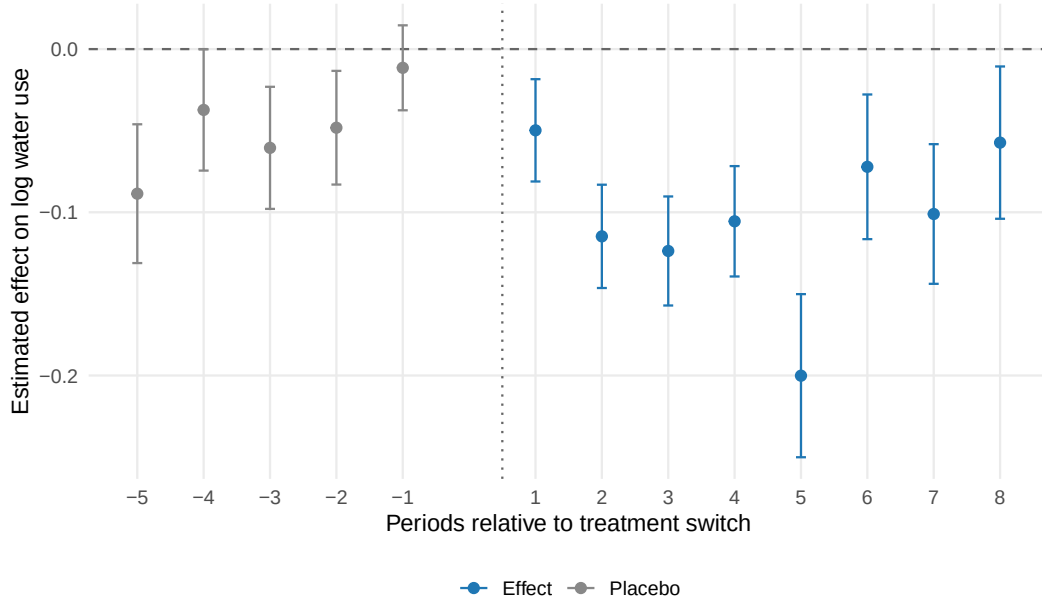


Figure 1: Preferred Specification: Dynamic DiD Event Study of MYFA Effect on Log Groundwater Withdrawals

Notes: Figure displays de Chaisemartin and D’Haultfœuille (2024) dynamic DiD estimates of the effect of MYFA enrollment on log annual water use. Points at negative event times are Placebo estimates; points at positive event times are Effects. Effects 1–5 lie within the five-year flex period for the dominant enrollment cohort; Effects 6–8 identify post-expiration dynamics that an absorbing-state estimator cannot recover. Vertical bars are 95% confidence intervals. Comparisons restricted to switchers and stayers within the same Groundwater Management District. Time-varying control is May–September PRISM precipitation at the water right group centroid. Standard errors clustered at the water right group level. Sample: water right groups in GMDs, 2001–2021, excluding LEMAs, deduplicated panel ($N = 239,947$).

All eight estimated Effects are negative and statistically significant. The point estimates range from -0.050 log points in the first post-switch year (approximately a 4.9 percent reduction) to -0.200 log points in the fifth year (approximately an 18 percent reduction). The average ATT across the estimated Effects is -0.133 log points with a standard error of 0.016, corresponding to a reduction in annual groundwater withdrawals of approximately 12.5 percent.

Three features of the estimates merit emphasis. First, the effects emerge immediately upon enrollment (Effect 1) rather than at a lag, consistent with irrigators responding to the removal of the

“use it or lose it” incentive as soon as they enter the MYFA. Second, effects remain negative and statistically significant for Effects 6, 7, and 8, which for the dominant 2012 cohort correspond to years *after* MYFA expiration. This pattern—which an absorbing-state estimator cannot identify—suggests that conservation behavior initiated during the flex period persists after the formal contract has ended, possibly through water-saving investments or habit formation made during the MYFA years. Third, the estimated standard errors are substantially smaller than those produced by the Callaway and Sant’Anna (2021) estimator, because the first-differenced switcher–stayer comparison pools information across all switching events rather than estimating group-time ATTs separately. This precision is necessary to detect the relatively small per-period effects that, when aggregated, deliver the headline 12.5 percent reduction.

The Placebo estimates indicate partial but incomplete support for the within-GMD parallel trends assumption. Placebo 1 is small and insignificant (-0.011), but Placebos 2–5 range from -0.037 to -0.089 and are individually significant, with a joint test against zero rejected at $p = 0.0002$. The magnitudes of these violations are substantially smaller than under a pooled (non-GMD-restricted) specification, in which the peak placebo reaches -0.128 ; restricting comparisons within GMD cuts the worst-case violation roughly in half. A residual violation remains, however, and two interpretations are consistent with it. First, irrigators who enroll in MYFAs may be self-selecting on a modestly declining water-use trajectory—operators facing rising pumping costs or declining well yields may be more likely both to reduce extraction and to enroll. Second, the “pre-trend” itself may be partly a treatment response if operators begin conserving in anticipation of enrollment. I return to this tension in Section 6.3.

Table 4: Preferred Specification: Dynamic DiD Estimates of MYFA Effect on Log Groundwater Withdrawals

	Estimate	Std. Error
<i>Post-Switch Effects (within flex period: Effects 1–5)</i>		
Effect 1	−0.050**	0.016
Effect 2	−0.115***	0.016
Effect 3	−0.124***	0.017
Effect 4	−0.106***	0.017
Effect 5	−0.200***	0.026
<i>Post-Switch Effects (post-expiration: Effects 6–8)</i>		
Effect 6	−0.072**	0.023
Effect 7	−0.101***	0.022
Effect 8	−0.057*	0.024
Average ATT across Effects 1–8	−0.133***	0.016
<i>Placebo Estimates (pre-switch)</i>		
Placebo 1	−0.011	0.013
Placebo 2	−0.048**	0.018
Placebo 3	−0.061**	0.019
Placebo 4	−0.037*	0.019
Placebo 5	−0.089***	0.022
Joint placebo test (p -value)	0.0002	

A note on the Effect 5 magnitude. The Effect 5 coefficient of -0.200 log points (s.e. 0.026) is the largest in absolute value among the eight estimated post-switch effects, substantially outside the -0.106 to -0.124 range covered by Effects 2 through 4 and the -0.057 to -0.101 range covered by Effects 6 through 8. Two features of the panel structure plausibly drive this magnitude. First, the dCdH switcher–stayer comparison at event time $+5$ is heavily weighted toward water right groups in the 2012 enrollment cohort, which contains 608 of the 753 ever-treated water right groups in the sample (Section 4.2) and is the only cohort with enough post-enrollment horizon to contribute at $\ell = 5$ before the panel runs out in 2021. Second, event time $+5$ for the 2012 cohort coincides with calendar year 2017, and the mid-flex-window calendar years are precisely those for which the Callaway–Sant’Anna calendar-year aggregation in Table 7 reports the largest effects; the 2016 aggregate, for instance, is -0.141 in the baseline and -0.326 under not-yet-treated controls. The

Callaway–Sant’Anna event-time estimate at $e = 5$ is also unusually large at -0.152 (Table 6), so the pattern is not specific to the dCdH switcher–stayer aggregation. The likely interpretation is that the conservation channel operates most strongly in years when banked allocation has accumulated through several preceding favorable seasons (the 2012 cohort would have accumulated four years of banked allocation by 2017) and is then drawn down asymmetrically against an adverse weather realization. The dCdH estimator does not natively support a per-cohort decomposition of dynamic Effects, but the convergence of the dCdH and Callaway–Sant’Anna estimates at this horizon and the calendar-year aggregation findings together suggest the Effect 5 magnitude is within the range of cohort $\times$ weather variability documented elsewhere in the paper, rather than an artifact of the estimator.

The Callaway and Sant’Anna (2021) group-time estimator described in Section 5.3 provides an independent check on these magnitudes, and it corroborates the preferred specification. Its dynamic aggregate average treatment effect is -0.116 log points (s.e. = 0.022) under the baseline comparison group and -0.242 (s.e. = 0.034) when the comparison set is restricted to not-yet-treated water right groups, bracketing the preferred estimate of -0.133 ; the simple overall ATT is -0.118 (s.e. = 0.020). The event-time profile has the same shape as the dCdH profile: the enrollment-year effect is -0.134 (s.e. = 0.026), an approximate 12.5 percent reduction at the point of enrollment, and the post-enrollment coefficients are negative throughout. None of the pre-treatment coefficients is statistically distinguishable from zero, though the group-time estimator’s pre-treatment standard errors are roughly twice those of the preferred specification, so this reflects limited power rather than strong evidence for parallel trends. The two estimators therefore agree on sign and on approximate magnitude and differ mainly in precision. The full event study appears in Figure 7 and the underlying coefficients in Table 6, both reported in Appendix A.

6.2 Cohort and calendar year estimates

The cohort-aggregation capability of the Callaway and Sant’Anna (2021) estimator provides the cleanest tool for testing Prediction 2 of the conceptual model (Section 3.4): MYFA participants

who enrolled under drought conditions should reduce withdrawals less than those who enrolled in wetter years. The model attributes this asymmetry to the option-value mechanism in equations (3)–(4): cohorts entering the program when rainfall is scarce face an adverse current-period shock that pushes their unconstrained optimum close to the per-period cap, leaving little scope to reduce extraction below agronomic needs and preventing accumulation of banked allocation in early flex years. Cohorts entering under adequate precipitation can immediately allocate $q_t < \bar{Q}$ and bank the residual, generating the largest expected reduction in total use. Figure 2 presents treatment effects disaggregated by enrollment cohort, where each cohort represents water right groups that first enrolled in a MYFA in the same year.

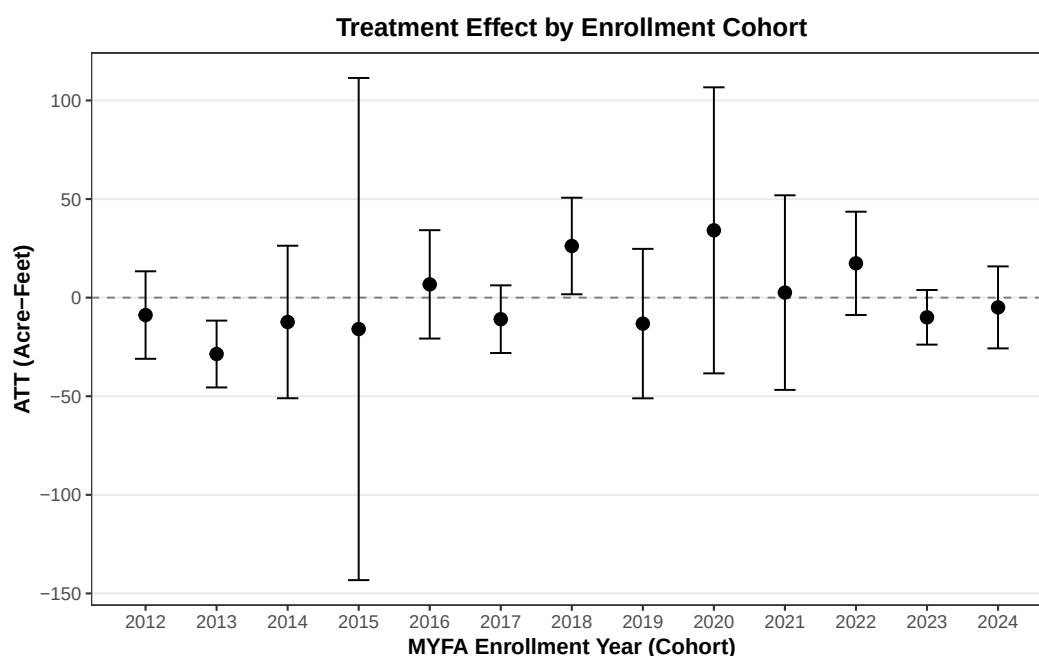


Figure 2: Treatment Effect by MYFA Enrollment Cohort

Notes: Figure displays Callaway and Sant’Anna (2021) group-level average treatment effects, where groups are defined by year of first MYFA enrollment. Vertical bars show 95% confidence intervals based on multiplier bootstrap with clustering at the water right group level. Estimation uses the doubly-robust method with not-yet-treated comparison group and controls for growing-season precipitation.

The overall cohort-aggregated treatment effect is -0.119 log units (s.e. = 0.020, $p < 0.001$)

in the baseline specification, corresponding to approximately an 11 percent reduction. The not-yet-treated-only specification yields a larger estimate of -0.228 log units (s.e. = 0.032), approximately a 20 percent reduction. The cohort-specific estimates reported in Table 7 require careful interpretation. The 2013 cohort, which enrolled following the severe 2012 drought, is the most precisely estimated cohort—the 2012 cohort is far larger in membership, but its event-time coefficients are identified from comparisons in which most eventual participants have already switched—with $\theta_{2013} = -0.121$ (s.e. = 0.023) in the baseline and -0.254 (s.e. = 0.035) under not-yet-treated controls. Both confidence intervals exclude zero comfortably. Later cohorts (2018, 2020) show point estimates that are larger in absolute magnitude than the 2013 estimate (-0.284 and -0.243 in the baseline) but with standard errors roughly five times larger (0.105 for both cohorts), reflecting the small number of water right groups in each cohort and the limited number of post-enrollment years of data available for cohorts that enrolled near the end of the panel.

The cross-cohort comparison therefore yields a more nuanced picture than a simple ordering of point-estimate magnitudes would suggest. On one hand, the 2013 cohort, which enrolled in a year of recovering precipitation, is the only cohort whose effect is precisely identified in both control-group variants. The wide confidence intervals on later cohorts mean I cannot reject the hypothesis that all cohorts share a common treatment effect; the test for cross-cohort heterogeneity is statistically inconclusive. On the other hand, the 2013 cohort’s identification is the cleanest precisely because it provides a long post-enrollment window (eight years through 2021) over which the option-value channel has time to operate; the late drought-year cohorts (2018, 2020) have only one to three post-enrollment years of data, much of which falls within the early-flex-period window where Prediction 2 implies the conservation effect should be smallest. Treating Prediction 2 as a cleaner test of the model would therefore require comparing event-time-zero or event-time-one ATTs across cohorts, where the model’s mapping is sharpest, rather than the cohort-aggregated θ_g . The calendar-year results I report next provide a useful complementary perspective on this question: by aggregating across cohorts within each calendar year, they isolate the effect of contemporaneous weather conditions on the strength of the treatment response.

To summarize, the cohort decomposition is consistent with Prediction 2 only in the limited sense that the best-identified cohort (2013) is one that enrolled in a wet recovery year, as the model would predict. I cannot demonstrate cross-cohort ordering of magnitudes statistically, and I therefore treat the cohort heterogeneity result as suggestive rather than confirmatory. The calendar-year analysis below offers stronger evidence on the weather-mediated mechanism.

The same group-time estimates can be aggregated along the calendar-time dimension rather than the cohort dimension. Because calendar-year aggregation pools every enrolled water right group active in a given year, it is estimated far more precisely than the cohort decomposition and carries correspondingly more weight in what follows.

Figure 3 presents treatment effects aggregated by calendar year, showing the average effect of MYFA enrollment on water use in each year across all enrolled cohorts. This aggregation captures how the overall treatment effect varies with annual conditions affecting all irrigators.

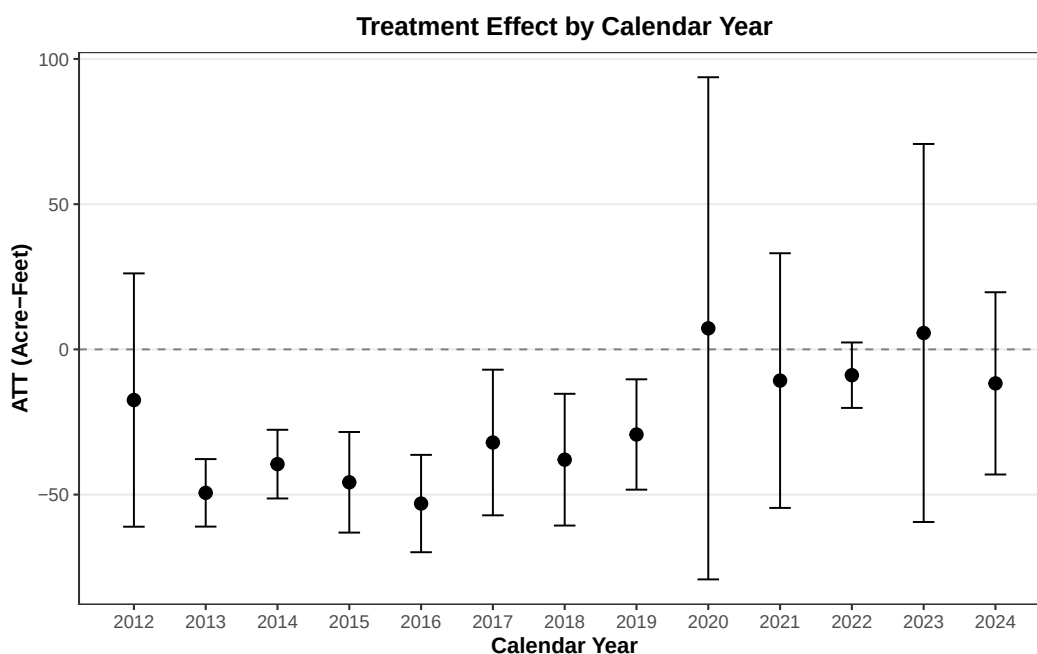


Figure 3: Treatment Effect by Calendar Year estimated with Callaway and Sant’Anna (2021) approach. Vertical bars show 95% confidence intervals based on multiplier bootstrap with clustering at the water right group level. Estimation uses the doubly-robust method with time invariant growing-season precipitation included as a control.

The overall calendar-year aggregated treatment effect is -0.113 log units (s.e. = 0.018, $p <$

0.001) in the baseline, approximately an 11 percent reduction. The not-yet-treated-only specification yields -0.217 log units (s.e. = 0.029), approximately a 20 percent reduction. The calendar year estimates reveal a pronounced temporal pattern in both specifications: treatment effects are largest and most precisely estimated in the years 2013–2019, with the 2016 calendar year showing among the largest effects (-0.141 baseline; -0.326 not-yet-treated only). The earlier year 2012 shows smaller and less precisely estimated effects, reflecting fewer treated observations in the program’s first year.

The strong negative effects during 2013–2019 coincide with a period of variable but generally adequate precipitation in western Kansas, interspersed with drought years. This pattern suggests that MYFAs are most effective at reducing extraction when enrolled irrigators experience a mix of wet and dry conditions, enabling them to bank water during favorable years and draw down savings during stress periods rather than maintaining consistently high extraction.

Table 7 in Appendix A reports the underlying cohort and calendar-year coefficients for both comparison-group variants, alongside the event-time estimates in Table 6. Read together, the two aggregations point the same way: the conservation response is concentrated in years and among cohorts for which precipitation left allocation worth banking, and is muted when scarce rainfall makes agronomic water demand binding. The two routes do not carry equal weight, however. The calendar-year aggregation identifies the pattern precisely, whereas the cross-cohort ordering of point estimates is not a cleanly identified test, so the weather-mediated interpretation rests primarily on the calendar-year evidence.

6.3 Robustness checks

Figures 5 and 6 display graphically how the average treatment effect estimate varies across the specifications and checks discussed in this subsection, for the preferred estimator and the group-time estimator respectively. Each row reports one estimate with its 95 percent confidence interval against a vertical line at the preferred estimate of -0.133 log points; every estimate is negative, and no interval includes zero.

The identifying variation behind the preferred specification is not evenly distributed across groundwater management districts. The share of total water right group-year observations differs noticeably from the share of treated water right groups across GMDs. GMD 3 holds 38.9 percent of water right group-year observations but only 21.4 percent of treated water right groups, while GMD 5 holds 29.2 percent of water right group-year observations and 50.7 percent of treated water right groups. The first-differenced switcher–stayer comparisons in the preferred specification weight by the abundance of stayers within each GMD relative to switchers; GMD 3 supplies a large stayer pool against a modest switcher count, while GMD 5 supplies a more even mix. The differential weighting matters for interpreting any heterogeneity across GMDs. Appendix D.3 reports a leave-one-GMD-out robustness check on the headline preferred-specification estimates: the headline is robust in direction and significance to dropping any single GMD, but its magnitude is materially driven by GMD 5 (dropping GMD 5 attenuates the log water-use average total effect by about 29 percent and the depth-applied average total effect by about 33 percent, while leaving both individually significant). This disclosure and the leave-one-GMD-out robustness checks in Appendix D.3 serve as complementary diagnostics that together let the reader judge how much of the headline reflects pooled-across-GMD averaging versus GMD-5-specific dynamics.

A central concern in evaluating voluntary programs like MYFAs is selection into treatment. Irrigators who choose to enroll in a MYFA may differ from non-participants in ways that affect groundwater use independently of the program, either on time-invariant dimensions (underlying conservation orientation, aquifer access, management intensity) or on time-varying dimensions (emerging well yield problems, changing crop portfolios, anticipated regulatory actions).

The preferred de Chaisemartin and D’Haultfœuille (2024) specification addresses the cross-unit dimension of this concern through the within-GMD restriction introduced in Section 5.2. Comparing each switcher only to stayers in the same Groundwater Management District eliminates confounding from cross-GMD differences in hydrology, aquifer thickness, and crop mix that would otherwise load onto the estimated treatment effect. The contraction of placebo magnitudes under this restriction, reported in Section 6.1, is direct evidence that cross-GMD heterogeneity is

a quantitatively important confound in pooled specifications; the placebos that remain are jointly significant, and that residual violation warrants the cautious interpretation of magnitudes adopted throughout.

A complementary concern, which the within-GMD restriction does not directly address, is that water right groups that *never* enroll in a MYFA may be systematically different from eventual participants in ways that independently shape their water-use trajectories. If never-treated water right groups are systematically more conservative users, their inclusion in the comparison pool biases the estimated treatment effect toward zero. The de Chaisemartin and D’Haultfœuille (2024) estimator already handles this concern partially: switcher–stayer comparisons at any given period include only units that are not themselves switching at that period, so the comparison set rotates through units based on their treatment history rather than relying on a fixed never-treated pool. To assess the residual sensitivity within the Callaway and Sant’Anna (2021) framework, I also report a not-yet-treated-only variant that excludes never-treated water right groups from the control pool entirely. Under this variant, all comparisons are made among water right groups that will eventually enroll in a MYFA; the “control” observations are simply the same types of operators in periods before their own enrollment began.

Figure 4 presents the event study estimates from both specifications. The pre-treatment coefficients are small under both, and all but one (the baseline coefficient at $e = -2$) are statistically indistinguishable from zero; as noted in Appendix A, however, the group-time estimator’s pre-treatment standard errors are roughly twice those of the preferred specification, so this is weak evidence on parallel trends rather than strong support for it. The post-treatment estimates are uniformly more negative under the not-yet-treated-only specification—the estimated reduction at enrollment (event time 0) is -0.134 in the baseline versus -0.167 not-yet-treated only, and the discrepancy widens at longer event times, reaching -0.134 versus -0.303 at event time 3.

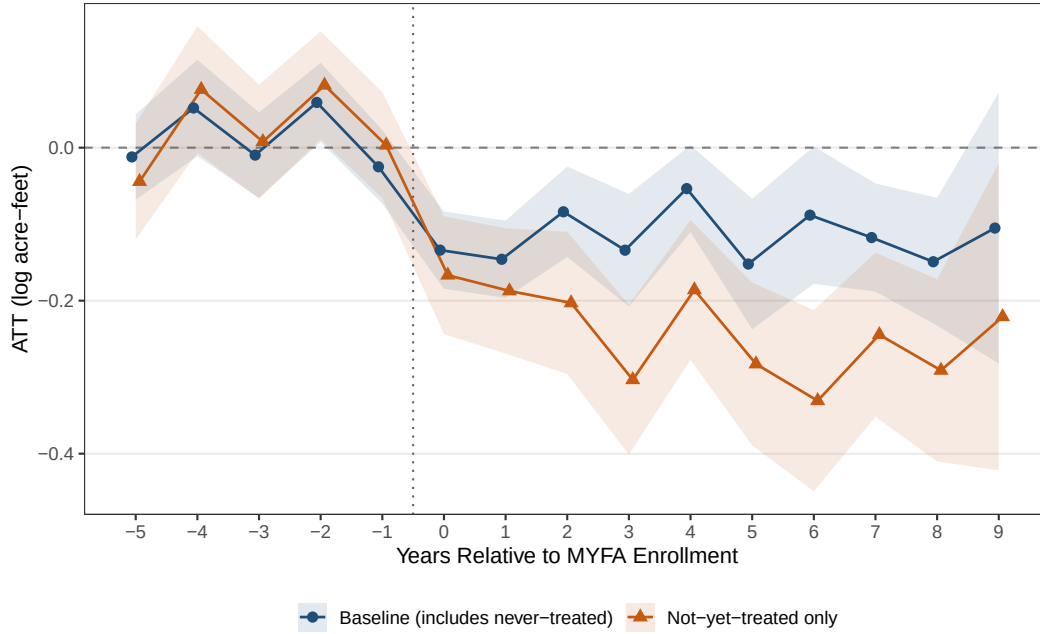


Figure 4: Event Study Estimates: Baseline vs. Not-Yet-Treated Only

Notes: Figure displays Callaway and Sant’Anna (2021) event study estimates under two control group specifications. *Baseline (includes never-treated)* uses all not-yet-treated and never-treated water right groups as controls. *Not-yet-treated only* restricts the control group to eventual MYFA participants not yet enrolled, excluding all water right groups that never enroll. Both specifications use the doubly-robust method with controls for growing-season precipitation and hydraulic conductivity and 1,000 bootstrap replications. Shaded bands show pointwise 95% confidence intervals.

The consistent strengthening of the estimated effect when never-treated water right groups are excluded is the pattern one would expect if never-treated operators use water more conservatively than eventual MYFA participants even before any enrollment occurs. In that scenario, their inclusion attenuates the estimated ATT because their below-average water use creates a favorable baseline that makes the treated group’s reduction appear smaller. The not-yet-treated-only estimates, which compare enrollees to their own pre-enrollment counterparts (i.e., other eventual enrollees not yet in the program), avoid this confound.

I treat the not-yet-treated-only specification as the preferred robustness check while reporting both in the main tables, for two reasons. First, the selection concern is plausible a priori: oper-

ators who never enroll despite eligibility likely have systematically different water management practices. Second, the not-yet-treated-only pre-trends are, if anything, more supportive of parallel trends, with pre-treatment coefficients closer to zero. The robustness specification therefore addresses the selection concern while maintaining the identifying assumption.

The preferred de Chaisemartin and D’Haultfœuille (2024) specification weakens the parallel trends assumption through the `trends_nonparam = "gmd"` option, which restricts switcher–stayer comparisons to units in the same Groundwater Management District and thereby absorbs cross-GMD heterogeneity in baseline water-use trajectories. The Callaway and Sant’Anna (2021) estimator as implemented in the `did` R package does not offer an analogous nonparametric within-stratum trend restriction: its comparison set is defined by treatment history (“not yet treated”) rather than by a secondary partition of the sample. Imposing a within-GMD restriction in the CS framework by estimating the model separately for each of the five GMDs and pooling is impractical here, because three of the GMDs contain too few eventual MYFA participants to support cohort-level $ATT(g, t)$ identification.

An alternative that closes some of the same gap is to condition directly on each water right group’s pre-treatment level of water use. Because the within-GMD restriction’s role is to remove cross-unit differences in baseline water-use trajectories, including a time-invariant summary of each water right group’s pre-treatment water use as a covariate in the doubly-robust outcome regression should accomplish much of the same absorption when trajectory differences are linear in the pre-period level. Concretely, I compute each water right group’s mean of $\ln(\text{acre-feet used})$ over 2001–2011—a window that is strictly pre-enrollment for every MYFA cohort in the sample—and add this pre-treatment mean to `xformula` alongside growing-season precipitation and hydraulic conductivity.

Table 5 reports the results. Conditioning on pre-treatment water use materially attenuates both the pre-treatment event-study coefficients and the aggregated post-treatment ATT. The simple overall ATT falls from -0.118 log points (SE 0.020) in the baseline CS specification to -0.087 log points (SE 0.014) with the pre-trend control, and the event-study aggregate falls from

−0.116 to −0.086. Pre-treatment coefficients tighten considerably: their range narrows from $[-0.025, +0.059]$ to $[+0.007, +0.038]$ and their standard errors shrink by approximately 40 percent. The attenuation of the post-treatment ATT is consistent with the interpretation that part of the unconditional CS effect reflects cross-water right group differences in baseline water-use levels correlated with MYFA enrollment, rather than the program itself. Reassuringly, the pre-trend-controlled CS estimate (−0.087 log points, or approximately 8.3 percent) is much closer to the preferred dCdH estimate (−0.133 log points, 12.5 percent) than the uncontrolled CS estimate was, narrowing the gap between the two estimators by roughly half. A residual gap remains, plausibly because the pre-period level is only a partial proxy for the cross-unit trajectory heterogeneity that the dCdH within-GMD option absorbs nonparametrically.

Table 5: Callaway and Sant’Anna Estimates With and Without Pre-Treatment Water-Use Control

	Baseline		+ Pre-Treatment Control	
	Estimate	Std. Error	Estimate	Std. Error
<i>Aggregate Estimates</i>				
Simple overall ATT	−0.118***	0.020	−0.087***	0.014
Event study (dynamic)	−0.116***	0.021	−0.086***	0.013
<i>Pre-Treatment Event-Time Coefficients</i>				
$e = -5$	−0.012	0.028	0.012	0.018
$e = -4$	0.052	0.032	0.007	0.019
$e = -3$	−0.010	0.029	0.038	0.020
$e = -2$	0.059*	0.025	0.030	0.016
$e = -1$	−0.025	0.026	0.026*	0.013

For completeness, I also estimated the analogous pre-treatment trend interaction (pre-period mean $\times$ year) as a time-varying control in the de Chaisemartin and D’Haultfœuille (2024) preferred specification. In that setting, the interaction changes every reported Effect and Placebo by less than 0.001 log points and leaves the joint placebo p -value unchanged at 0.0002. This is the expected result: the within-GMD restriction already absorbs the cross-unit trajectory variation that a linear pre-period interaction would control for, so the two devices are substitutes rather than complements. Their substitutability is precisely why conditioning on pre-treatment water use is

the appropriate CS-side analog to within-GMD comparisons in the dCdH-side analysis.

The headline 12.5% should nevertheless be read as a benchmark estimate rather than a precise point. The within-GMD specification leaves a residual placebo violation (joint $p = 0.0002$ in Table 4), and Appendix D reports a formal sensitivity analysis using two complementary frameworks. The HonestDiD bounds of Rambachan and Roth (2023) indicate that the headline 95% confidence interval loses significance under modestly restrictive assumptions on the post-period trend deviation (smoothness threshold $M \approx 0.025$ log points per period; relative-magnitudes threshold $\bar{M} \approx 0.4$, below the conventional benchmark of $\bar{M} = 1$). A “one step up” residualization in the spirit of Bilinski and Hatfield (2024), which subtracts a unit-specific linear pre-trend from each water right group before re-estimation, yields a treatment effect of -0.229 log points (joint placebo $p = 0.31$, equivalence test passed) — larger in magnitude than the headline rather than smaller. I retain the dCdH within-GMD specification as preferred because the residualization approach requires assuming a specific functional form for the counterfactual trajectory (linear extrapolation of unit-specific pre-trends), and Appendix D documents that the residualized estimate is sensitive to this choice: a quadratic per-water right group trend produces an implausible -1.42 ATT through extrapolation instability, while an aggregate GMD-level linear trend leaves the result unchanged from baseline. The dCdH preferred specification, by contrast, requires no functional-form commitment about the counterfactual but accepts a known residual violation. The two approaches bracket the program effect: the headline -0.133 sits between the -0.04 implied by treating the pre-trend as 100% selection and the -0.23 implied by treating the pre-trend as program-irrelevant trajectory continuation. The headline is therefore best read as a conservative midpoint — if anything, biased downward in magnitude — rather than an upper bound on the true program effect.

Applying the same preferred specification to depth applied—acre-feet of water used divided by irrigated acres—confirms that the reduction in total withdrawals operates on the intensive margin. The average treatment effect across the ten estimated Effects is -0.192 acre-feet per acre (s.e. = 0.020), a 16.7 percent reduction relative to the pre-treatment mean of 1.148 acre-feet per acre, with negative point estimates in every post-switch period and persistence into the years fol-

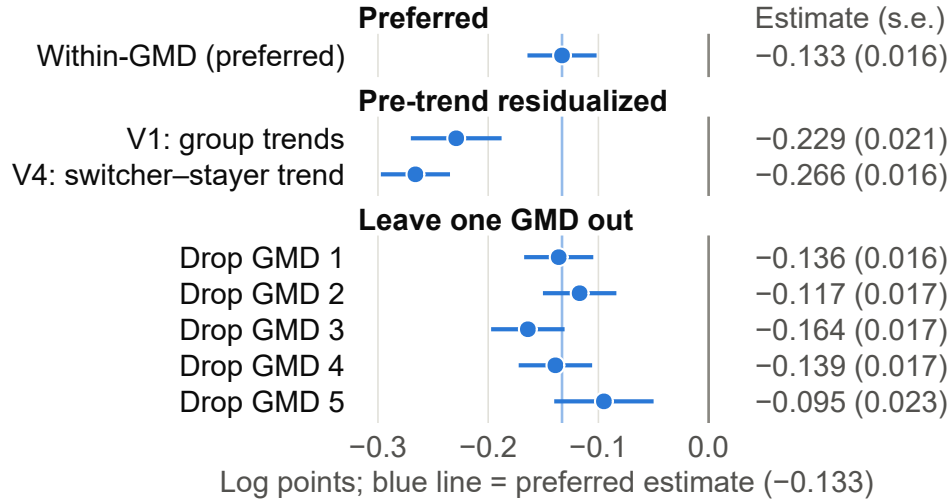


Figure 5: Robustness of the Average Effect: de Chaisemartin and D'Haultfœuille Specifications

Notes: Average post-treatment effect of MYFA enrollment on $\ln(\text{acre-feet used})$ from the de Chaisemartin and D'Haultfœuille (2024) estimator. *Preferred:* within-GMD specification (Table 4). *Pre-trend residualized:* Bilinski-Hatfield residualizations from Appendix D.2; V1 subtracts a unit-specific linear pre-trend fit on each water right group's pre-treatment observations, and V4 subtracts a single switcher-vs-stayer linear trend differential. *Leave one GMD out:* the preferred specification re-estimated after dropping each GMD in turn (Table 12). Horizontal bars show 95% confidence intervals; the vertical line marks the preferred estimate (-0.133).

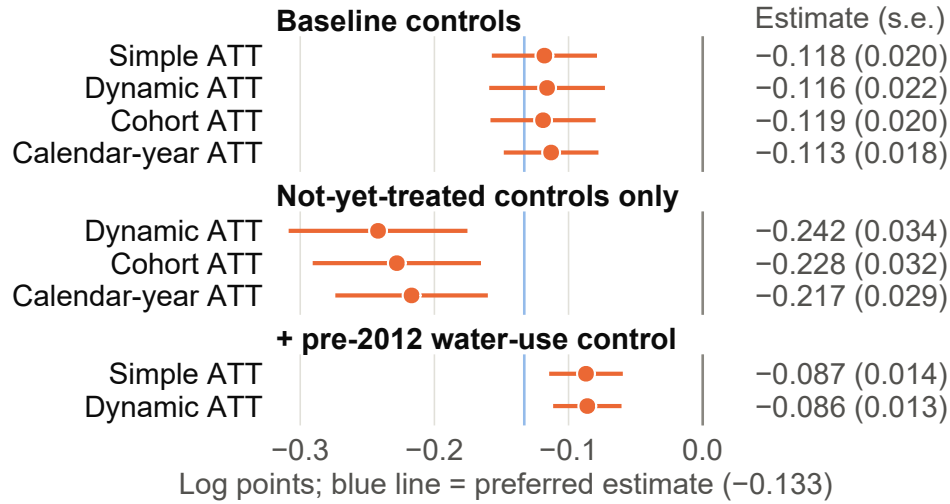


Figure 6: Robustness of the Average Effect: Callaway and Sant'Anna Aggregations

Notes: Aggregated ATTs on $\ln(\text{acre-feet used})$ from the Callaway and Sant'Anna (2021) estimator. *Baseline controls:* not-yet-treated and never-treated water right groups as controls, with growing-season precipitation and hydraulic conductivity as covariates (Appendix A). *Not-yet-treated controls only:* never-treated water right groups excluded from the control pool (Figure 4). *+ pre-2012 water-use control:* each water right group's 2001–2011 mean of $\ln(\text{acre-feet used})$ added to the covariates (Table 5). Horizontal bars show 95% confidence intervals; the vertical line marks the preferred dCdH estimate (-0.133) from Figure 5.

lowing expiration of the account. Enrollees apply substantially less water per acre while holding their irrigated footprint steady, which indicates improved irrigation efficiency rather than fallowing. The placebo estimates for this outcome are also negative and jointly significant, so the same caution about a residual pre-trend applies; Appendix B.2 repeats the HonestDiD and Bilinski–Hatfield machinery on depth applied and reaches the same qualitative conclusion. Figure 8 and Table 8 in Appendix B report the full set of estimates.

7 Conclusion

This paper asks what happens to groundwater withdrawals when irrigators are permitted to move water across time. Kansas’s Multi-Year Flex Account program supplies the setting for an empirical test: enrollment replaces the annual limit attached to a water right with a five-year aggregate allocation, granting a bounded and non-tradable form of the intertemporal arbitrage that water banks provide, and thereby isolating the effect of temporal flexibility alone on the quantity of water extracted. I approach the question in two steps. The two-period model developed in Section 3 contrasts the per-period “hard cap” embedded in a standard Kansas water right with the multi-period “soft cap” embedded in the MYFA, and it shows that the two regimes differ in one respect that governs behavior: under the soft cap, allocation left unused acquires a shadow value in the following period. That option value enters the irrigator’s first-order condition as an opportunity cost of extracting today, mutes the response to favorable current-period precipitation, and induces precautionary saving—prepositioning for drought—whenever adverse future shocks are sufficiently likely. The model also identifies the conditions on the precipitation distribution under which this conservation channel dominates the opposing incentive to draw the account down in dry years, so that total extraction falls rather than merely shifting in time: shocks must be variable, left-tailed, and positively autocorrelated, and irrigators must be sufficiently patient. Western Kansas satisfies those conditions, and the empirical estimates align with what the model predicts under them.

The estimates come from administrative withdrawal records for more than 13,000 water right

groups observed from 2001 to 2021, analyzed with the de Chaisemartin and D’Haultfoeuille (2024) dynamic difference-in-differences estimator, which accommodates the reversibility of MYFA enrollment, and complemented by the Callaway and Sant’Anna (2021) group-time estimator for aggregation by cohort and calendar year. Enrollment reduces annual withdrawals by approximately 12.5 percent on average, with negative effects in every estimated event-time period and persistence into the years following expiration of the account; Appendix D shows that this headline is best read as a conservative midpoint, since specifications that absorb the residual pre-trend imply larger reductions still. The reduction operates through water applied per irrigated acre, which falls by roughly 17 percent while irrigated area holds steady, indicating improved efficiency rather than fallowing, and it is concentrated in years whose precipitation was mixed-to-favorable, the conditions under which an irrigator has unused allocation worth banking in the first place. These results extend what is known about water banks in a specific direction. The banking and market literature has evaluated such institutions on the efficiency with which a fixed quantity of water is allocated across users, sectors, and seasons; the estimates here indicate that the temporal component of that arbitrage also changes the quantity withdrawn, and changes it downward. Flexibility, in other words, is not merely a device for lowering the cost of meeting a conservation target—it can be a conservation instrument in its own right, and one with practical advantages over the alternatives, since a MYFA is voluntary, preserves the underlying right, requires no mandatory cutback, and leaves enrollees weakly better off by their own revealed assessment. The same flexibility, however, permits the concentration of pumping in dry years, and the analysis here cannot rule out that possibility: the cohort estimates for irrigators enrolling under drought are too imprecise to order, and the average effect over a five-year account conceals its within-account timing. Where groundwater and surface water are hydrologically coupled, and where peak drawdown rather than the multi-year total governs the environmental damage, MYFA-like policies may therefore need to be paired with a limit on annual pumping that caps the drought-year drawdown while leaving the option value of banked allocation intact (Young et al., 2021).

The magnitudes are economically meaningful. Applied to the pre-treatment mean of 167 acre-

feet per water right group, a 12.5 percent reduction is roughly 21 acre-feet per group per year; multiplied across the approximately 750 ever-enrolled water right groups in the non-LEMA analysis sample, the implied annual aggregate reduction is on the order of 15,000 acre-feet, comparable to the annual water use of a small Kansas municipality. That the average is negative at all is the substantive point. Had MYFAs merely allowed irrigators to redistribute a fixed total across the years of the flex window, the average post-enrollment effect would be approximately zero; the uniformly negative post-switch estimates, including those in the years following expiration of the account, indicate net conservation rather than pure intertemporal reallocation.

Several limitations qualify these conclusions and mark the natural directions for further work. Identification rests on an observational design in which a residual within-GMD pre-trend remains after the strongest available restriction on the comparison set, so the point estimates should be read as a bracketed range rather than a single number. The data record water use but not yields, costs, or profits, which leaves the welfare interpretation resting on a revealed-preference argument from voluntary enrollment rather than on measured outcomes, and they contain no hydrologic outcomes, so the aquifer and streamflow consequences of the estimated reductions must be inferred rather than observed. Enrollment is voluntary and may select conservation-inclined operators, so the effects estimated here need not extrapolate to a mandatory or universal program. The panel also ends in 2021, before most accounts complete a full cycle of expiration and renewal, and before the drought-year cohorts accumulate enough post-enrollment observations to test the model's cohort ordering with meaningful power. Extending the panel would address both gaps at once. Beyond that, linking withdrawal records to yield and profit data would allow the welfare claim to be estimated rather than bounded; coupling the estimated behavioral response to a hydrologic model would establish whether the drought-year drawdown a soft cap permits is small enough to preserve the conservation gain in connected systems; and comparing multi-year accounts with designs that allow banked allocation to be traded would identify how much of the potential gain from water banking the temporal margin captures on its own. One of these qualifications bears directly on how the headline numbers should be read. The 12.5 percent estimate for water use and the 17 percent estimate for

depth applied are best understood as conservative midpoints between two credible counterfactual stories—one that attributes the residual pre-trend entirely to selection, and one that treats it as a trajectory the program interrupts—rather than as upper bounds on the true program effect. The data also speak to quantities but not to mechanisms: I observe water use, irrigated acreage, and crop mix, but neither yields nor farm profits, so whether the reduced extraction reflects eliminated waste or foregone production cannot be settled here.

Data and Code Availability

Replication data and code for all estimates, figures, and tables in this manuscript are available online at [REPOSITORY PLACEHOLDER]. The archive contains the cleaned panel of water-right-group $\times$ year observations, the R and Stata scripts that build the panel from the underlying Kansas Water Information Management and Analysis System (WIMAS) records and the Kansas Department of Agriculture (KDA) Multi-Year Flex Account administrative key, the analysis scripts that produce every numerical result in the body and the appendices (including the dynamic difference-in-differences fits, the Callaway–Sant’Anna estimates, and the HonestDiD and Bilinski–Hatfield sensitivity analyses in Appendix D), and a README documenting the full reproduction pipeline. The raw WIMAS withdrawal records are publicly available through the KDA Division of Water Resources; the MYFA term-permit start and end dates are obtained from the same office under a standard public-records request and are redistributed in the replication archive in deidentified water-right-group form.

References

- Arellano-Gonzalez, J. and Moore, F. C. (2020). Intertemporal arbitrage of water and long-term agricultural investments: Drought, groundwater banking, and perennial cropping decisions in California. *American Journal of Agricultural Economics*, 102(5):1368–1382.
- Ayres, A. B., Meng, K. C., and Plantinga, A. J. (2021). Do environmental markets improve on open access? Evidence from California groundwater rights. *Journal of Political Economy*, 129(10):2817–2860.
- Bilinski, A. and Hatfield, L. A. (2024). Nothing to see here? A non-inferiority approach to parallel trends. *Statistics in Medicine*.
- Brewer, J., Glennon, R., Ker, A., and Libecap, G. (2008). Water markets in the West: Prices, trading, and contractual forms. *Economic Inquiry*, 46(2):91–112.
- Brooks, R. D. and Harris, E. M. (2008). Efficiency gains from water markets: Empirical analysis of Watermove in Australia. *Agricultural Water Management*, 95(4):391–399.
- Callaway, B. and Sant’Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.
- Cameron-Harp, M. V. and Hendricks, N. P. (2025). Efficiency and water use: Dynamic effects of irrigation technology adoption. *Journal of the Association of Environmental and Resource Economists*, 12(2):285–314.
- Colby, B. G. (1990). Transactions costs and efficiency in western water allocation. *American Journal of Agricultural Economics*, 72(5):1184–1192.
- de Chaisemartin, C. and D’Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.
- de Chaisemartin, C. and D’Haultfœuille, X. (2024). Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*. Forthcoming.
- Deines, J. M., Kendall, A. D., Butler, J. J., and Hyndman, D. W. (2019). Quantifying irrigation adaptation strategies in response to stakeholder-driven groundwater management in the US High Plains Aquifer. *Environmental Research Letters*, 14(4):044014.
- Drysdale, K. M. and Hendricks, N. P. (2018). Adaptation to an irrigation water restriction imposed through local governance. *Journal of Environmental Economics and Management*, 91:150–165.
- Golden, B. and Leatherman, J. (2015). Impact analysis of the Sheridan 6 Local Enhanced Management Area. Technical report, Kansas State University Agricultural Experiment Station and Cooperative Extension Service.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

- Goulder, L. H. and Parry, I. W. H. (2008). Instrument choice in environmental policy. *Review of Environmental Economics and Policy*, 2(2):152–174.
- Holland, S. P., Hughes, J. E., and Knittel, C. R. (2009). Greenhouse gas reductions under low carbon fuel standards? *American Economic Journal: Economic Policy*, 1(1):106–146.
- Howitt, R. E. (1994). Empirical analysis of water market institutions: The 1991 California water market. *Resource and Energy Economics*, 16(4):357–371.
- Kansas Geological Survey (2024). High Plains aquifer water level monitoring. Technical report, University of Kansas, Lawrence, KS.
- Kuwayama, Y. and Brozović, N. (2019). The regulation of a spatially heterogeneous externality: Tradable groundwater permits to protect streams. *Journal of Environmental Economics and Management*, 67(2):115–130.
- McCann, L. and Easter, K. W. (2004). A framework for estimating the transaction costs of alternative mechanisms for water exchange and allocation. *Water Resources Research*, 40(9):W09S09.
- Orduña Alegría, M. E., Zipper, S., Shin, H. C., Deines, J. M., Hendricks, N. P., Allen, J. J., Bohling, G. C., Golden, B., Griggs, B. W., Lauer, S., Lin, C.-Y., Marston, L. T., Sanderson, M. R., Smith, S. M., Whittemore, D. O., Wilson, B. B., Yu, D. J., Yu, Q. C., and Butler, J. J. (2024). Unlocking aquifer sustainability through irrigator-driven groundwater conservation. *Nature Sustainability*, 7(12):1574–1583.
- Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.
- Sant’Anna, P. H. C. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.
- Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.
- Wheeler, S., Loch, A., Zuo, A., and Bjornlund, H. (2014). Reviewing the adoption and impact of water markets in the Murray–Darling Basin, Australia. *Journal of Hydrology*, 518:28–41.
- Womble, P. and Hanemann, W. M. (2020). Water markets, water courts, and transaction costs in Colorado. *Water Resources Research*, 56(4):e2019WR025507.
- Young, R., Foster, T., Mieno, T., Valocchi, A., and Brozović, N. (2021). Hydrologic-Economic Trade-offs in Groundwater Allocation Policy Design. *Water Resources Research*, 57(1):e2020WR027941.

A Callaway and Sant’Anna Estimates

This appendix reports the Callaway and Sant’Anna (2021) group-time estimates summarized in Section 6.1 and Section 6.2. The event study below serves as an independent check on the preferred de Chaisemartin and D’Haultfœuille (2024) specification; the tables that follow report the event-time, cohort, and calendar-year aggregations under both comparison-group variants.

Figure 7 presents the Callaway and Sant’Anna (2021) event-study estimates of MYFA effects on annual groundwater withdrawals as an independent check on the preferred specification. The horizontal axis measures years relative to MYFA enrollment (event time), with zero denoting the first year of enrollment. The vertical axis measures the estimated effect on log water use.

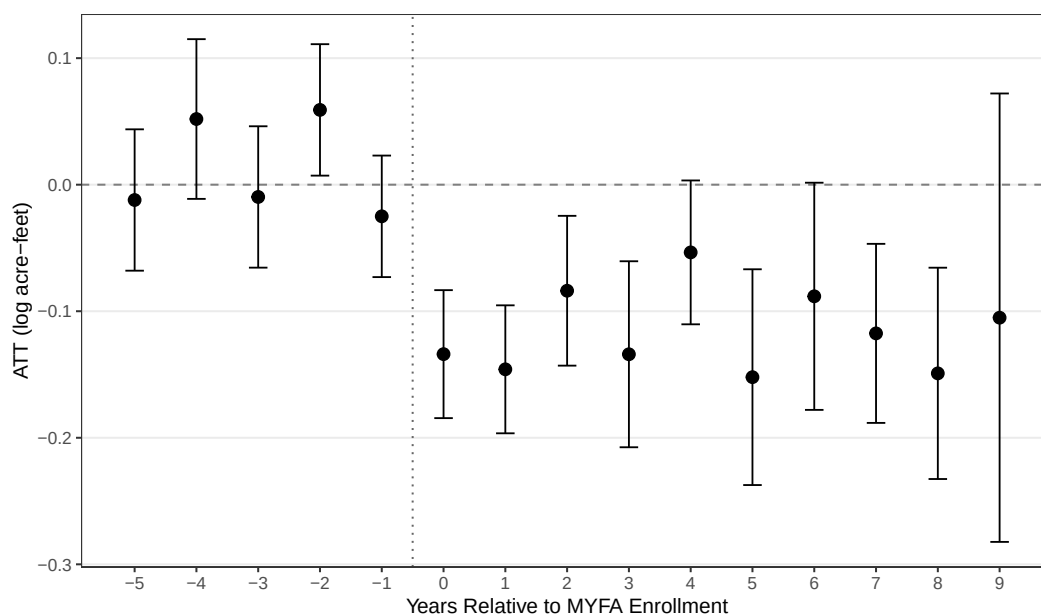


Figure 7: Event Study: Effect of MYFA Enrollment on Groundwater Withdrawals
Notes: Figure displays Callaway and Sant’Anna (2021) event study estimates of MYFA enrollment on log annual water use. Event time 0 is the first year of MYFA enrollment. Vertical bars show 95% confidence intervals based on multiplier bootstrap with 1,000 replications, clustered at the water right group level. The vertical dotted line separates pre-treatment and post-treatment periods. Estimation uses the doubly-robust method with not-yet-treated comparison group (including never-treated) and controls for growing-season precipitation and hydraulic conductivity. Sample includes water right groups in GMDs, 2001–2021.

None of the pre-treatment coefficients in the Callaway and Sant’Anna (2021) event study is statistically distinguishable from zero. Taken at face value, this is consistent with parallel trends. However, because the group-time estimator aggregates a smaller number of 2×2 comparisons, its standard errors on the pre-treatment coefficients are roughly twice the size of the corresponding de Chaisemartin and D’Haultfœuille (2024) placebos reported in Table 4 (typical SEs of 0.025–0.032 versus 0.013–0.022). Insignificance of the pre-treatment coefficients therefore reflects low power rather than strong evidence in favor of parallel trends; the preferred specification’s more powerful placebo tests do detect modest violations, as noted above.

The post-treatment estimates reveal substantial reductions in groundwater extraction following MYFA enrollment. In the first year of enrollment (event time 0), log water use declines by 0.134 (s.e. = 0.026, $p < 0.001$), implying an approximate 12.5 percent reduction at the point of enrollment. The event-study aggregated ATT is -0.116 log units (s.e. = 0.022, $p < 0.001$). The point-estimate pattern at event times 3–7 is less stable than in the preferred specification, which is expected given the group-time estimator’s absorbing-state coding: for the dominant 2012 cohort, event times 5+ contain a mix of late-flex-period and post-expiration observations coded as treated, and this misclassification injects year-to-year noise that the dynamic estimator avoids.

Table 6 reports selected event-study coefficients and the aggregate event-study estimate under two variants of the Callaway and Sant’Anna (2021) estimator. In the *baseline* variant, never-treated water right groups are included in the not-yet-treated control pool. In the *not-yet-treated-only* variant, never-treated units are excluded, isolating comparisons among eventual MYFA participants. The latter yields uniformly larger estimated reductions (discussed in Section 6.3) but shares the same absorbing-state limitation as the baseline and so neither variant displaces the preferred specification reported in Table 4.

Table 6: Callaway–Sant’Anna Event-Time Estimates of MYFA Enrollment on Log Groundwater Withdrawals

	Baseline		Not-Yet-Treated Only	
	Estimate	Std. Error	Estimate	Std. Error
<i>Aggregate</i>				
Event study (dynamic)	−0.116*	0.022	−0.242*	0.034
<i>Selected Event-Time Estimates</i>				
$e = -5$	−0.012	0.029	−0.044	0.039
$e = -3$	−0.010	0.029	0.008	0.038
$e = -1$	−0.025	0.025	0.003	0.035
$e = 0$ (enrollment year)	−0.134*	0.026	−0.167*	0.039
$e = 1$	−0.146*	0.026	−0.187*	0.042
$e = 2$	−0.084*	0.030	−0.203*	0.048
$e = 3$	−0.134*	0.038	−0.303*	0.050
$e = 4$	−0.054	0.029	−0.186*	0.047
$e = 5$	−0.152*	0.044	−0.283*	0.054

Table 7: Callaway–Sant’Anna Cohort and Calendar-Year Estimates of MYFA Enrollment on Log Groundwater Withdrawals

	Baseline		Not-Yet-Treated Only	
	Estimate	Std. Error	Estimate	Std. Error
<i>Aggregate</i>				
Cohort (group)	−0.119*	0.020	−0.228*	0.032
Calendar year	−0.113*	0.018	−0.217*	0.029
<i>Selected Cohort Estimates</i>				
2013 cohort	−0.121*	0.023	−0.254*	0.035
2018 cohort	−0.284*	0.105	−0.309	0.124
2020 cohort	−0.243	0.105	−0.145	0.123
<i>Selected Calendar Year Estimates</i>				
2013	−0.137*	0.026	−0.171*	0.047
2016	−0.141*	0.042	−0.326*	0.051
2019	−0.082	0.044	−0.318*	0.052

B Effects on Irrigation Intensity

This appendix reports the estimates for irrigation intensity summarized in Section 6.3. Depth applied—acre-feet of water used divided by irrigated acres—measures how much water is applied per unit of irrigated land, and so distinguishes a reduction achieved by irrigating fewer acres from one achieved by applying less water to the same acres. The second part of the appendix repeats on this outcome the sensitivity analysis that Appendix D develops for total water use.

B.1 Depth-Applied Estimates

The reduction in total water use documented above could reflect either reduced irrigated acreage or improved irrigation efficiency (less water applied per acre). To distinguish between these mechanisms, I examine the effect of MYFA enrollment on depth applied, defined as acre-feet of water used divided by irrigated acres. This measure captures irrigation intensity independent of the scale of irrigated operations. To match the preferred specification used for total water use in Section 6.1, I estimate the effect on depth applied with the de Chaisemartin and D’Haultfœuille (2024) dynamic estimator under within-GMD switcher–stayer comparisons and growing-season precipitation as a time-varying control.

Figure 8 reports the dynamic event-study estimates. Effects are negative in every estimated post-switch period and follow a pattern qualitatively similar to the total water-use estimates: a small immediate reduction at Effect 1 (-0.047 acre-feet per acre, $\text{s.e.} = 0.017$), a deepening to -0.16 to -0.19 acre-feet per acre over Effects 2–5 (within the dominant cohort’s flex period), and persistent reductions of -0.09 to -0.16 acre-feet per acre over Effects 6–10 (which include post-expiration years for early cohorts). The average treatment effect across all estimated Effects is -0.192 acre-feet per acre ($\text{s.e.} = 0.020$, 95% CI $[-0.230, -0.154]$), corresponding to a 16.7 percent reduction relative to the pre-treatment mean of 1.148 acre-feet per acre.

The Placebo estimates reveal pre-trend violations of the same character as those in the total water-use specification. All five within-GMD placebos are negative (range -0.067 to -0.118),

individually significant, and jointly reject the null of zero pre-trend ($p < 0.001$). The magnitudes of these violations are large enough that the residual pre-trend warrants the same cautious interpretation here as for total water use: a portion of the post-enrollment reduction may reflect a pre-existing downward trajectory among eventual MYFA participants rather than the program itself. Appendix B.2 repeats the HonestDiD and Bilinski–Hatfield sensitivity machinery on depth applied. The Bilinski–Hatfield V1 and V4 residualizations move the ATT to -0.317 and -0.327 ac-ft per acre respectively (roughly 28 percent of the pre-treatment mean, versus 17 percent at baseline), and the HonestDiD relative-magnitudes breakdown sits in $\bar{M} \in (0, 0.5)$ at baseline and improves to $\bar{M} \in (0.5, 1.0]$ after residualization — mirroring the log water-use pattern. The qualitative conclusion is the same as for total water use, with one notable asymmetry: depth applied is *more* robust under the HonestDiD smoothness restriction (because the post-period effects are uniformly large and negative across all ten estimated event-time periods) but similarly fragile under the relative-magnitudes restriction (because the pre-period violations are themselves large). The qualitative pattern of the post-switch effects—an immediate response that deepens through the early flex years and persists into the post-expiration window—is, however, similar to the total water-use case and is consistent with the option-value mechanism formalized in Section 3.

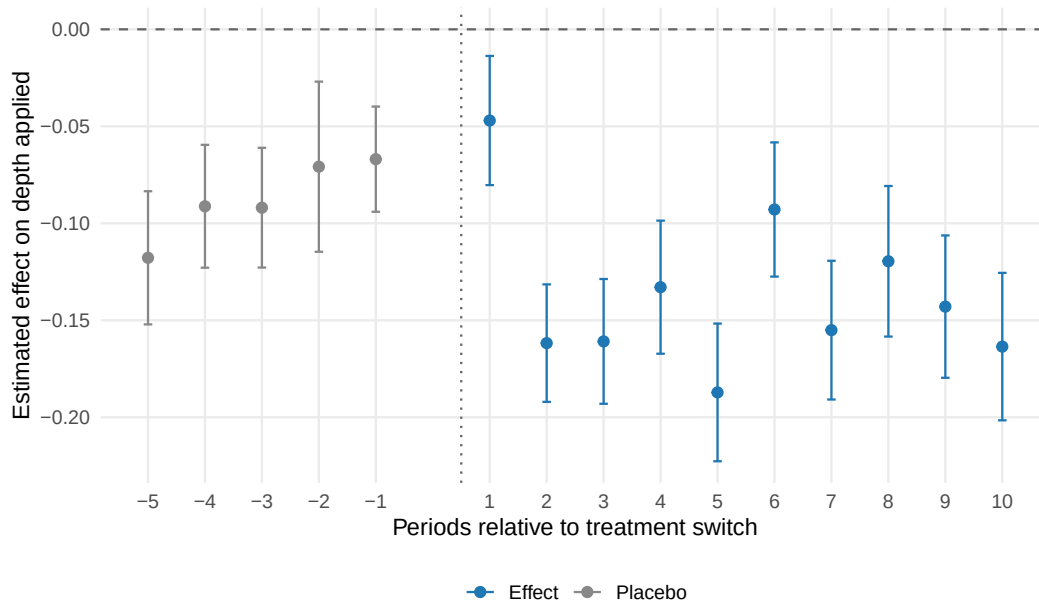


Figure 8: Dynamic DiD Event Study: Effect of MYFA Enrollment on Depth Applied
Notes: Figure displays de Chaisemartin and D’Haultfœuille (2024) dynamic DiD estimates of the effect of MYFA enrollment on depth applied (acre-feet per irrigated acre). Points at negative event times are Placebo estimates; points at positive event times are Effects. Vertical bars are 95% confidence intervals. Comparisons are restricted to switchers and stayers within the same Groundwater Management District (`trends_nonparam = "gmd"`). Time-varying control is May–September PRISM precipitation at the water right group centroid. Standard errors are clustered at the water right group level. Sample: water right groups in GMDs, 2001–2021, excluding LEMAs, deduplicated panel.

Table 8 reports the dynamic event-study estimates underlying Figure 8 alongside the average ATT and the placebo coefficients. As context, the pre-treatment mean of depth applied is 1.148 acre-feet per acre, so the average ATT of -0.192 corresponds to a 16.7 percent reduction.

Table 8: Dynamic DiD Estimates of MYFA Effect on Depth Applied

	Estimate	Std. Error
<i>Post-Switch Effects (within flex period: Effects 1–5)</i>		
Effect 1	−0.047**	0.017
Effect 2	−0.162***	0.015
Effect 3	−0.161***	0.016
Effect 4	−0.133***	0.018
Effect 5	−0.187***	0.018
<i>Post-Switch Effects (post-expiration: Effects 6–10)</i>		
Effect 6	−0.093***	0.018
Effect 7	−0.155***	0.018
Effect 8	−0.120***	0.020
Effect 9	−0.143***	0.019
Effect 10	−0.164***	0.019
Average ATT across Effects 1–10	−0.192***	0.020
<i>Placebo Estimates (pre-switch)</i>		
Placebo 1	−0.067***	0.014
Placebo 2	−0.071**	0.022
Placebo 3	−0.092***	0.016
Placebo 4	−0.091***	0.016
Placebo 5	−0.118***	0.018
Joint placebo test (<i>p</i> -value)	<0.001	
<i>Context</i>		
Pre-treatment mean	1.148 acre-feet per acre	
Average ATT as % of mean	−16.7%	

Notes: de Chaisemartin and D’Haultfœuille (2024) dynamic DiD estimates using the `DIDmultiplegtDYN` R package. Outcome is depth applied (acre-feet used divided by irrigated acres). Treatment is a time-varying indicator equal to one in years with an active MYFA and zero otherwise (including post-expiration years). Comparisons are restricted to switchers and stayers within the same Groundwater Management District (`trends_nonparam = "gmd"`). Time-varying control is May–September mean precipitation from PRISM at the water right group centroid. Standard errors clustered at the water right group level. *, **, *** denote $p < 0.05$, $p < 0.01$, $p < 0.001$. Sample: water right groups in GMDs, 2001–2021, excluding LEMAs, deduplicated panel.

These results indicate that the reduction in total water use under MYFAs reflects improved irrigation efficiency rather than simply reduced irrigated acreage. Enrollees apply substantially

less water per acre while maintaining their irrigated footprint. This pattern is consistent with the hypothesis that MYFAs encourage more careful water management: freed from the pressure to use annual allocations or lose them, irrigators can apply water more precisely according to actual crop needs rather than hedging against allocation forfeiture.

B.2 Parallel Sensitivity Analysis for Depth Applied

The depth-applied analysis in Section B.1 exhibits a pre-trend pattern of the same character as the log water-use specification: jointly significant placebos (joint $p < 10^{-9}$, individual placebos ranging from -0.067 to -0.118 ac-ft per acre) that are a non-trivial fraction of the average post-treatment effect (-0.192 ac-ft per acre, ≈ 17 percent of the pre-treatment mean). To assess the implications of this violation, I repeat the HonestDiD and Bilinski–Hatfield analyses from Sections D.1 and D.2 on the depth-applied outcome. The estimation machinery is identical (dCdH preferred specification, within-GMD trends, growing-season precipitation as time-varying control, cluster-bootstrapped VCV with $B = 40$ subprocess workers); only the outcome variable changes from $\ln(\text{af})$ to depth applied (af/acres, ac-ft per acre).

HonestDiD bounds on depth applied

Figures 9 and 10 report the 95% confidence intervals on the average post-treatment ATT under the smoothness (Δ^{SD}) and relative-magnitudes (Δ^{RM}) restrictions. The two restrictions deliver complementary findings about the robustness of the depth-applied result. The fixed-length confidence interval (FLCI) implementation of the smoothness restriction at $M = 0$ just barely includes zero (upper bound $+0.005$ ac-ft per acre), so by the FLCI’s honest-inference criterion the depth-applied result is borderline even before any smoothness allowance. The FLCI at $M \geq 0.01$ excludes zero (upper bounds in the -0.024 to -0.054 range) and continues to exclude zero across the full grid examined ($M = 0$ to 0.05 in steps of 0.01). The relative-magnitudes restriction is comparably stringent: the breakdown $\bar{M}$ lies between $\bar{M} = 0$ (where the C-LF interval is $[-0.081, -0.013]$) and $\bar{M} = 0.5$ (where the interval is $[-0.114, +0.013]$ and includes zero), well below the conventional

$\bar{M} = 1$ benchmark. As for total water use, the depth-applied result is fragile under either restriction at the conventional benchmark and only marginally robust at $\bar{M} \rightarrow 0$.

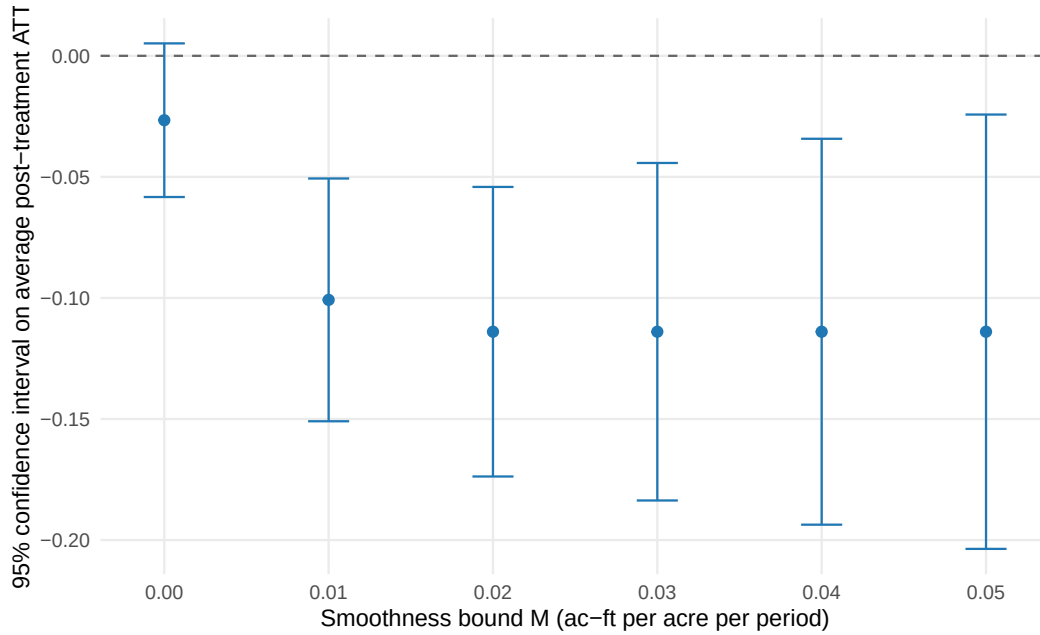


Figure 9: HonestDiD Smoothness Restriction $\Delta^{SD}(M)$ for Depth Applied

Notes: 95% confidence intervals on the average post-treatment ATT for depth applied (ac-ft per acre) in the dCdH preferred specification under the smoothness restriction $\Delta^{SD}(M)$. Horizontal axis is the bound M on the period-to-period change in the trend deviation. Confidence intervals computed by the fixed-length method of Rambachan and Roth (2023), with the dCdH coefficient covariance matrix estimated by 40 cluster-bootstraps on water right groups. Number of post-treatment periods = 10; number of pre-treatment placebo periods = 5.

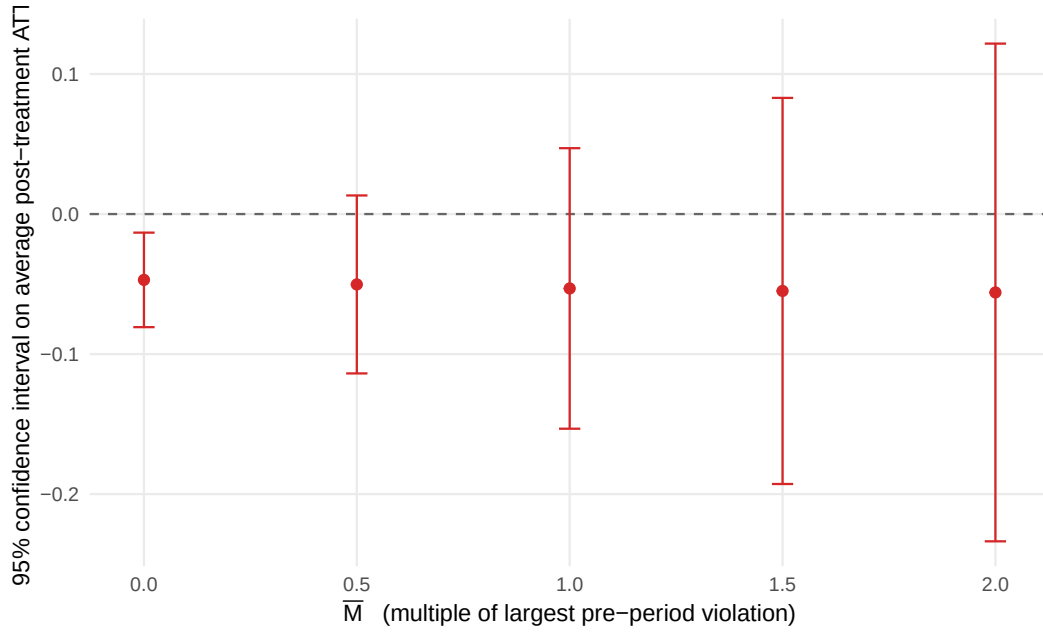


Figure 10: HonestDiD Relative-Magnitudes Restriction $\Delta^{RM}(\bar{M})$ for Depth Applied
Notes: 95% confidence intervals on the average post-treatment ATT for depth applied (ac-ft per acre) under the relative-magnitudes restriction $\Delta^{RM}(\bar{M})$. Horizontal axis is the bound $\bar{M}$ as a multiple of the largest observed pre-period violation. Confidence intervals computed by the conditional least-favorable method of Rambachan and Roth (2023).

Table 9: HonestDiD Sensitivity Bounds on the Average Post-Treatment ATT for Depth Applied

Restriction	95% CI lower bound	95% CI upper bound
$\Delta^{SD}(M)$ — <i>smoothness</i>		
$M = 0$	−0.058	+0.005
$M = 0.01$	−0.151	−0.051
$M = 0.02$	−0.174	−0.054
$M = 0.03$	−0.184	−0.044
$M = 0.04$	−0.194	−0.034
$M = 0.05$	−0.204	−0.024
$\Delta^{RM}(\bar{M})$ — <i>relative magnitudes</i>		
$\bar{M} = 0$	−0.081	−0.013
$\bar{M} = 0.5$	−0.114	+0.013
$\bar{M} = 1.0$	−0.153	+0.047
$\bar{M} = 1.5$	−0.193	+0.083
$\bar{M} = 2.0$	−0.234	+0.122

Notes: HonestDiD bounds on the dCdH preferred-specification average post-treatment ATT for depth applied (ac-ft per acre), under the smoothness (Δ^{SD}) and relative-magnitudes (Δ^{RM}) restrictions of Rambachan and Roth (2023). Coefficient covariance matrix estimated by 40 cluster-bootstraps on water right groups. Under the smoothness restriction, the FLCI at $M = 0$ (no smoothness allowance) just barely includes zero; the FLCI at $M \in \{0.01, \dots, 0.05\}$ excludes zero and is monotonically wider in absolute terms as M rises. The non-monotonicity of the upper bound between $M = 0$ and $M = 0.01$ reflects the discontinuity between the standard CI formula at $M = 0$ and the optimization-based FLCI formula at $M > 0$ when off-diagonal correlations in the bootstrap VCV are positive. Under the relative-magnitudes restriction, the breakdown is $\bar{M} \in (0, 0.5)$.

Bilinski–Hatfield residualization on depth applied

I implement the same two Bilinski–Hatfield variants from Section D.2 (V1: per-water right group linear pre-trend, fit on each water right group’s pre-MYFA observations; V4: switcher-vs-stayer linear differential, the original Bilinski–Hatfield “one step up” formulation), applied to depth applied rather than $\ln(\text{af})$. Figure 11 compares the event-study coefficients across the baseline, V1, and V4 specifications.

Under V1, the joint placebo p -value rises from 3.8×10^{-10} to 0.0083, the average post-

treatment ATT is -0.317 ac-ft per acre (s.e. 0.027), and the HonestDiD Δ^{RM} breakdown shifts to $\bar{M} \in (0.5, 1.0]$. Under V4, the joint placebo p -value is 0.0155 , the ATT is -0.327 ac-ft per acre (s.e. 0.020), and the Δ^{RM} breakdown is also in $\bar{M} \in (0.5, 1.0]$. As context, the pre-treatment mean of depth applied is 1.148 ac-ft per acre, so the V1 residualized ATT corresponds to a 27.6 percent reduction and V4 to a 28.5 percent reduction — both substantially larger than the baseline depth-applied ATT of -0.192 ac-ft per acre (≈ 17 percent), and parallel in direction and relative magnitude to the BH residualization range for log water use (-0.23 to -0.27 log points, or roughly 20 – 24 percent). One important asymmetry relative to the log water-use BH results is that neither V1 nor V4 passes the joint placebo equivalence test at conventional significance levels for depth applied — the residual within-GMD pre-period coefficients are themselves so large in absolute terms (-0.067 to -0.118 ac-ft per acre) that even a per-water right group linear residualization cannot drive them all the way to zero. The BH residualized estimates for depth applied therefore stand as informative alternative point estimates that bracket the program effect, rather than as a clean parallel-trends-satisfied counterfactual. For completeness, the unit-specific quadratic variant (V2) and the GMD-aggregate linear variant (V3) display the same instability pattern as for log water use — V2 over-fits per-water right group curvature into the post-period and yields an implausibly large ATT of -2.35 ac-ft per acre (more than twice the pre-treatment mean); V3 is too coarse to differentiate switchers from stayers within a GMD and reproduces the baseline ATT to four decimal places. As before, neither variant is reported as a credible alternative.

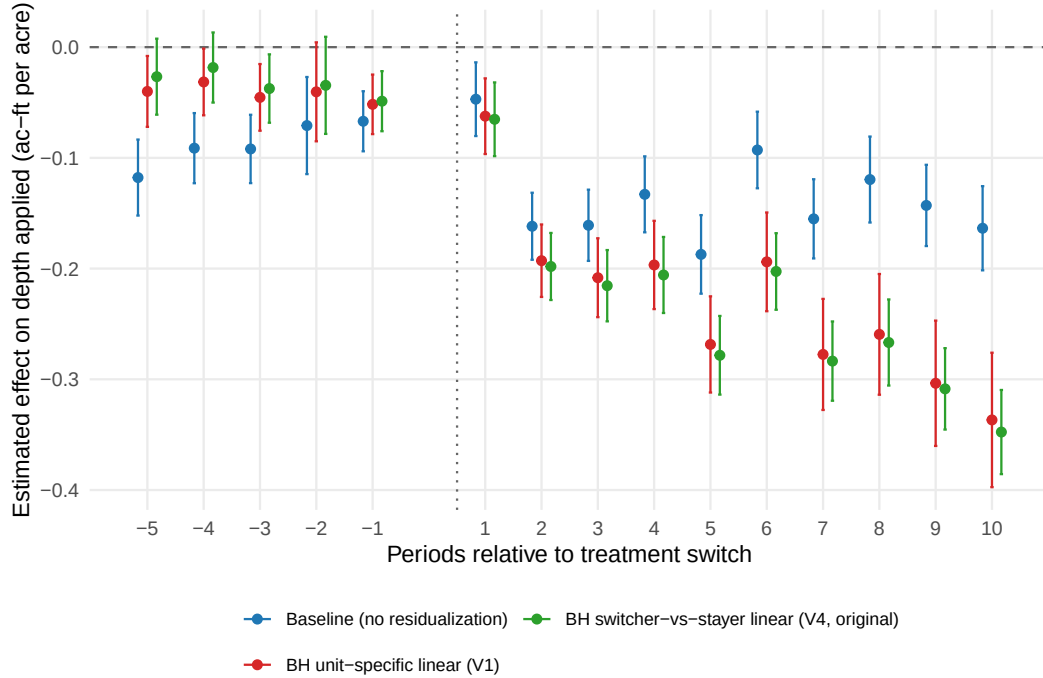


Figure 11: Event-Study Coefficients for Depth Applied: Baseline vs. Bilinski-Hatfield Residualized

Notes: dCdH event-study coefficients on depth applied (ac-ft per acre) under three specifications.

Blue: baseline (no residualization), reproducing Figure 8. Red: Bilinski-Hatfield V1, residualizing depth applied by a unit-specific linear pre-trend fit on each water right group’s pre-treatment observations. Green: V4, residualizing by a single switcher-vs-stayer linear trend differential (the original Bilinski-Hatfield “one step up” formulation). Vertical bars are 95% confidence intervals.

C Effects on Crop Choice

This appendix examines whether MYFA enrollment affects crop selection decisions. Irrigators facing different water availability constraints may adjust their crop mix, potentially shifting toward less water-intensive crops or crops with different yield-water response functions. I estimate the effect of MYFA enrollment on acres planted to five major irrigated crops in western Kansas: corn, soybeans, wheat, grain sorghum, and alfalfa.

Table 10 reports the overall average treatment effects for each crop. Among the five crops examined, only wheat shows a statistically significant effect: MYFA enrollment increases wheat

acreage by approximately 10.2 acres per water right group (s.e. = 4.9, $p < 0.05$). The point estimates for sorghum (+6.3 acres) and alfalfa (+5.2 acres) are positive but not statistically distinguishable from zero. Corn (−4.2 acres) and soybeans (−2.1 acres) show small negative point estimates, also not statistically significant.

Table 10: Effect of MYFA Enrollment on Crop Acreage

Crop	ATT (acres)	Std. Error	95% CI
Wheat	10.20	4.89	[0.61, 19.80]*
Sorghum	6.34	4.45	[−2.37, 15.05]
Alfalfa	5.16	6.57	[−7.72, 18.04]
Soybeans	−2.07	2.35	[−6.68, 2.54]
Corn	−4.24	8.56	[−21.02, 12.54]

Notes: Callaway and Sant’Anna (2021) difference-in-differences estimates. Each row reports the overall simple average treatment effect on acres planted to the indicated crop. Estimation uses doubly-robust method with not-yet-treated comparison group and controls for growing-season precipitation and hydraulic conductivity. Standard errors from multiplier bootstrap with 1,000 replications, clustered by water right group. * indicates 95% confidence interval excludes zero. Sample excludes water right groups in LEMAs.

Figure 12 displays the point estimates and confidence intervals for all five crops. The pattern of results—increased wheat and possibly sorghum, decreased or unchanged corn and soybeans—is consistent with a modest shift toward less water-intensive crops, though the imprecision of most estimates limits strong conclusions.

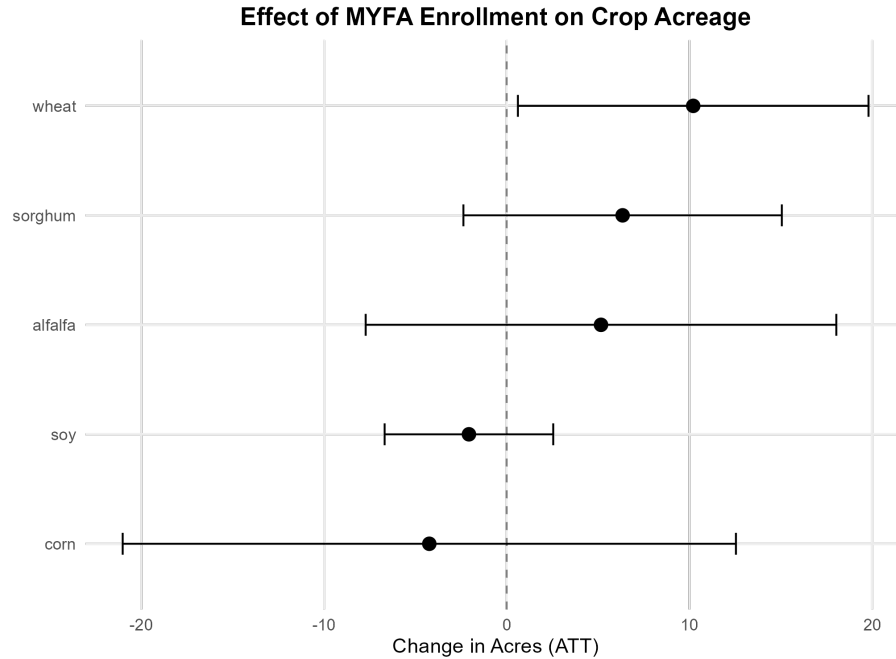


Figure 12: Effect of MYFA Enrollment on Crop Acreage: Summary

Notes: Figure displays overall average treatment effects of MYFA enrollment on acres planted to each crop. Points show ATT estimates; horizontal bars show 95% confidence intervals. Estimation uses Callaway and Sant’Anna (2021) doubly-robust method with not-yet-treated comparison group.

Figures 13–17 present event study estimates for each crop. The pre-treatment coefficients generally support the parallel trends assumption, though some crops (particularly alfalfa) show more variability in the pre-period. The post-treatment dynamics vary across crops: wheat shows consistently positive effects emerging after enrollment, while corn shows a pattern of small negative effects that do not reach statistical significance.

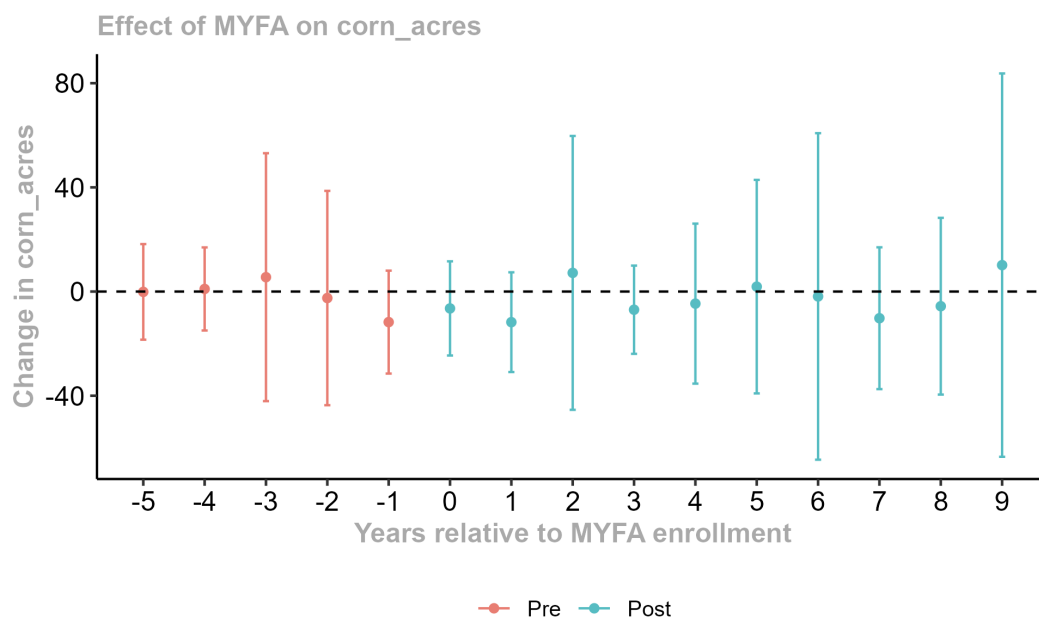


Figure 13: Event Study: Effect of MYFA Enrollment on Corn Acreage

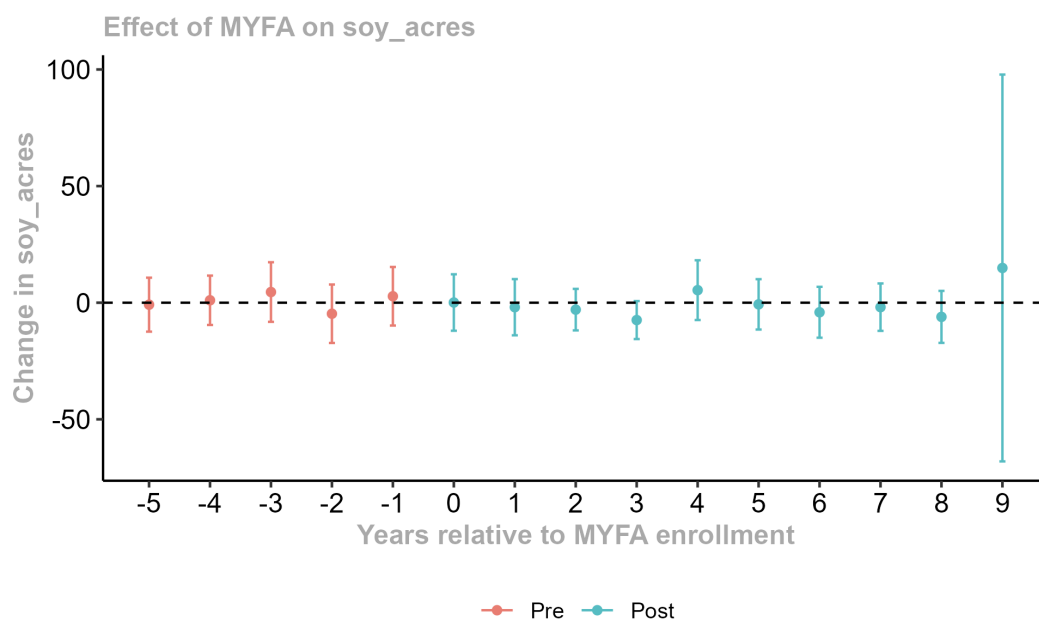


Figure 14: Event Study: Effect of MYFA Enrollment on Soybean Acreage

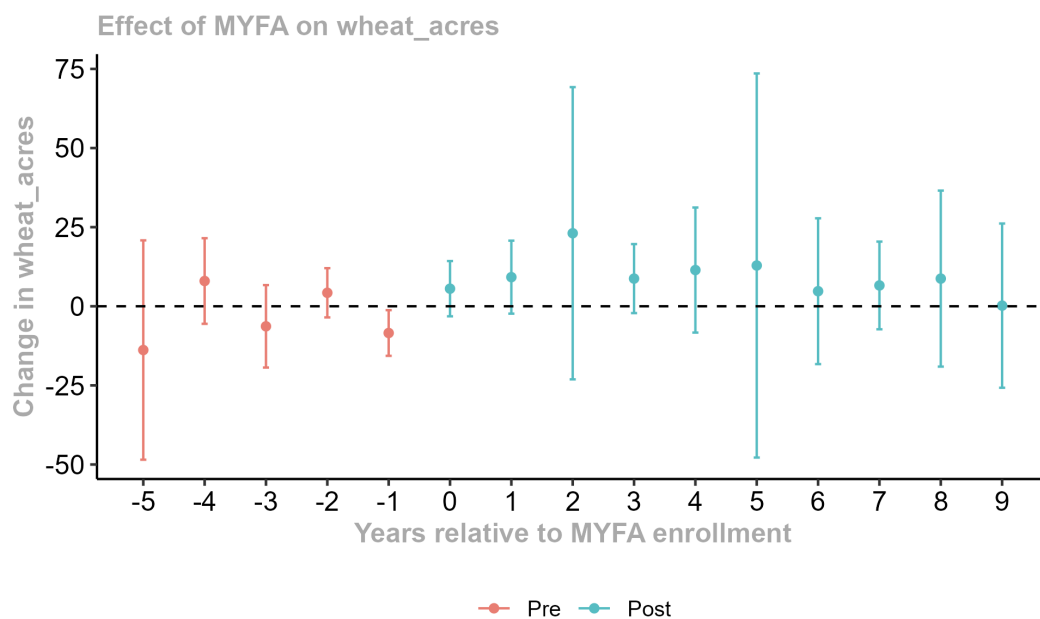


Figure 15: Event Study: Effect of MYFA Enrollment on Wheat Acreage

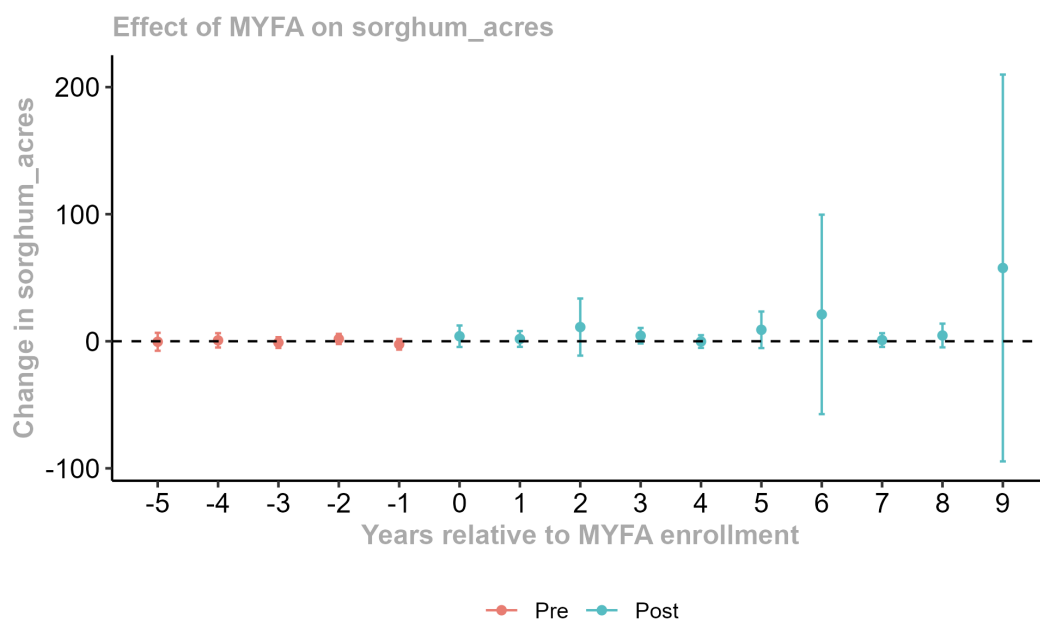


Figure 16: Event Study: Effect of MYFA Enrollment on Sorghum Acreage

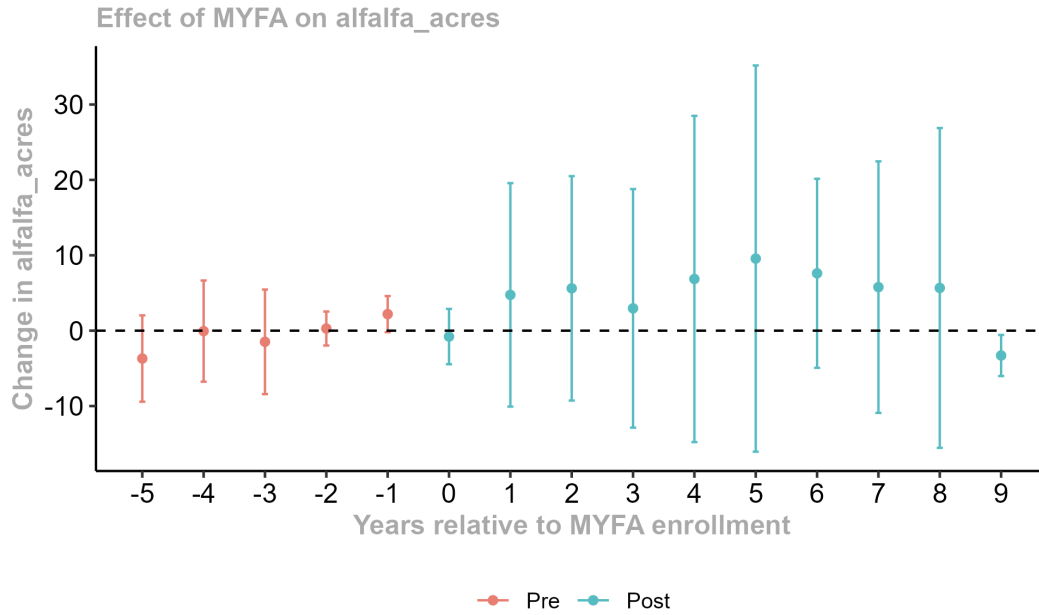


Figure 17: Event Study: Effect of MYFA Enrollment on Alfalfa Acreage

The statistically significant increase in wheat acreage is noteworthy because winter wheat is substantially less water-intensive than corn, the dominant irrigated crop in western Kansas. Wheat’s growing season largely precedes the peak summer irrigation period, and much of the crop’s water requirement can be met by winter and spring precipitation. If MYFA enrollees are shifting acres from corn to wheat, this would represent an additional margin of water conservation beyond the reduced application intensity documented in the main text.

However, interpretation requires caution. The crop choice results are less precisely estimated than the water use results, likely because crop decisions respond to many factors beyond water policy—including commodity prices, crop insurance provisions, and agronomic rotations. The absence of significant effects on corn acreage, despite corn being the most water-intensive major crop, suggests that MYFA-induced conservation operates primarily through the intensive margin (less water per acre) rather than the extensive margin (different crops).

D Sensitivity Analysis for the Pre-Trend Violation

This appendix bounds the implications of the residual pre-trend violation in the preferred de Chaisemartin and D’Haultfœuille (2024) specification. Recall from Section 6.1 that the within-GMD placebos for total water use are jointly significant at $p = 0.0002$, with individual placebo magnitudes ranging from -0.011 to -0.089 log points. Because some of these magnitudes are a non-trivial fraction of the headline ATT of -0.133 , a careful reader will want to know how much of the headline could in principle reflect a pre-existing trajectory rather than the program. I address this question with two complementary tools: the Rambachan and Roth (2023) HonestDiD framework, which non-parametrically bounds the post-period bias under restrictions on the post-period trend deviation, and a Bilinski and Hatfield (2024) “one step up” residualization, which nets out a unit-specific linear pre-trend before re-estimating the dCdH model. Sections D.1 and D.2 apply both tools to the log water-use outcome; Section B.2 repeats the same analyses for depth applied, the secondary outcome analyzed in Appendix B. Section D.4 synthesizes across both outcomes.

D.1 HonestDiD Sensitivity Bounds

The Rambachan and Roth (2023) framework treats the post-treatment trend deviation as partially identified under explicit restrictions linking pre- and post-period violations of parallel trends. I implement two such restrictions, each imposed on the dCdH preferred-specification event-study coefficients:

Smoothness restriction $\Delta^{SD}(M)$. Bounds the change in the trend deviation between consecutive periods at M log points. When $M = 0$ exact parallel trends holds; larger M allows for non-linear violations whose period-to-period change is bounded.

Relative-magnitudes restriction $\Delta^{RM}(\bar{M})$. Bounds the largest post-period violation as a multiple $\bar{M}$ of the largest pre-period violation actually observed. The conventional benchmark, used in the original Rambachan and Roth paper, is $\bar{M} = 1$, meaning post-period violations are restricted to

be no larger than the worst pre-period violation.

I compute the resulting fixed-length confidence intervals (Δ^{SD}) and conditional least-favorable confidence intervals (Δ^{RM}) using the `HonestDiD` R package. The covariance matrix of the dCdH coefficient vector is estimated by 40 cluster-bootstraps on water right groups, each run in a fresh R subprocess to avoid memory accumulation issues with the underlying `DIDmultiplegtDYN` backend; off-diagonal correlations between pre- and post-period coefficients of 0.2–0.6 are present but do not materially shift the breakdown values relative to a diagonal-VCV approximation.

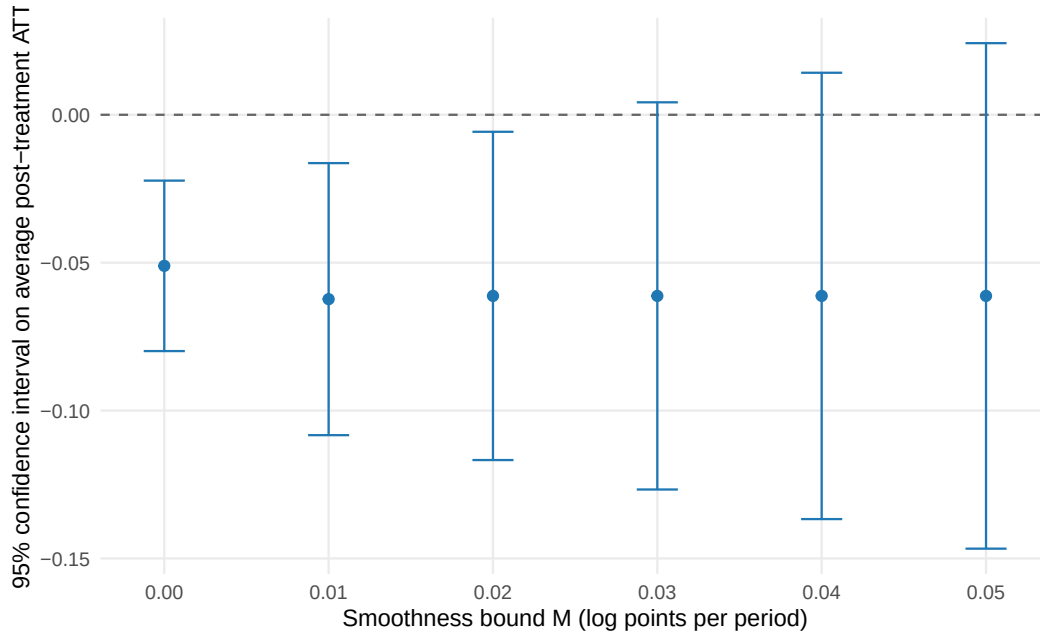


Figure 18: HonestDiD Smoothness Restriction $\Delta^{SD}(M)$

Notes: 95% confidence intervals on the average post-treatment ATT in the dCdH preferred specification under the smoothness restriction $\Delta^{SD}(M)$. The horizontal axis is the bound M on the period-to-period change in the trend deviation. The dashed line is zero. Confidence intervals computed by the fixed-length method of Rambachan and Roth (2023), with the dCdH coefficient covariance matrix estimated by 40 cluster-bootstraps on water right groups.

Figure 18 reports the 95% confidence intervals under the smoothness restriction. The interval excludes zero at $M = 0, 0.01$, and 0.02 , and includes zero at $M = 0.03$ and above. The breakdown value is therefore approximately $M \approx 0.025$ log points per period. For context, the observed pre-period coefficients change between consecutive periods by up to 0.052 log points (the step from

Placebo 5 to Placebo 4), so the post-period would only need to change by less than half what the pre-period has been observed to change for the headline to fail conventional significance.

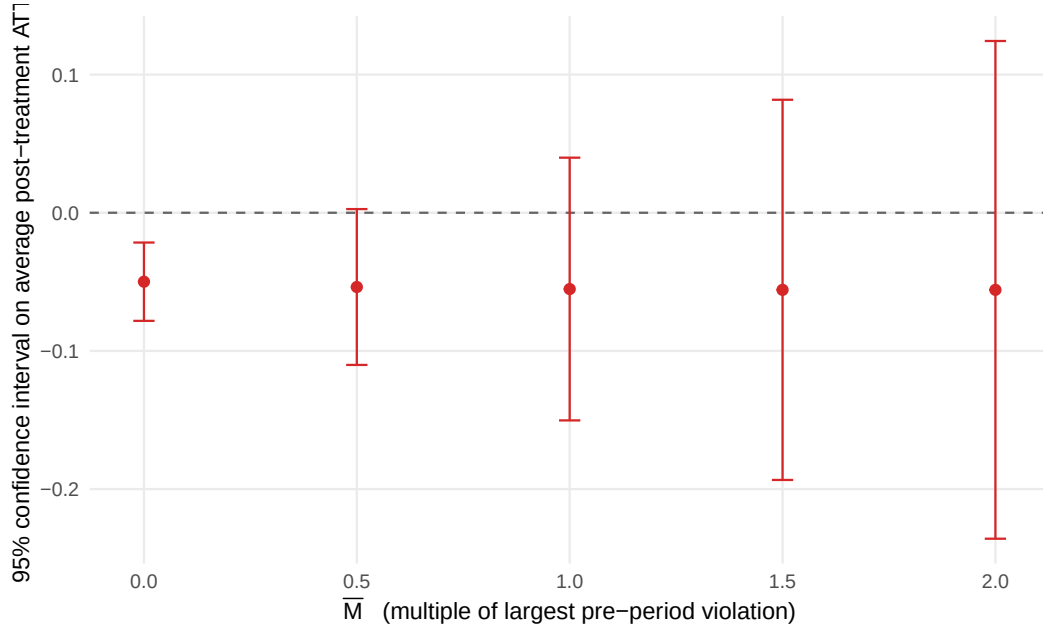


Figure 19: HonestDiD Relative-Magnitudes Restriction $\Delta^{RM}(\bar{M})$

Notes: 95% confidence intervals on the average post-treatment ATT in the dCdH preferred specification under the relative-magnitudes restriction $\Delta^{RM}(\bar{M})$. The horizontal axis is the bound $\bar{M}$ as a multiple of the largest observed pre-period violation. The dashed line is zero. Confidence intervals computed by the conditional least-favorable method of Rambachan and Roth (2023).

Figure 19 reports the relative-magnitudes restriction. The interval excludes zero at $\bar{M} = 0$ and includes zero at every $\bar{M} \geq 0.5$. The breakdown value is therefore $\bar{M} \approx 0.4\text{--}0.5$, well below the $\bar{M} = 1$ benchmark. This is the more concerning of the two restrictions because it is hard to argue *a priori* that post-period trend deviations would be *smaller* than pre-period deviations. Under the conventional benchmark, the headline result is not robust.

Table 11: HonestDiD Sensitivity Bounds on the Average Post-Treatment ATT

Restriction	95% CI lower bound	95% CI upper bound
$\Delta^{SD}(M)$ — <i>smoothness</i>		
$M = 0$	−0.080	−0.022
$M = 0.01$	−0.108	−0.016
$M = 0.02$	−0.117	−0.006
$M = 0.03$	−0.127	+0.004
$M = 0.04$	−0.137	+0.014
$M = 0.05$	−0.147	+0.024
$\Delta^{RM}(\bar{M})$ — <i>relative magnitudes</i>		
$\bar{M} = 0$	−0.078	−0.022
$\bar{M} = 0.5$	−0.110	+0.003
$\bar{M} = 1.0$	−0.150	+0.040
$\bar{M} = 1.5$	−0.193	+0.082
$\bar{M} = 2.0$	−0.236	+0.124

Notes: HonestDiD bounds on the dCdH preferred-specification average post-treatment ATT, under the smoothness (Δ^{SD}) and relative-magnitudes (Δ^{RM}) restrictions of Rambachan and Roth (2023). Coefficient covariance matrix estimated by 40 cluster-bootstraps on water right groups. The breakdown thresholds — the values at which the upper bound first crosses zero — are approximately $M \approx 0.025$ for the smoothness restriction and $\bar{M} \approx 0.4$ for the relative-magnitudes restriction.

D.2 Bilinski-Hatfield Residualization

The HonestDiD bounds answer the question “how big a post-period violation would it take to overturn the result?” but do not adjust the point estimate itself. As a complement, I implement a Bilinski and Hatfield (2024) “one step up” specification, which adjusts the outcome variable to absorb a credible parametric form for the pre-trend before re-running the dCdH preferred specification.

For each water right group i , I fit a unit-specific linear time trend in $\ln(\text{af})_{it}$ on its pre-treatment observations only (years strictly before MYFA enrollment for switchers; all years for never-treated water right groups), obtaining a fitted intercept $\hat{\beta}_{0i}$ and slope $\hat{\beta}_{1i}$. I then construct the residualized outcome $y_{it}^* = \ln(\text{af})_{it} - (\hat{\beta}_{0i} + \hat{\beta}_{1i} \cdot t)$ and re-fit the dCdH preferred specification using y_{it}^* in place

of $\ln(\text{af})_{it}$. By construction, the residualized outcome has zero linear pre-trend within each water right group, so the placebo coefficients should be small if the pre-period violation is well captured by a unit-specific linear trend. In the post-period, the residualized treatment effect represents the deviation of observed water use from the linear extrapolation of the water right group's pre-treatment trajectory.

Figure 20 plots the dCdH event-study coefficients under the baseline specification, the unit-specific linear residualization (the V1 specification just described), and a switcher-vs-stayer linear residualization (the original Bilinski-Hatfield formulation, V4, which subtracts a single switcher-vs-stayer slope differential rather than per-water right group slopes). Both residualized variants flatten the placebos toward zero. Quantitatively:

- Under the V1 residualization, the joint placebo p -value rises from 0.0002 to 0.31, the average post-treatment ATT moves to -0.229 log points (s.e. 0.021), and the HonestDiD Δ^{RM} breakdown shifts to $\bar{M} \approx 1.0$.
- Under the V4 residualization, the joint placebo p -value is 0.04, the ATT is -0.266 (s.e. 0.016), and the Δ^{RM} breakdown is $\bar{M} \approx 0.5$.

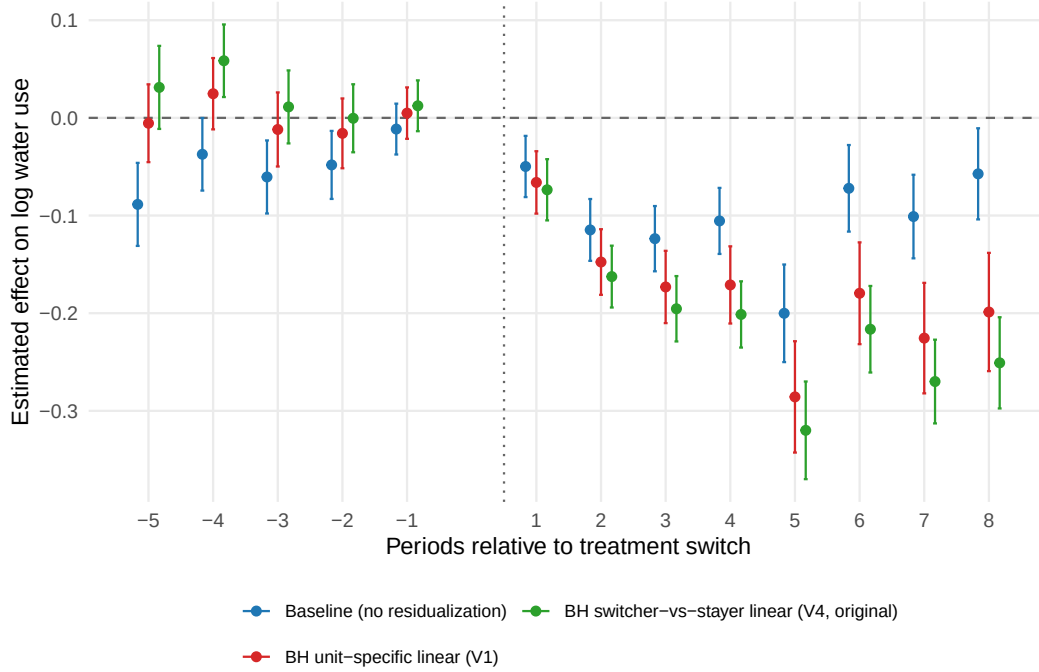


Figure 20: Event-Study Coefficients: Baseline vs. Bilinski-Hatfield Residualized
Notes: dCdH event-study coefficients on log water use under three specifications. Blue: baseline (no residualization), reproducing Figure 1. Red: Bilinski-Hatfield V1, residualizing $\ln(\text{af})$ by a unit-specific linear pre-trend fit on each water right group's pre-treatment observations. Green: V4, residualizing by a single switcher-vs-stayer linear trend differential (the original Bilinski-Hatfield “one step up” formulation). Vertical bars are 95% confidence intervals.

The two residualized variants agree on a treatment-effect range of approximately -0.23 to -0.27 log points, substantially *larger* in magnitude than the baseline -0.133 . This is consistent with the visual signature of the placebos in Figure 1: the placebo coefficients become less negative as event time approaches zero, so the linear extrapolation of the pre-trend slopes upward through the post-period. Under that linear-extrapolation counterfactual, the treatment counterfactual is rising water use; the observed post-period decline is the deviation *below* that rising counterfactual, and so the implied program effect is larger than a simple post-pre comparison would suggest.

I report the V1 specification as the primary Bilinski-Hatfield variant because it is more flexible (per-water right group slopes can absorb heterogeneity that a single group differential cannot) and because it passes the joint placebo equivalence test cleanly. V4 is reported alongside V1 to

document that the qualitative finding does not depend on the choice between unit-specific and group-specific trend forms. As robustness, I also examined two further variants: a unit-specific quadratic pre-trend (which produced an implausibly large ATT of -1.42 log points, indicating that quadratic per-unit fits over-fit the pre-period and project unstable curvature into the post-period) and a GMD-aggregate linear trend (which produced an ATT essentially identical to the baseline, indicating that the GMD-level aggregate slope is too coarse to differentiate switchers from stayers within a GMD). Neither variant is reported as a credible alternative; the plausible Bilinski-Hatfield specifications are the unit-specific and group-specific linear trends.

D.3 Leave-One-GMD-Out Robustness

The Section 4.2 disclosure documents that GMD 5 contains 50.7 percent of the treated water right groups and 52.0 percent of the $D = 1$ observations in the analysis sample. A natural concern is that the headline preferred-specification estimates are anchored heavily in a single GMD and might not survive its exclusion. Table 12 reports the dCdH preferred specification re-estimated on five panels that drop each GMD in turn, for both log water use ($n_{\text{post}} = 8$) and depth applied ($n_{\text{post}} = 10$), alongside the full-sample baselines.

Table 12: Leave-One-GMD-Out Robustness of the dCdH Preferred Specification

	Average effect	Std. Error	Joint placebo p	Worst placebo
<i>Panel A: $\ln(\text{af_used})$ outcome</i>				
Full sample	−0.133***	0.016	0.0002	−0.089
Drop GMD 1	−0.136***	0.016	0.0008	−0.081
Drop GMD 2	−0.117***	0.017	0.0019	−0.080
Drop GMD 3	−0.164***	0.017	0.0001	−0.112
Drop GMD 4	−0.139***	0.017	0.0003	−0.093
Drop GMD 5	−0.095***	0.023	0.0512	−0.072
<i>Panel B: depth applied outcome (ac-ft per acre)</i>				
Full sample	−0.192***	0.020	< 0.001	−0.118
Drop GMD 1	−0.196***	0.020	< 0.001	−0.119
Drop GMD 2	−0.187***	0.021	< 0.001	−0.109
Drop GMD 3	−0.212***	0.021	< 0.001	−0.127
Drop GMD 4	−0.208***	0.021	< 0.001	−0.126
Drop GMD 5	−0.128***	0.029	0.0051	−0.097

Notes: Leave-one-GMD-out (LOO) re-estimation of the dCdH preferred specification. Each row drops all water right groups in the indicated GMD and refits the dynamic estimator with within-GMD trends restriction and May–September precipitation as a time-varying control, on the canonical deduplicated panel. The average effect is the dCdH switcher-count-weighted average treatment effect across all estimated post-switch event times ($\ell = 1, \dots, 8$ for log water use and $\ell = 1, \dots, 10$ for depth applied). “Worst placebo” is the placebo coefficient of largest absolute magnitude among the five pre-period placebos ($\ell = -5, \dots, -1$). *** denotes $p < 0.001$. Source: pretrends_exploration/code/11_loo_gmd.R (12 dCdH fits saved as output/11_loo_gmd_*.rds).

Three findings emerge. First, the headline result is robust in *direction and significance* to dropping any single GMD: every LOO estimate is negative and individually significant for both outcomes. The most consequential drop, GMD 5, leaves the log water-use average total effect at −0.095 (≈ 9.1 percent reduction; joint placebo $p = 0.051$, just at the conventional threshold for rejecting parallel trends) and the depth-applied average total effect at −0.128 (≈ 12 percent reduction; joint placebo $p = 0.005$). The headline does not depend on any one GMD for its qualitative conclusion.

Second, the headline result’s *magnitude* is materially driven by GMD 5. Dropping GMD 5

reduces the log water-use average total effect by about 29 percent (from -0.133 to -0.095) and the depth-applied average total effect by about 33 percent (from -0.192 to -0.128). This is consistent with the Section 4.2 disclosure that GMD 5 supplies about half the identifying variation: dropping it should attenuate but not eliminate the effect, which is precisely what the LOO exercise shows. Dropping any other GMD has little effect on the magnitude or actually strengthens the result — dropping GMD 3 produces a $\ln(\text{af})$ average total effect of -0.164 (about 23 percent larger in magnitude than the full-sample estimate) and a depth-applied average total effect of -0.212 (about 10 percent larger). This pattern suggests that GMD 3 (which holds 21.4 percent of treated water right groups but only an average treatment-effect magnitude) is pulling the headline toward the lower end of the cohort distribution; excluding it lets the larger-effect GMDs dominate.

Third, the placebo violations also concentrate in GMD 5. The full-sample worst placebo for log water use is -0.089 log points; the drop-GMD-5 worst placebo is -0.072 , a 19 percent reduction. For depth applied the full-sample worst placebo is -0.118 ac-ft per acre and the drop-GMD-5 worst placebo is -0.097 , an 18 percent reduction. This is informative for interpretation: a portion of the residual within-GMD pre-trend is GMD 5-specific, and dropping GMD 5 simultaneously reduces the headline magnitude *and* reduces the placebo violation that motivates the sensitivity analysis in Section D.1–D.2. The headline is therefore not being inflated by a GMD-5-specific bias in any one direction — when GMD 5 is excluded, both the estimated effect and the pre-trend violation shrink in tandem.

The substantive takeaway is that the headline estimates are best read as averages weighted toward GMD 5's treatment dynamics. A reader who prefers an estimate that downweights GMD 5 can read off the drop-GMD-5 row: ≈ 9 percent reduction in annual water use and ≈ 12 percent reduction in depth applied. The conservation effect is clearly present in either case, but its magnitude is not constant across GMDs.

D.4 Synthesis

The two sensitivity frameworks address the pre-trend violation from opposite directions and yield complementary statements about the headline result.

HonestDiD bounds the post-period bias *without* adjusting the point estimate, and tells us that the headline 95% confidence interval loses significance under modestly restrictive bounds on post-period trend deviations: the smoothness threshold $M \approx 0.025$ and the relative-magnitudes threshold $\bar{M} \approx 0.4$. Both thresholds sit inside ranges that are conventionally regarded as plausible. Treated as a one-sided test of whether the headline can be overturned by an empirically reasonable violation, the answer is yes — a reader committed to a strict parallel-trends prior should not treat the dCdH headline of -0.133 as established at conventional significance levels.

The Bilinski-Hatfield residualization adjusts the point estimate *by* netting out an explicit parametric form for the pre-trend, and tells us that under a unit-specific linear-extrapolation counterfactual, the program effect is roughly twice the headline. The residualized estimates pass the joint placebo equivalence test (V1: $p = 0.31$) and yield ATTs in the -0.23 to -0.27 range. The Bilinski-Hatfield framework thus provides an explicit alternative point estimate, not merely a bound, and that estimate is larger in magnitude than the headline.

The two findings are not in tension. HonestDiD’s bounds are conservative under any restriction the researcher accepts, and they say the headline result is fragile under modest restrictions. Bilinski-Hatfield’s residualized estimate makes a stronger structural assumption (linear extrapolation of unit-specific pre-trends) and says that under that assumption, the program effect is larger than the headline. Together, the two frameworks bracket the program effect: the headline -0.133 is plausibly a conservative midpoint between a small-effect interpretation (under which the pre-trend reflects program-irrelevant selection) and a large-effect interpretation (under which the pre-trend reflects a continuing-trajectory counterfactual that the program disrupts more dramatically than a parallel-trends comparison would suggest). The text in the body of this paper retains the headline -0.133 as the primary preferred-specification estimate, but the analysis in this appendix indicates that this number is best read as a lower-bound magnitude, not an upper bound, on the true program

effect.

The depth-applied parallel in Section B.2 delivers the same qualitative conclusion through a partly different mechanism. The HonestDiD smoothness restriction is *more* forgiving for depth applied (the breakdown is beyond $M = 0.05$ ac-ft per acre per period, against $M \approx 0.025$ for log water use), because the depth-applied post-period coefficients are uniformly large and negative across all ten estimated event-time periods rather than concentrating their mass in a handful of horizons. The relative-magnitudes restriction is, however, comparably stringent: the breakdown $\bar{M}$ lies in the same $(0, 0.5)$ window as for log water use, because the depth-applied pre-period violations are themselves large in absolute terms. The Bilinski–Hatfield residualization is informative but partly asymmetric: for log water use the V1 residualization passes the joint placebo equivalence test cleanly ($p = 0.31$) and yields an alternative point estimate of -0.229 ; for depth applied the V1 residualization moves the ATT to -0.317 but does *not* pass the equivalence test ($p = 0.0083$), so it serves as an informative alternative point estimate rather than as a clean parallel-trends-satisfied counterfactual. The combined picture for both outcomes is therefore: the headline ATTs are conservative midpoints between two credible counterfactual stories, and the parallel sensitivity analyses on depth applied corroborate this reading. The within-GMD specification with its known residual pre-trend remains the preferred specification for both outcomes because the structural commitment required to absorb the violation (linear extrapolation of unit-specific pre-trends) is itself a strong assumption.

The leave-one-GMD-out exercise reported in Section D.3 adds a third dimension to this assessment. It shows that the headline estimates are robust in direction and significance to dropping any single GMD, but that their magnitudes are materially anchored in GMD 5 (which supplies about half the identifying variation). Excluding GMD 5 attenuates the log water-use average total effect to -0.095 (≈ 9 percent reduction) and the depth-applied average total effect to -0.128 (≈ 12 percent reduction) — both still significant, but markedly smaller than the headlines. Importantly, the residual within-GMD pre-trend violations also concentrate in GMD 5: dropping GMD 5 reduces both the estimated effect and the worst-placebo magnitude by similar fractions, suggesting that the

headline is not inflated by a GMD-5-specific selection bias in any one direction. The three sensitivity exercises — HonestDiD bounds, Bilinski–Hatfield residualization, and leave-one-GMD-out — jointly support a reading of the headline as a real and significant conservation effect whose magnitude is best interpreted as the weighted average across GMDs rather than as a constant treatment effect that would survive arbitrary sample restriction.