Autonomous Control, Decision Authority, and Firm Performance*

Cevat Giray Aksoy† Prithwiraj Choudhury‡ Anton Grahed§

September 5, 2026

Abstract

Autonomous technologies are changing how firms allocate routine decisions between workers and software. We study this shift in a randomized experiment at a hydropower plant of a large publicly listed energy firm. Experimental days are assigned to standard manual operation, autonomous control, or an information-enhanced manual regime in which operators receive an automated reservoir-information update but retain responsibility for decisions and execution. The plant operates against a predetermined electricity generation schedule fixed before the operating mode is revealed. Autonomous control reduces schedule deviations by 26% relative to standard manual operation and significantly outperforms the information-enhanced regime. High-frequency data point to faster corrective response as an important mechanism. Autonomous control reduces response time by 57% and increases the share of deviations resolved within ten minutes by 16 percentage points. Deviation-related settlement costs fall by 28%.

Keywords: autonomous control; firm performance; automation; decision authority
JEL: C93; D22; O33

*Can Göçer provided outstanding research assistance. We are grateful to the firm's employees for their significant time and effort in preparing for and implementing the experiment. We also thank Ralph De Haas and attendees at the CEPR Workshop on Incentives, Management and Organisation and seminars at the EBRD and King's College London for helpful comments and suggestions. The views expressed are those of the authors and do not represent those of the EBRD.

†European Bank for Reconstruction and Development (EBRD); King's College London; CEPR.
‡London School of Economics.
§EBRD.

1 Introduction

Autonomous technologies are expanding the set of decisions that firms can delegate to software. In many production environments, digital systems no longer only measure conditions, generate forecasts, or execute commands chosen by workers. They also increasingly monitor an operating state, determine a routine response, implement it directly, and continue adjusting as conditions change (Baldea et al., 2025). The economic significance of this shift is organizational as well as technological. As autonomy expands, firms must decide which recurring decisions should remain with workers and which can be delegated to software.[1]

This choice matters because human judgment and autonomous systems have different comparative advantages. Workers can draw on experience and exercise discretion in unusual situations that are difficult to codify. But requiring workers to intervene in every routine adjustment can create delays between a change in conditions and the firm's response. Delegating routine decisions to software can reduce this friction, while leaving workers responsible for cases in which judgment is especially valuable. The resulting trade-off is fundamentally organizational: how to allocate authority, information, and problem-solving between workers and software (Aghion and Tirole, 1997; Garicano, 2000; Dessein, 2002; Dessein and Santos, 2006). When decisions are time-sensitive, that allocation can directly affect performance because delays in processing information and coordinating a response are themselves costly (Bolton and Farrell, 1990; Van Zandt, 1999).

Recent experiments on AI at work primarily study technologies that assist workers with tasks such as writing, customer support, and medical decision-making (Agarwal et al., 2023; Noy and Zhang, 2023; Brynjolfsson et al., 2025). Related work studies algorithmic predictions and recommendations and the role of human discretion (Hoffman et al., 2018; Kleinberg et al., 2018; Angelova et al., 2026), while theory has begun to study explicitly how decision authority should be allocated between humans and artificial intelligence (Athey et al., 2020). Yet we know much less about what happens inside firms when a recurring set of decisions and their execution are delegated to an autonomous system.[2]

We provide randomized field evidence on this organizational margin by reallocating decision authority over a recurring real-time production task from workers to software. The experiment takes place at a reservoir hydropower plant within a large publicly listed energy firm. Before each operating day, the plant receives an hourly generation schedule that is fixed before the experimental operating mode is revealed. During the day, the national transmission-system operator, which manages the electricity grid, can also issue dispatch instructions that change the plant's required output. The plant must therefore keep realized generation close to the

[1] Recent industry evidence points to rapid movement in this direction. In a 2025 survey of 216 senior manufacturing executives in North America and Europe, 74% expected AI agents or software to manage between 11% and 50% of routine production decisions within the next three years (Tata Consultancy Services, 2025). A 2026 survey of 400 senior executives in the energy and chemicals sector across 12 countries reported an average autonomous-operations maturity of 3.52 out of 5, between 'Advanced Regulatory' and 'Select Autonomy', and about 32% identified advancing autonomy as a critical priority over the next five years (Schneider Electric, 2026).

[2] Evidence from industrial technology adoption addresses a related question, but typically relies on variation in technology adoption across firms, plants, industries, or countries (Hubbard, 2000; Bresnahan et al., 2002; Bartel et al., 2007; Graetz and Michaels, 2018). Such comparisons bundle technology choice with persistent differences in equipment, workers, management, and operating conditions.

required level as conditions evolve.[3] Meeting these requirements involves a continuous corrective process: the plant must detect a deviation, assess current conditions, determine and implement the appropriate correction, and then reassess whether further action is needed. This setting allows us to vary who carries out that process while holding the physical plant and production environment fixed.

We use a day-level *switchback* design in which each 24-hour operating window is randomly assigned to one of three modes. The experiment spans 87 days and 2,088 operating hours, with near-complete compliance. Under *Control*, trained operators carry out this corrective process manually using the plant's existing systems. Under *Treatment*, these routine functions are delegated to an autonomous system, which continuously monitors the operating state, determines corrections, sends commands directly to the generating units, and reassesses whether further adjustment is needed. Operators remain present throughout, retain the right to return the plant to manual operation, and handle exceptional cases. Treatment shifts routine decision authority and execution to the autonomous system while preserving human oversight.

A third arm, *Information Only*, helps interpret why Treatment improves performance. Relative to Control, Treatment changes both the information used in real time and who makes and executes routine decisions. Under Information Only, operators retain full decision authority and execution but receive an automated, centralized update on reservoir water level. Because reservoir level is already observable under Control, this arm does not provide new information; instead, it makes an important operating input more salient and easier to access. Information Only therefore asks whether improving the delivery of information, while keeping decisions and execution with operators, can reproduce the Treatment gains. Since Treatment also draws on a broader set of operating information, the Treatment–Information Only comparison is best interpreted as the full autonomous regime relative to an information-enhanced manual regime.

Delegating the routine response to the autonomous system improves performance. Treatment reduces the hourly absolute deviation between realized and required generation by 0.537 MWh/hr, or 25.6% relative to the Control mean. The estimate is highly stable across specifications and remains statistically significant under alternative inference procedures. Information Only produces a smaller reduction of 0.207 MWh/hr, approximately 9.9% of the Control mean, which is not statistically distinguishable from zero. Treatment reduces deviations by an additional 0.330 MWh/hr relative to Information Only, and this difference is statistically significant at the 1% level. The full autonomous regime substantially outperforms both standard manual operation and the specific information-enhanced manual regime tested alongside it.

The high-frequency data show where the organizational advantage of autonomous control emerges within the corrective process. We define the start of a deviation episode using the plant's operating alert threshold and apply the same definition across all three arms.[4] Treatment and Control begin to diverge over the next five to fifteen minutes. The key difference is how quickly corrective action starts. For episodes followed by corrective action, Treatment reduces the time

[3] Deviations are economically costly because excess and shortfall imbalances are settled at unfavorable prices in the balancing market, while sufficiently large departures from required generation incur additional schedule-deviation penalties.

[4] The plant uses an absolute deviation of 3.6 MWh/hr as its operating alert threshold. For the analysis, an episode is considered resolved once the deviation falls back below a resolution threshold of 1.8 MWh/hr, half the onset threshold, as an arm-neutral convention.

from deviation onset to the first action by 7.26 minutes, or 57% relative to the Control mean. Once action has begun, however, the remaining time until the deviation falls back below the resolution threshold is essentially unchanged. Consistent with this faster initial response, across all deviation episodes Treatment raises the share brought back below the resolution threshold within ten minutes from about 53% under Control to about 69%. The evidence therefore points to faster initiation of corrective action as an important channel: delegating the routine corrective response to the autonomous system shortens the time between a deviation arising and the firm beginning to respond.

The Information-Only arm helps sharpen this interpretation. If the faster response under Treatment were driven mainly by making reservoir information more salient, Information Only should also reduce the time to the first corrective action. It does not produce a detectable reduction in response time. Although Information Only improves some outcomes, including how quickly deviations return below the resolution threshold and deviation-related settlement costs, it does not reproduce the much faster initiation of corrective action under Treatment. The larger gain therefore appears to come from combining information processing with the reallocation of routine decision authority and execution to the autonomous system.

The improvement also has meaningful financial value for the firm. Valuing deviations using the settlement rules and market prices actually faced by the firm, Treatment reduces deviation-related settlement costs by approximately 28%. Under our baseline annualization assumptions, this corresponds to an estimated gross benefit of approximately $20,600 per plant-year. Because we do not observe implementation, licensing, or maintenance costs, we cannot estimate the realized net return. Instead, we ask how much the firm could spend up front on the technology and still recover that cost from the savings it generates. Assuming the same annual benefit for ten years and a 10% real discount rate, the firm could spend up to about $126,000 per plant and still break even before ongoing costs; whether the investment yields a positive net return depends on ongoing licensing, monitoring, and maintenance costs, which we do not observe.

The Treatment effect is also robust to alternative levels of aggregation, sample restrictions, and outcome definitions, and is essentially unchanged when we restrict the sample to compliant hours. We find no robust evidence that the estimated Treatment effect is driven by carryover across experimental days. The estimate remains similar when we account for lagged assignments and persistent hydrological conditions or exclude the opening hours of each experimental day, and we find little evidence that current assignment predicts next-day reservoir conditions or scheduling outcomes.

We contribute to a growing literature on how algorithms and artificial intelligence reshape decision-making inside firms. A closely related literature studies settings in which algorithms provide recommendations but workers retain discretion over whether to follow, override, or modify them. These studies show that departures from algorithmic prescriptions are common and can either improve or reduce performance depending on the task and the information available to workers (Kawaguchi, 2020; Kesavan and Kushwaha, 2020; Sun et al., 2021; Kesavan et al., 2025). More recent experiments on generative AI show that these technologies can raise worker productivity and improve task performance while leaving workers responsible for the final output (Noy and Zhang, 2023; Brynjolfsson et al., 2025). Our experiment studies a different

organizational margin: who holds decision authority over a recurring production process. Routine decisions and their execution are delegated to the autonomous system, while workers retain oversight and responsibility for exceptional cases. The experiment provides field evidence on a central question in the theory of human–AI organization (Athey et al., 2020): what happens to firm performance when routine decision authority and execution are delegated from workers to an autonomous system.

The mechanism connects the experiment to classic work on information processing inside organizations (Bolton and Farrell, 1990; Radner, 1993; Van Zandt, 1999): autonomous control accelerates the information-to-action process in a real production environment. The within-plant design also provides direct evidence on technology and performance in physical production. By repeatedly randomizing the operating regime within the same plant, we avoid comparisons across production units that differ persistently in physical technology, workers, management, and operating conditions, while observing adjustment at the frequency at which the relevant decisions are made. In this sense, the paper extends the insider-econometrics tradition of using within-firm variation to study determinants of performance (Lazear, 2000; Ichniowski and Shaw, 2003; Bandiera et al., 2007; Bloom et al., 2013) to the allocation of routine decision authority between workers and autonomous systems.

2 Setting and Experimental Design

We study a setting in which performance depends on how quickly and consistently a firm converts real-time information into corrective action. The experiment changes who carries out this response while holding the plant and its physical production technology fixed. This section describes the operating problem, the autonomous technology, and the randomized switchback design.

2.1 The firm, plant, and operating problem

Our setting is a reservoir hydropower plant operated by a large publicly listed energy firm. The plant operates against an hourly generation schedule and must also respond to dispatch instructions from the national transmission-system operator (hereafter the grid operator). Its central operating objective is therefore to keep realized generation close to the required output in each hour. Deviations are costly because they expose the firm to balancing-market settlement costs and, when sufficiently large, additional schedule-deviation penalties. Figure B.1 shows the plant and its hydropower infrastructure.

Under normal operation, operators, who are trained engineers, manage this schedule-following process manually. They monitor realized generation, reservoir conditions, equipment status, the planned generation schedule, and grid instructions using the plant's existing operating systems. Figure B.2 shows the control-room environment and operating interfaces used to monitor these conditions. When realized output deviates from the required level, an operator must detect the deviation, determine the appropriate correction, enter a start, stop, or setpoint command, and then reassess the plant to judge whether another correction is needed.[5] The relevant object is therefore a corrective process rather than a single operating decision.

[5] A setpoint is the output level that a generating unit is instructed to maintain.

This process is demanding because operating conditions change continuously and corrective actions are time-sensitive. Delays in detecting a deviation, determining a response, or implementing it leave the plant away from its required output, and a single correction need not be sufficient.

2.2 The technology and experimental arms

Treatment activates an autonomous control system that leaves the plant's physical production technology unchanged and instead changes how operating information is converted into decisions and actions.

The system continuously combines information on realized generation, reservoir conditions, equipment status, the planned generation schedule, and instructions from the grid operator. It compares realized output with the required level, determines whether a routine correction is needed, selects the appropriate start, stop, or setpoint command, and sends that command directly to the plant. It then reassesses the operating state and can issue further corrections when needed.

The defining feature of the technology is therefore that it both determines and executes routine corrective actions. It is not simply an information dashboard or the automated execution of commands chosen by an operator. In engineering terms, this sequence makes it a closed-loop control system.[6] The control logic evaluated in the experiment is rule-based rather than learning-based: it applies predetermined decision rules to the continuously observed plant state. These rules were fixed before the experiment began and were not modified during the study period, so the Treatment regime is the same on every Treatment day.

The experiment uses a day-level switchback design in which the same plant rotates across three operating modes. An experimental day is a 24-hour operating window running from 08:00 to 08:00 the following day.

Under *Control*, the plant follows the standard manual regime described above: operators use the existing information and operating systems and remain responsible for every stage of the corrective process. The autonomous system is inactive.

Under *Information Only*, the plant continues to operate manually, but operators receive an automated, centralized update on reservoir water level. The update is pushed proactively rather than left for operators to retrieve.[7] Reservoir level is an important operating input because it affects the plant's ability to adjust generation. Operators nevertheless retain full responsibility for interpreting the information, deciding whether and how to respond, and executing any correction.

Under *Treatment*, the autonomous system described above carries out the routine corrective response directly. Operators remain present to monitor the system and handle exceptional cases. They do not review or veto individual system commands; their intervention right, which they retain throughout the experiment, is to return the plant to full manual operation in the event of a system malfunction. Table A.1 compares the three operating modes in greater detail.

[6] Appendix B.1 describes the technology in more detail, including the rule-based control loop and a separate machine-learning analytics layer used elsewhere in the broader platform. The latter supports monitoring and prediction but is distinct from the routine closed-loop execution studied here.

[7] The automated update arrives just under once per hour on Information-Only days, and take-up is incomplete on a minority of days.

The three-arm design gives each comparison a clear interpretation. Treatment relative to Control identifies the total effect of autonomous operation relative to the plant's standard manual regime. Information Only relative to Control measures the effect of improving the delivery of reservoir information while leaving decision authority and execution with operators. Treatment relative to Information Only compares the full autonomous regime with this information-enhanced manual regime. Because Treatment draws on a broader set of operating information, this contrast does not isolate automation while holding information fixed.

2.3 Randomization, timing, and implementation

Randomization was conducted in permuted blocks of three consecutive experimental days. Each block contains one Control, one Information-Only, and one Treatment day, with their order randomized within the block. The experiment ran for 87 consecutive days from December 23, 2025 to March 19, 2026, producing 29 days in each arm. The realized assignments are well interleaved throughout the study window; Figure A.1 shows the full sequence.

The generation schedule used to measure performance is set independently of experimental assignment. The scheduling desk prepares and fixes the plant's day-ahead hourly schedule before the 24-hour experimental window begins, and the randomization sequence, held by the research team, was not available to the desk, so the schedule could not be adjusted in response to the assigned operating mode. Each day's assignment was provided by the research team to the on-site engineer responsible for switching the operating mode. The engineer switched the mode at 08:00 and informed the operating staff at that point, so operators learned the day's mode only once the window began and the day-ahead schedule for it had already been fixed; operators did not have access to the assignment sequence and were not told future assignments in advance. Any subsequent within-day dispatch instructions are issued by the grid operator and are likewise independent of assignment. Because operators necessarily work under the day's mode, our estimates capture the effect of assignment to each operating regime, including any behavioral response by operators to the assigned mode. We verify the schedule's independence directly in Section 3 (Panel A of Table A.2).

Implementation compliance was near complete. The experimental protocol was followed for the full 24-hour window on 79 of the 87 experimental days and for 97.1% of the 2,088 experimental hours overall.[8] We estimate intention-to-treat effects based on the original randomized assignment; restricting the sample to the 2,027 protocol-compliant hours produces essentially the same estimate (Section 5.4).

2.4 Identification and carryover

Identification comes from randomized changes in operating mode within a single plant. Because the plant repeatedly operates under Control, Information Only, and Treatment, the design does not rely on comparisons across plants that may differ in equipment, workforce, management,

[8]The small number of non-compliant hours occurred at the end of eight experimental days (four Control, two Information Only, and two Treatment). On the affected Treatment and Information-Only days, the autonomous system or automated information feed was switched off for the remaining overnight hours; on the affected Control days, the corresponding administrative stand-down did not alter the plant's standard manual operating mode.

location, or other persistent characteristics. Assignment occurs at the experimental-day level, while outcomes are observed hourly, allowing us to measure performance as operating conditions and required output change over time.

The main design-specific concern is whether the effect of one operating mode persists into the following experimental day. For example, Treatment could change reservoir conditions in a way that affects subsequent performance even after the plant switches to another mode. Section 5.4 tests for such carryover using lagged assignments, design-based randomization inference, and specifications that account for persistent hydrological conditions, and shows that the contemporaneous Treatment estimate remains similar when we exclude the opening hours of each experimental window and when we aggregate the data to the day level. These checks follow recent work on the design and analysis of switchback experiments (Bojinov et al., 2022).

3 Data, Outcomes, and Balance

Our analysis combines high-frequency plant operations data, generation schedules, electricity-market data, the randomized assignment sequence, and timestamped operational logs. The core operational data come from the plant's supervisory control and data acquisition (SCADA) system and record realized electricity generation, reservoir water levels, and flow rates. We merge these data with the plant's hourly generation schedule and electricity-market prices and link each observation to the assigned operating mode. The final sample contains 87 experimental days and 2,088 hourly observations.

For the mechanism analysis, we combine the SCADA data with timestamped operational records of corrective actions and plant responses; corrective actions are drawn from a common command log with the same time base in all three arms. We use the high-frequency generation data to identify deviation episodes with a common onset threshold in all three arms and align the operational records to deviation onset. This allows us to measure response time, defined as the interval from deviation onset to the first corrective action, and the subsequent time from that action until the deviation returns below the resolution threshold. We also calculate the share of deviations brought back below the resolution threshold within ten and thirty minutes of onset.

Our primary outcome is the hourly absolute deviation between realized and required generation. In ordinary hours, this is

$$D_{h,t} = \left| Y_{h,t} - \widehat{Y}_{h,t} \right|,$$

where $Y_{h,t}$ denotes realized generation in hour h of experimental day t, and $\widehat{Y}_{h,t}$ denotes required generation in that hour, based on the plant's day-ahead schedule, which is fixed before the experimental window begins (Section 2.3), adjusted for any within-day dispatch instructions issued by the grid operator. In hours in which the plant provides frequency-control services and is permitted to operate within a band (approximately 15% of hours), the absolute deviation is zero when realized generation lies inside the permitted band and equals the absolute distance to the nearest band boundary otherwise; excluding these hours entirely produces a somewhat larger Treatment effect and leaves the conclusion unchanged (Figure A.2). Lower values indicate that the plant follows its required production profile more closely. Throughout, generation and

deviations are expressed in MWh per hourly settlement interval, written MWh/hr; because balancing settlement is hourly, this equals the plant's average megawatt output over the hour.

The outcome captures an economically important operating margin. Under Control, the mean absolute schedule deviation is 2.10 MWh/hr, about 7% of mean scheduled generation of approximately 29 MWh/hr. The production profile also changes substantially within the day: hourly targets range from roughly 6 (and zero in a small number of hours) to more than 100 MWh/hr, and the target changes by about 13 MWh/hr on average from one hour to the next. Following this profile therefore requires frequent adjustment. Section 5.3 translates the estimated reduction in schedule deviations into monetary terms using the settlement costs faced by the firm.

We conduct several checks on the construction of the outcome. Excluding the small number of hours with zero scheduled generation produces a slightly larger Treatment effect, so the result is not driven by periods in which the plant is effectively scheduled to be off. We further examine over- and under-generation separately, log(1 + absolute deviation), price-weighted deviations, and indicators for large or persistent misses. These exercises, reported in Section 5.4, test whether the main result depends on the outcome definition and help locate where in the deviation distribution the gains arise.

Predetermined operating conditions are broadly balanced across the three experimental arms (Table A.3). Average absolute schedule deviations in the four- and eight-week pre-experiment windows defined in the notes to Table A.3 are nearly identical across arms, and starting reservoir levels and day-ahead electricity prices are also balanced. The largest standardized difference is 0.33 standard deviations, for on-site staffing (Treatment below Control); controlling for staffing leaves the Treatment estimate essentially unchanged (Section 5.4). Because our primary outcome is defined relative to the generation schedule, balance in that schedule is particularly important. Mean scheduled generation is approximately 28–29 MWh/hr in all three arms, the Treatment and Information-Only indicators do not predict scheduled generation (Table A.2), and both the mean and maximum hour-to-hour changes in scheduled generation are balanced across arms. The estimated differences in absolute schedule deviations therefore do not reflect systematically different production targets.

4 Empirical Strategy

We estimate intention-to-treat effects of randomized assignment using the following specification:

$$D_{h,t} = \alpha + \beta_1 T_t + \beta_2 \mathrm{IO}_t + \gamma_{d(t)} + \delta_h + X'_{h,t}\theta + \varepsilon_{h,t}, \tag{1}$$

where $D_{h,t}$ is the absolute schedule deviation in hour h of experimental day t, defined in Section 3. T_t is an indicator for assignment to Treatment and IO_t is an indicator for assignment to Information Only, with Control as the omitted category. Because operating mode is randomly assigned, β_1 identifies the effect of autonomous control relative to standard manual operation, while β_2 identifies the effect of Information Only relative to Control. We also report the contrast $\beta_1 - \beta_2$, which compares the full autonomous regime with the information-enhanced manual regime.

The specification includes day-of-week fixed effects, $\gamma_{d(t)}$, and hour-of-day fixed effects, δ_h, which absorb recurring differences across days of the week and systematic intraday variation in operating conditions. The vector $X_{h,t}$ contains the hourly day-ahead electricity price and two predetermined measures of schedule difficulty: the day's mean and maximum hour-to-hour change in scheduled generation. These controls are included to improve precision rather than identification. Because assignment was randomized within three-day blocks, we also report a specification with randomization-block fixed effects, which identifies the arm effects using only within-block variation (Figure A.2).

We do not control for contemporaneous reservoir water level or flow rate because Treatment can itself affect these variables through the plant's operating decisions. Conditioning on them could therefore absorb part of the Treatment effect. All controls in equation (1) are predetermined with respect to the assigned operating mode. As shown in Table 1, the estimated Treatment effect is stable when these controls are excluded. Column (4) of Table 1, which includes both sets of controls, is our main specification.

In the main specifications, standard errors are clustered by experimental day, the unit of randomization. We also report randomization-inference p-values (Young, 2019), obtained from 2,000 re-randomization draws that preserve the same permuted-block assignment procedure used in the experiment. Section 5.4 reports several additional inference checks: week-level clustering, which allows for dependence across nearby experimental days; an unrestricted permutation test based on 2,000 draws; and specifications that aggregate the data to one observation per experimental day, so that the number of observations matches the number of randomized units. For the family of secondary outcomes, we report Romano–Wolf stepdown p-values that adjust for multiple comparisons.

For the mechanism analysis, we estimate the same experimental-arm contrasts using outcomes defined at either the experimental-day or deviation-episode level. Day-level outcomes use the same day-of-week fixed effects and heteroskedasticity-robust standard errors, while standard errors for episode-level outcomes are clustered by experimental day. Section 3 defines the corresponding mechanism outcomes.

5 Results

This section presents the results in four steps. Section 5.1 reports the effect of autonomous control on schedule conformance and briefly examines where the gains are concentrated. Section 5.2 then examines the mechanisms behind this improvement, using high-frequency generation data and operational logs to examine whether autonomous control changes the deviations initially faced by the plant or instead improves the corrective response once a deviation occurs. Section 5.3 translates the estimated improvement in schedule conformance into monetary benefits and assesses the implementation cost at which the technology would break even. Finally, Section 5.4 presents a series of robustness checks, addressing the main design-specific concern in the switchback experiment and showing that the results are stable across alternative inference procedures, specifications, samples, and outcome definitions.

5.1 Main Results

Figure 1 provides a first view of the main result. Across all 2,088 hourly observations, the raw mean absolute schedule deviation is 2.10 MWh/hr under Control, 1.92 MWh/hr under Information Only, and 1.59 MWh/hr under Treatment. Figure A.3 shows the corresponding raw distribution of log(1 + absolute deviation): the Treatment distribution is shifted toward smaller deviations relative to both Control and Information Only, while Information Only lies between the other two arms but remains closer to Control.

Table 1 reports the regression estimates. In our main specification, Treatment reduces the hourly absolute schedule deviation by 0.537 MWh/hr, equivalent to a 25.6% reduction relative to the Control mean of 2.099 MWh/hr. The estimate is statistically significant at the 1% level and is highly stable across the specifications in Table 1, ranging from 25.6% to 26.1% of the Control mean. Inference based directly on the experimental assignment procedure gives the same conclusion: re-randomizing the three operating modes across days using the permuted-block design produces a randomization-inference p-value of 0.001. The specification with randomization-block fixed effects described in Section 4 yields a Treatment estimate close to our main estimate, at −0.482 MWh/hr (SE 0.093), a 23% reduction relative to the Control mean (Figure A.2).

Information Only moves the outcome in the same direction, but the estimated improvement is considerably smaller. It reduces the hourly absolute schedule deviation by 0.207 MWh/hr, or approximately 9.9% of the Control mean and about two-fifths of the Treatment effect. The estimate is not statistically significant at conventional levels. We therefore do not interpret this result as showing that improved information delivery has no value.[9] The correct reading is narrower: the experiment does not detect an improvement under Information Only comparable to that generated by the full autonomous regime.

The direct comparison with Information Only confirms this reading. Treatment lowers absolute schedule deviations by an additional 0.330 MWh/hr, with the gap statistically significant at the 1% level. This result is stable across all specifications in Table 1 and under randomization inference. Thus, the full autonomous regime improves schedule conformance relative to both standard manual operation and the information-enhanced manual regime.

Before turning to mechanisms, we ask whether the Treatment effect is concentrated in particular operating conditions. We find little evidence of systematic heterogeneity along predetermined margins: there is no evidence that the effect increases with hydrological persistence (Table A.4), and it is negative whether required hour-to-hour changes in scheduled generation are relatively small or large, with both estimates negative and their confidence intervals overlapping substantially (Figure A.2). We also decompose the effect by the direction of the deviation. The point estimate is larger for over-generation than for under-generation, but the difference between the two effects is not statistically significant (Table A.5).

[9]The experiment is well powered for the Treatment effect but not for a small Information-Only effect. The ex post minimum detectable effect at 80% power and a 5% two-sided level, computed from the realized experimental standard errors, is about 0.35 MWh/hr for the Treatment relative to Control and 0.39 MWh/hr for Information Only (about 18% of the Control mean), so the design could not reliably detect an Information-Only effect of the magnitude estimated (0.207 MWh/hr). We therefore rely on the directly estimated and statistically significant Treatment-minus-Information-Only contrast ($p = 0.007$) rather than on the Information-Only level effect.

Treatment does, however, substantially reduce the incidence of the largest operating errors: it lowers the probability of an extreme scheduling miss by about 61% and the probability of a sustained high-deviation episode by about 58% (Table A.6). The secondary-outcome results therefore show particularly large reductions in extreme and sustained deviations rather than a uniform shift across all hours, which motivates the mechanism analysis below.

5.2 Mechanisms

We next explore the mechanisms behind the improvement in performance under autonomous control. We first ask whether Treatment changes the deviations that initially arise or instead changes how the plant responds once a deviation has occurred. We then examine how quickly any difference between Treatment and Control emerges after a deviation begins. Finally, we use operational logs to identify where in the corrective process the two regimes differ and whether improving the delivery of reservoir information through Information Only produces a similar response. Together, these analyses help distinguish between changes in the initial deviation, the timing of corrective action, and the remaining adjustment once action has begun. Because deviation episodes are defined from realized generation, which can itself respond to Treatment, we interpret the episode-based analysis as evidence on where the Treatment advantage emerges within the corrective process rather than as a causal decomposition of the overall effect.

The timing of the Treatment effect provides a first way to distinguish between these explanations. If autonomous control improves performance by preventing or dampening developing deviations before they grow large, differences should already be visible outside deviation episodes or when a deviation first begins. If instead its main advantage lies in the response once a deviation is under way, the effect should be concentrated within active deviation episodes and emerge after the deviation has started. Figure A.4 descriptively splits the raw Treatment–Control mean gap between active deviation-episode hours and periods in which the plant is operating close to target. Essentially all of the raw gap is concentrated in deviation-episode hours, consistent with the Treatment advantage emerging during the corrective response rather than as a general shift during calm periods.

We then compare measured deviation severity when an episode first crosses the common threshold. At onset, the average deviation is 4.795 MWh/hr under Treatment and 4.834 MWh/hr under Control, a difference of -0.04 MWh/hr (SE 0.05; randomization-inference $p = 0.49$) that is not statistically significant. Because an episode is defined to begin when the deviation crosses the common 3.6 MWh/hr threshold, similar measured severity at the crossing is partly built into the design, so we do not read the near-zero onset gap on its own as evidence that Treatment cannot affect how deviations form. We therefore place more interpretive weight on the subsequent divergence between the two regimes.

Figure 2 shows how the Treatment advantage builds over the minutes after a deviation begins, reporting the estimated effect on the deviation in five-minute bins over the first hour after onset. The effect is smaller in the first five minutes (about -0.28 MWh/hr) and grows to roughly -0.7 MWh/hr within five to fifteen minutes, as the corrective response takes hold; it remains negative across the rest of the first hour, while the Information-Only effect stays close to zero in

every bin. This timing is consistent with the Treatment advantage emerging in the corrective response after a deviation begins.

Figure 2 shows when the improvement appears, but not what changes within the corrective response. We therefore turn to the operational records. They let us answer three questions, the second and third of which are summarized in Table 2 and Figure A.5, respectively. First, did Information Only change the delivery of operating information as intended? Second, did autonomous control change how quickly the first corrective action was taken? Third, was this earlier response accompanied by a greater probability that the deviation was resolved quickly?

The first check confirms that Information Only changed the intended information channel. Operational records show that operators engaged with the dedicated automated water-level update on about 86% of Information-Only days. Reservoir levels were already observable through the plant's standard operating systems under Control, so the intervention changes the delivery and salience of this information rather than access to it. Operators under Information Only nevertheless remain responsible for deciding what action to take and executing that action. We can therefore examine whether improving this information channel reproduces the changes in the corrective response observed under Treatment, which integrates a broader operating state.

The clearest difference appears in response time. We measure response time from the moment a deviation crosses the same 3.6 MWh/hr threshold under all three operating modes to the first corrective action. For episodes followed by corrective action, the Control mean of the day-level response-time measure is 12.74 minutes. Treatment reduces this measure by 7.26 minutes, or 57%, and the effect is statistically significant at the 1% level. Information Only reduces it by just 0.48 minutes, and the estimate is not statistically significant. The Treatment–Information Only difference is 6.79 minutes and is also statistically significant at the 1% level (Table 2). Autonomous control therefore begins corrective action substantially sooner, whereas improving the delivery of reservoir information does not produce a detectable change in the timing of the first action. Response time is defined only for episodes followed by a corrective action, so the set of episodes entering this measure can differ across arms. Appendix B.2 shows that the estimate remains large under a conservative exercise that assigns episodes returning below the threshold before any corrective action a response time equal to their time to return below the threshold.

We next ask whether this faster initial response translates into faster resolution of deviations. Across all deviation episodes, 53% are brought back below the resolution threshold within ten minutes under Control, compared with about 69% under Treatment and 59% under Information Only (Figure A.5). Within thirty minutes, the corresponding shares are 90%, 96%, and 93%. Treatment brings deviations back below the threshold significantly faster than Control across the full distribution (log-rank test, design-based randomization-inference $p < 0.001$, permutation at the experimental-day level). Information Only is also faster than Control ($p = 0.013$), but its improvement is roughly two-fifths the size of Treatment's, mirroring the partial improvement it produces in the primary outcome. The advantage of autonomous control is largest early in the response and narrows as the horizon lengthens.

We then ask whether Treatment also shortens the remaining time from the first corrective action until the deviation falls back below the resolution threshold. The Control mean of this day-level measure is 1.46 minutes. This second measure is defined on the narrower subset of

action-followed episodes that also return below the resolution threshold within the experimental day. The estimated differences are only 0.008 minutes under Treatment and −0.002 minutes under Information Only (Table 2). The main timing difference therefore occurs before the first corrective action: autonomous control begins the response substantially sooner. This result is consistent with the event-time evidence in Figure 2, where the Treatment effect on the deviation builds over the first five to fifteen minutes after onset and persists across the response window. The same qualitative patterns are also robust to the choice of episode threshold (Appendix B.2).

Two qualifications are important for interpreting these results. First, Information Only does not reproduce the full set of information continuously available to the autonomous system. The Treatment–Information Only comparison therefore does not isolate automated execution while holding information fixed. Instead, it shows that the full autonomous regime produces a faster corrective response than an information-enhanced manual regime in which operators receive an automated reservoir-information update but retain responsibility for deciding and executing the response. Second, we do not measure the full sequence of later corrective actions, so we cannot determine whether Treatment also changes the number, size, or sequencing of subsequent adjustments. We therefore do not interpret the results as a complete decomposition of every channel through which autonomous control affects performance.

Taken together, the timing and operational evidence points to faster initiation of corrective action as an important channel behind the improvement: autonomous control begins corrective action much sooner after a deviation occurs and brings substantially more deviations back below the threshold within the first ten minutes, while the time from the first corrective action until the deviation returns below the threshold is essentially unchanged across operating modes. The specific reservoir-information enhancement tested here does not produce a response comparable to the full autonomous regime. The larger gain therefore appears to come from combining information processing with the reallocation of routine decision authority and execution to the autonomous system.

5.3 Economic Value and Break-Even Cost

We next ask whether the improvement in schedule conformance is economically meaningful. Because we do not observe the cost of implementing the autonomous system, we cannot calculate a realized net return. Instead, we estimate the monetary benefit generated by the Treatment effect and then ask how large the implementation cost could be before the investment would no longer break even.

When the plant generates more or less electricity than required, the difference must be settled in the balancing market. These deviations are costly because excess generation is typically sold at a discount, shortfalls are purchased at a premium, and sufficiently large deviations incur an additional penalty. Autonomous control therefore generates a financial benefit to the extent that it reduces these deviation-related costs.

We calculate this benefit in two steps. First, we use the market prices and settlement rules observed in each experimental hour to convert the plant's deviation into a monetary cost. Second, we estimate the Treatment effect on this hourly cost using the same specification as in our main analysis, expressing the outcome in monetary terms. We value the Treatment-minus-Control

contrast throughout, because Control, the plant's standard manual regime, is the status quo that an adoption decision would replace; the Information-Only arm serves as an experimental benchmark for the information channel rather than a separately deployed alternative.

To formalize the calculation, let $\Delta_{h,t}$ denote the signed counterpart of the schedule-deviation measure defined in Section 3, using the same permitted-band adjustment in frequency-control hours, with positive values indicating over-generation and negative values indicating under-generation. Define $O_{h,t} = (\Delta_{h,t})^+$ and $U_{h,t} = (-\Delta_{h,t})^+$, where $(x)^+ = \max\{x, 0\}$. Let $P^{DA}_{h,t}$ denote the day-ahead electricity price, $P^{B}_{h,t}$ the balancing-market price, and $S_{h,t}$ scheduled generation. The hourly cost of a deviation is

$$\begin{aligned} \text{Cost}_{h,t} = {} & O_{h,t}\left[P^{DA}_{h,t} - 0.97\min(P^{DA}_{h,t}, P^{B}_{h,t})\right] \\ & + U_{h,t}\left[1.03\max(P^{DA}_{h,t}, P^{B}_{h,t}) - P^{DA}_{h,t}\right] \\ & + 0.03\max(P^{DA}_{h,t}, P^{B}_{h,t})\left(|\Delta_{h,t}| - 0.05S_{h,t}\right)^+ . \end{aligned} \tag{2}$$

The first two terms capture the cost of over- and under-generation relative to the day-ahead price. The final term captures the additional schedule-deviation penalty, which applies only to the portion of the absolute deviation above 5% of scheduled generation.[10]

Applying our main specification to this monetary outcome, Treatment reduces deviation-related settlement costs by 139.0 local-currency units per hour, or 27.8% relative to the Control mean of 499 local-currency units per hour, and the estimate is statistically significant at the 1% level (Table A.7). Information Only reduces these costs by 85.9 local-currency units per hour and is statistically significant at the 5% level. Although Treatment produces a significantly larger reduction in the primary schedule-deviation outcome than Information Only, the Treatment–Information Only difference in total settlement costs (53.1 local-currency units per hour) is not statistically distinguishable from zero.

Information Only reduces settlement costs significantly even though its effect on the primary performance measure is imprecise, because the two outcomes weight the same deviations differently. Our main outcome treats each MWh of absolute deviation equally, whereas the monetary outcome assigns greater weight to deviations that occur when the associated settlement costs are higher. The monetary calculation therefore asks a different question: what is the financial value of the observed change in the pattern of deviations?

Most of the estimated Treatment benefit comes from reducing over-generation. Treatment lowers the cost associated with over-generation by 109.8 local-currency units per hour, or 33.0% relative to the Control mean, while the estimated reduction in the under-generation component is smaller and not statistically significant (Table A.7). The additional penalty applied when absolute deviations exceed 5% of scheduled generation also falls by 36.0 local-currency units per hour, or 39.3%. The over- and under-generation components sum to the total effect; the

[10] During the experimental period, over-generation was settled at 0.97 times the lower of the day-ahead and balancing-market prices, while under-generation was settled at 1.03 times the higher of the two prices. For deviations exceeding 5% of scheduled generation, an additional charge equal to 3% of the relevant market price applied to the portion of the absolute deviation above 5% of scheduled generation, as in the final term of equation (2). The same settlement rule applied throughout the experimental period, and we apply these rates throughout.

additional schedule-deviation penalty is already embedded in these directional cost measures and should therefore not be added to them separately. Figure 3 shows the estimated settlement-cost reduction and the resulting annual benefit.

Translating the hourly benefit into an annual figure requires one additional consideration specific to reservoir hydropower. When autonomous control reduces over-generation, it also conserves water. Outside periods of very high inflow, this water can remain in the reservoir and be used to generate electricity later, so avoiding unnecessary generation has economic value. During the spring spill season, however, the reservoir may already contain more water than can be stored. Conserving an additional unit of water then has little or no future value because that water may ultimately have to be released without generating electricity.

We account for this difference when annualizing the Treatment effect. The experimental period lies almost entirely outside the spill season: spill conditions occur in approximately 4% of experimental hours. For the annual calculation, we therefore apply the estimated benefit from lower over-generation to the approximately eight months of the year in which the reservoir is typically below its spill threshold. The under-generation component is treated as year-round because failing to produce committed generation remains costly regardless of reservoir conditions.

Under these assumptions, our baseline annualized gross-benefit estimate is \$20,574, or approximately \$20,600 per plant-year. Of this, about \$14,700 comes from lower over-generation and approximately \$5,900 from the estimated reduction in under-generation. Because the under-generation estimate is not itself statistically distinguishable from zero, a more conservative calculation that excludes it yields approximately \$14,700 per plant-year; we carry this uncertainty through the calculations below.[11]

The annual estimate is subject to uncertainty in the estimated Treatment effect as well as assumptions about market prices and the exchange rate. To illustrate the uncertainty associated with these inputs, we also report a Monte Carlo exercise that varies them simultaneously, drawing each independently as described in Table A.8. The resulting illustrative sensitivity distribution has a median annual benefit of approximately \$22,000, close to the \$20,600 point estimate, with a 5th–95th percentile range of roughly \$12,000 to \$34,500 under the stated parameter assumptions, holding the eight-month non-spill assumption fixed (Figure 3).

Finally, because the implementation cost of the technology is not observed, we calculate the cost at which adoption would break even. Discounting a constant annual benefit at 10% over ten years implies a break-even up-front investment of approximately \$126,000 per plant. For illustration, an implementation cost of \$60,000 would imply a discounted payback period of approximately 3.6 years. Table A.8 reports how these figures respond to the key assumptions: across alternative non-spill windows, discount rates, and asset lives, and when the imprecisely estimated under-generation component is dropped, the annual benefit ranges from about \$14,700 to \$24,200 and the break-even investment from about \$69,000 to \$187,400. These calculations are gross of ongoing licensing, monitoring, and maintenance costs.

[11] We convert the estimates using the average exchange rate during the experimental period. The eight-month non-spill period is an approximation based on the seasonal operation of the reservoir. The over-generation component of the estimated annual benefit scales proportionally with the assumed length of the non-spill period (Table A.8).

These estimates should be interpreted as an economic-value calculation from the firm's perspective rather than as a measure of social surplus. Some balancing-market payments redistribute resources within the electricity system, while the social value depends on the real operating costs avoided. The calculation is also partial-equilibrium: widespread adoption of autonomous control could itself affect balancing-market prices. Finally, deviations across plants within the same generation portfolio may partly offset one another, reducing the settlement cost attributable to any individual plant. Subject to these qualifications, the experimentally estimated improvement in schedule conformance generates a meaningful financial benefit for the plant.

5.4 Robustness

We conclude the results section by assessing the main concerns about the design and estimation. We first test for carryover across experimental days, the key concern specific to the switchback design. We then check whether the statistical conclusions depend on how standard errors and p-values are calculated. Finally, we vary the level of aggregation, empirical specification, estimation sample, compliance restriction, and definition of the performance outcome.

We begin with carryover. If Treatment continues to affect the plant after the assigned day ends, then assignment to Treatment on previous days should predict current schedule deviations even after the plant has switched to another operating mode. We therefore include indicators for Treatment and Information-Only assignment one, two, and three days earlier and test their cumulative effects (Table A.9). In the simpler lag specification (column (1) of Table A.9), the cumulative lagged Treatment coefficient is +0.736 MWh/hr and is statistically significant at the 5% level using standard errors clustered by experimental day. Taken literally, this would imply that performance deteriorates on days following Treatment.

This conventional significance test is difficult to interpret in the permuted-block design. Each three-day block contains exactly one Treatment, one Information-Only, and one Control day, so assignments on neighboring days are mechanically related. A block-preserving randomization exercise illustrates this feature: under the joint sharp null of no assignment effect at any horizon, the design generates cumulative lag sums ranging from about −0.99 to +0.99 between the 5th and 95th percentiles, and the observed +0.736 is not unusual ($p = 0.227$). Because this exercise also imposes a zero contemporaneous Treatment effect, we treat it as a diagnostic of the assignment structure rather than as a standalone test of no carryover.

We therefore assess carryover using a specification that allows the contemporaneous Treatment effect to remain unrestricted while accounting for persistent operating conditions. We add the current and previous three days' values of the predetermined hydrological-persistence measure together with fixed effects for the three-day randomization blocks. In the simpler lag specification, the contemporaneous Treatment coefficient is −0.360 MWh/hr. After accounting for persistent conditions and the block design, the contemporaneous coefficient is −0.551 MWh/hr, close to our main estimate of −0.537 MWh/hr, while the cumulative Treatment lag changes from +0.736 to −0.266 MWh/hr and is statistically indistinguishable from zero ($p = 0.588$; Table A.9). The corresponding cumulative Information-Only lag is −0.957 MWh/hr (SE 0.441) and is statistically different from zero at conventional levels, although it is not statistically distinguishable from the cumulative Treatment lag ($p = 0.150$). This multi-day pattern is not corroborated by the

direct previous-day check reported next: on Control days, a preceding Information-Only day is associated with a small and statistically insignificant +0.15 MWh/hr difference (SE 0.22).

We also test directly whether Treatment affects conditions at the start of the following experimental day. Day-t Treatment assignment has essentially no relationship with the reservoir level at the beginning of day $t+1$: the estimated coefficient is -0.002 meters (randomization-inference $p = 0.97$; $N = 86$ day-to-day transitions). Treatment also does not predict the following day's natural inflow or inflow variability. We also exclude the first four, eight, or twelve hours of each experimental window, when any effect inherited from the previous operating mode should be most likely to matter. The contemporaneous Treatment estimate remains negative and statistically significant across these restrictions (Figure A.2). As a further check, we restrict the sample to days currently operating under Control and ask whether the previous day's operating mode predicts current performance: relative to a preceding Control day, a preceding Treatment day is associated with a statistically insignificant +0.14 MWh/hr difference in the current Control-day deviation (SE 0.21; $N = 696$ plant-hours on 29 Control days). Together with the Information-Only estimate reported above, this indicates that performance under Control does not vary systematically with the immediately preceding operating mode. Taken together, these results provide no robust evidence that the estimated Treatment effect is driven by carryover across experimental days.

A related concern is whether the scheduling desk revises the next day's schedule in response to the current day's operating mode, which could contaminate the following day's outcome. We therefore also regress the day-$(t+1)$ schedule outcomes on day-t assignment (Panel D of Table A.2). Of the eight next-day schedule outcomes examined, only the schedule coefficient of variation shows a nominally significant response ($p = 0.014$; randomization-inference $p = 0.010$). Given that eight next-day outcomes are examined, we interpret this isolated result cautiously; next-day scheduled generation and hour-to-hour changes in scheduled generation are unaffected.

We next check whether the statistical significance of the Treatment effect changes when we use alternative ways of calculating standard errors and p-values. Our main specifications cluster standard errors by experimental day, the unit of randomization. Clustering instead at the week level leaves the conclusion unchanged. An unrestricted permutation test that randomly reallocates the 29 Treatment, 29 Information-Only, and 29 Control assignments across experimental days gives a p-value of 0.001. These results are consistent with the block-based randomization inference reported in Section 5.1, which also gives a p-value of 0.001 (Table A.10). The direct Treatment–Information Only contrast is likewise significant under design-based inference, with a block-preserving randomization-inference p-value of 0.005 and an unrestricted permutation p-value of 0.011, consistent with the Wald test ($p = 0.007$).

We also check whether the result depends on using hourly observations when Treatment is randomized at the day level or on a particular empirical specification. Aggregating the data to one observation per experimental day produces a similar Treatment effect. The estimate also remains stable when we use week fixed effects, linear or quadratic time trends, or fixed effects for the three-day randomization blocks. Varying the controls for schedule difficulty, including alternative functional forms for hour-to-hour changes in scheduled generation, additional measures of the shape of the daily generation schedule, and an ANCOVA specification controlling for

pre-experiment deviations, produces similar estimates, and the effect is also similar in the earlier and later parts of the experiment (Figure A.2).

We next vary the estimation sample and construction of the primary outcome. First, we exclude the small number of hours in which scheduled generation is zero, where any positive generation is mechanically recorded as a deviation from the target; the Treatment effect is slightly larger when these hours are removed. We also trim the largest 1% of deviations to ensure that a small number of extreme observations are not driving the estimate and use a simpler outcome that measures the distance from the hourly production target without adjusting for permitted operating bands. In addition, we exclude frequency-control hours, hours with within-day dispatch instructions, and hours affected by outages or trips, when standard schedule-following is harder to interpret, and restrict the sample to hours in which the plant has operational scope to correct a deviation. Across all of these alternative samples and outcome definitions, the estimated Treatment effect remains similar (Figure A.2).

We also check whether the small amount of non-compliance affects the estimate. Our main results are intention-to-treat effects based on randomized assignment. Restricting the sample to the 2,027 protocol-compliant hours produces an estimate of -0.552 MWh/hr, compared with -0.537 MWh/hr in our main specification. Restricting the sample to experimental days with full compliance gives the same conclusion (Figure A.2).

Finally, we consider alternative measures of the performance outcome. Treatment continues to reduce deviations when the outcome is expressed in logs, and it substantially reduces the probability of extreme scheduling misses and sustained high-deviation episodes (Table A.6). These extreme-miss and sustained-episode indicators use the 90th and 75th percentiles of the pooled deviation distribution. We also estimate effects on over- and under-generation separately, the deviation as a share of capacity, price-weighted deviations, and squared deviations. Because these outcomes form a broader family of secondary tests, we adjust for multiple comparisons. Over the secondary-outcome family, the effects on over-generation, the deviation as a share of capacity, price-weighted deviations, log(1 + absolute deviation), extreme misses, sustained episodes, and squared deviations remain statistically significant after the Romano–Wolf adjustment (Romano and Wolf, 2005), whereas the weaker under-generation contrast does not (Table A.11).

Table A.12 collects three further design and plant-operation checks. Panel A tabulates the next-day reservoir-state and hydrological-condition tests discussed above. Panel B shows that adding the predetermined on-site staffing level, the covariate with the largest standardized imbalance (0.33 SD below Control), leaves the Treatment estimate essentially unchanged. Panel C reports several available plant-operation outcomes. We find no statistically precise evidence that Treatment changes hour-to-hour movements in the number of units online, outage or trip incidence, or capacity utilization, and a joint test across the three outcomes does not reject equality with zero. These estimates are secondary and relatively imprecise, however, and they do not capture longer-run equipment wear or maintenance costs.

Across these checks, both the magnitude and statistical significance of the main Treatment effect remain highly stable.

6 Conclusion

We study the effect of autonomous control on firm performance. In a randomized switchback experiment at a hydropower plant, autonomous control reduces deviations from the plant's required generation schedule by about 26% relative to standard manual operation. The improvement is also significantly larger than under an Information-Only regime that improves the delivery of reservoir information while leaving operators responsible for deciding and executing the response.

The high-frequency evidence points to the corrective response as an important source of the Treatment advantage. Conditional on crossing a common threshold, deviations have similar measured severity at onset under Treatment and Control, but their paths subsequently diverge. Across all deviation episodes, autonomous control raises the share brought back below the resolution threshold within ten minutes from about 53% under Control to about 69%. For episodes followed by corrective action, it also reduces response time from deviation onset to the first corrective action by 57%. The gains are concentrated in the largest operating errors. By contrast, Information Only improves the delivery of reservoir information but produces little change in the timing of the first action and a much smaller improvement in how quickly deviations are brought back below the resolution threshold. Taken together, the evidence points to faster initiation of corrective action as an important channel behind the improvement.

These operating gains are economically meaningful. Valuing deviations using the settlement costs faced by the firm, we estimate an annualized gross benefit of approximately $20,600 per plant-year (about $14,700 excluding the imprecisely estimated under-generation component).

More broadly, the experiment shows how autonomous control can affect firm performance by reallocating routine decision authority and accelerating operational response. The mechanism highlighted here should be most relevant in settings where the operating state is observable in real time, routine corrective responses can be codified, and delays in responding are costly. These features characterize parts of process manufacturing, utilities, data centers, logistics, and other asset-intensive operations. In this setting, improving the delivery of one important operating input did not reproduce the full autonomous regime's larger improvement in schedule conformance or its much faster initiation of corrective action. Autonomous systems can take responsibility for routine corrections while workers remain responsible for monitoring, exceptions, and intervention when needed. Whether similar gains arise elsewhere, and how large they are, will depend on how predictable the operating environment is, how costly response delays are, and how readily routine decisions can be delegated. The experiment does not vary staffing as part of the intervention and therefore does not estimate employment effects. Instead, it estimates the consequences of reallocating routine production decisions to an autonomous system while retaining human monitoring, the right to return the plant to manual operation, and responsibility for exceptional cases.

References

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. NBER Working Paper 31422, National Bureau of Economic Research.

Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29.

Angelova, V., Dobbie, W., and Yang, C. S. (2026). Algorithmic recommendations and human discretion. *Review of Economic Studies*, 93(4):2250–2283.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings*, 110:80–84.

Baldea, M., Georgiou, A. T., Gopaluni, B., Mercangöz, M., Pantelides, C. C., Sheth, K., Zavala, V. M., and Georgakis, C. (2025). From automated to autonomous process operations. *Computers & Chemical Engineering*, 196:109064.

Bandiera, O., Barankay, I., and Rasul, I. (2007). Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *Quarterly Journal of Economics*, 122(2):729–773.

Bartel, A., Ichniowski, C., and Shaw, K. (2007). How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. *Quarterly Journal of Economics*, 122(4):1721–1758.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? Evidence from India. *Quarterly Journal of Economics*, 128(1):1–51.

Bojinov, I., Simchi-Levi, D., and Zhao, J. (2022). Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777.

Bolton, P. and Farrell, J. (1990). Decentralization, duplication, and delay. *Journal of Political Economy*, 98(4):803–826.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*, 117(1):339–376.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942.

Dessein, W. (2002). Authority and communication in organizations. *Review of Economic Studies*, 69(4):811–838.

Dessein, W. and Santos, T. (2006). Adaptive organizations. *Journal of Political Economy*, 114(5):956–995.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904.

Graetz, G. and Michaels, G. (2018). Robots at work. *Review of Economics and Statistics*, 100(5):753–768.

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *Quarterly Journal of Economics*, 133(2):765–800.

Hubbard, T. N. (2000). The demand for monitoring technologies: The case of trucking. *Quarterly Journal of Economics*, 115(2):533–560.

Ichniowski, C. and Shaw, K. (2003). Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices. *Journal of Economic Perspectives*, 17(1):155–180.

Kawaguchi, K. (2020). When will workers follow an algorithm? A field experiment with a retail business. *Management Science*, 67(3):1670–1695.

Kesavan, S. and Kushwaha, T. (2020). Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science*, 66(11):5182–5190.

Kesavan, S., Kushwaha, T., and Steele, D. (2025). Profit implications of judgmental adjustments to forecast inputs: Evidence from a large-scale field experiment. *Management Science*, 72(1):119–127.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1):237–293.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Radner, R. (1993). The organization of decentralized information processing. *Econometrica*, 61(5):1109–1146.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Schneider Electric (2026). Global autonomous maturity report. Technical report, Schneider Electric. Commissioned in partnership with Censuswide and Development Economics; survey of 400 senior executives in the energy and chemicals sector across 12 countries.

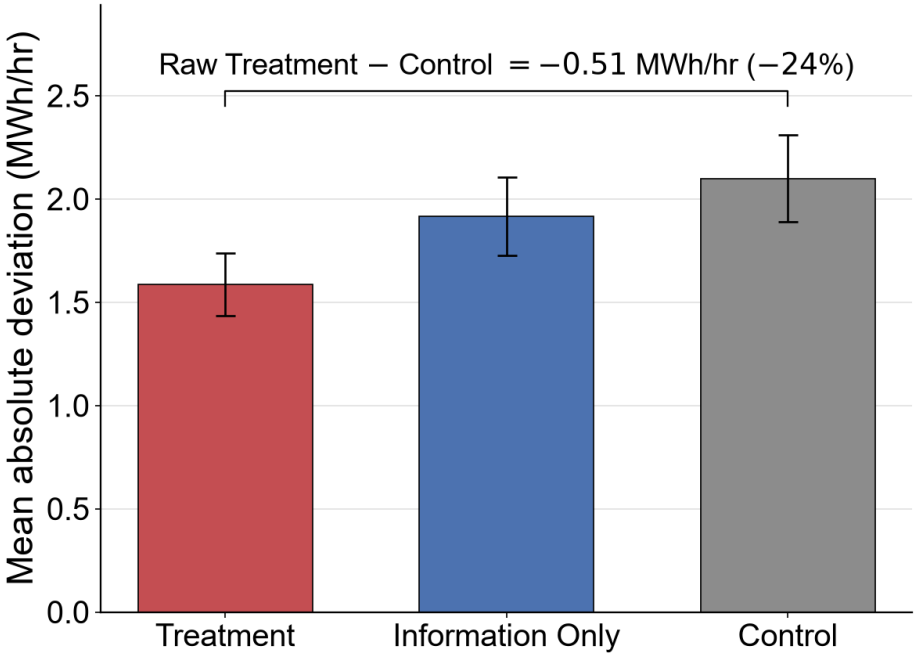
Sun, J., Zhang, D. J., Hu, H., and Van Mieghem, J. A. (2021). Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865.

Tata Consultancy Services (2025). Future-ready manufacturing study 2025. Technical report, Tata Consultancy Services. Study conducted in collaboration with Amazon Web Services; survey of 216 senior manufacturing executives in North America and Europe.

Van Zandt, T. (1999). Real-time decentralized information processing as a model of organizations with boundedly rational agents. *Review of Economic Studies*, 66(3):633–658.

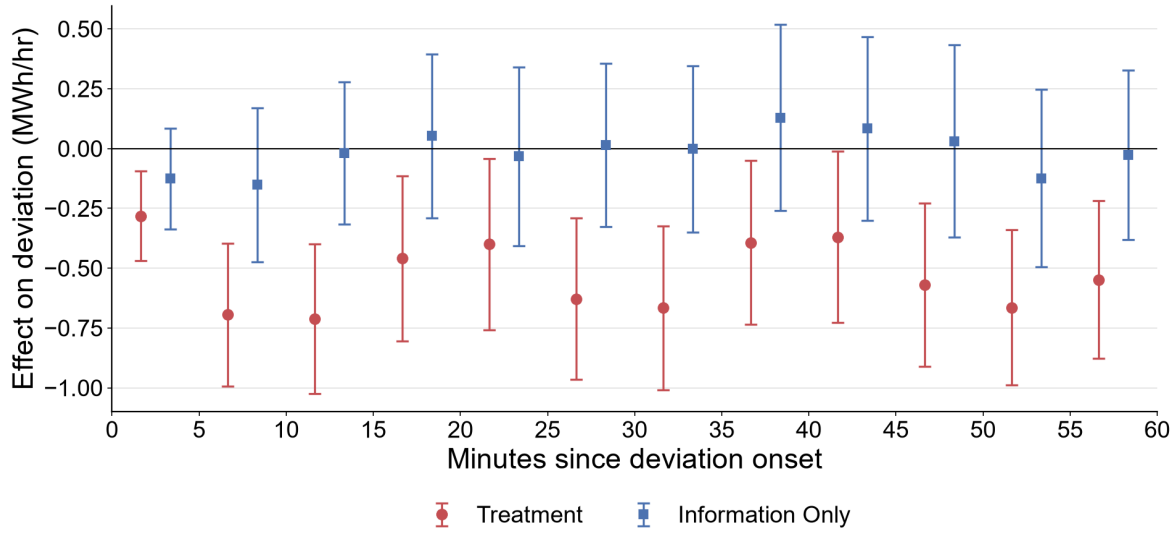
Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557–598.

Figure 1: Autonomous Control Reduces Schedule Deviations



Notes: The figure plots the raw, unadjusted mean absolute deviation (MWh/hr) between realized and required generation in each operating mode, pooled across the full sample (N = 2,088 hourly observations, 696 per arm across 87 experimental days): 2.10 under Control, 1.92 under Information Only, and 1.59 under Treatment. The primary comparison is Treatment versus Control: autonomous control lowers the mean absolute deviation by about 0.51 MWh/hr, roughly a quarter of the Control mean; the covariate-adjusted estimate is −0.54 MWh/hr, about 26% (Table 1). Whiskers are 95% confidence intervals from standard errors clustered at the experimental-day level. For reference, the raw Treatment − Information Only gap is −0.33 MWh/hr (covariate-adjusted −0.330, $p = 0.007$); formal pairwise tests across arms are reported in Table 1.

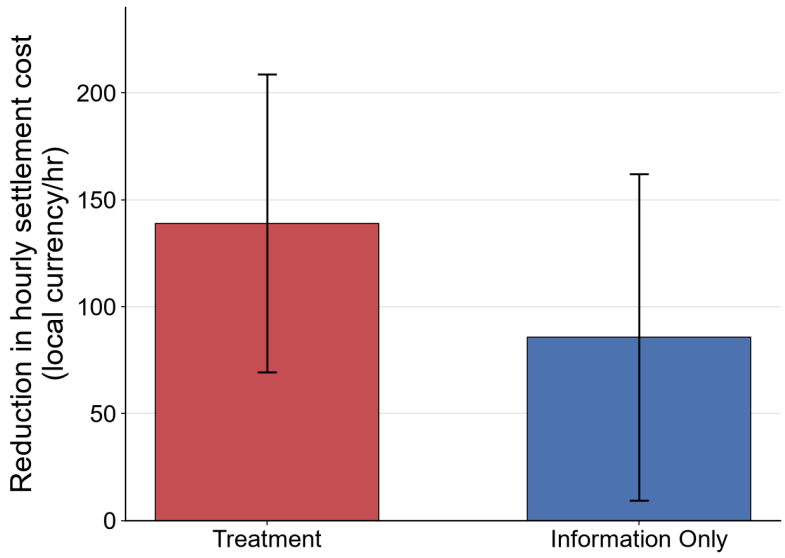
Figure 2: The Treatment Effect Builds After a Deviation Begins



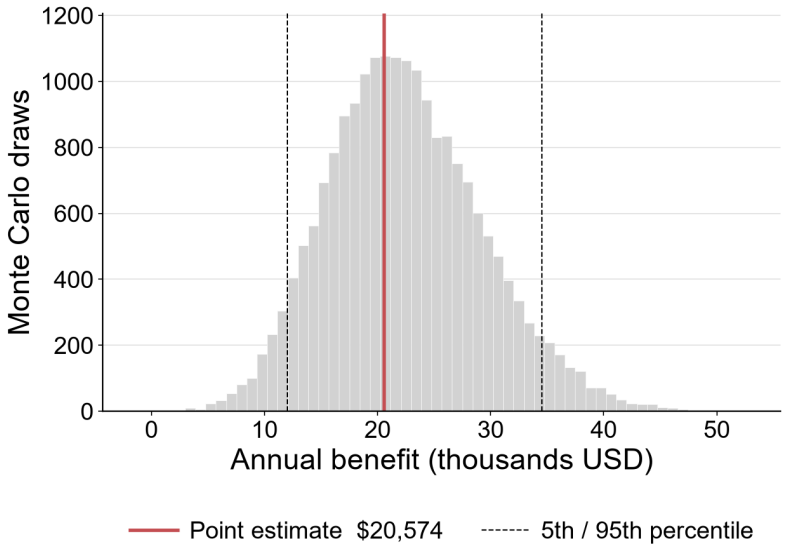
Notes: The figure plots the estimated effect on the generation deviation (MWh/hr) in five-minute bins over the first hour after a deviation begins, for Treatment (red circles) and Information Only (blue squares), each relative to Control, with 95% confidence intervals from standard errors clustered at the experimental-day level (87 clusters); the two arms are offset horizontally within each bin for legibility. The unit of observation is a deviation episode. For each five-minute event-time bin, we average each episode's absolute deviation over the minutes in that bin and regress this episode-level bin mean on Treatment and Information-Only indicators (Control omitted) with day-of-week fixed effects and no additional controls, clustering standard errors by experimental day; the plotted point and its 95% confidence interval are the estimated coefficient and interval for each arm. Minutes since deviation onset is the elapsed time from the moment the absolute deviation first crosses the common onset threshold (an absolute deviation of 3.6 MWh/hr) that defines a deviation episode, constructed as in Appendix B.2. Episodes remain in the sample after they resolve, entering later bins at their realized deviations, so the profile is not driven by differential episode survival across arms. The Treatment effect is smaller in the first five-minute bin (about −0.28 MWh/hr) and grows to about −0.7 MWh/hr within five to fifteen minutes as the corrective response takes hold, remaining negative across the rest of the first hour; the Information-Only effect stays close to zero and is not distinguishable from zero in any bin. Control is the zero baseline. Figure A.4 shows descriptively that the raw Treatment–Control gap is concentrated in active deviation-episode hours.

Figure 3: Autonomous Control Generates a Meaningful Financial Benefit

(A) Estimated reduction in the hourly settlement cost



(B) Illustrative distribution of the annual benefit



Notes: Panel (A) shows the experimentally estimated reduction in the hourly settlement cost of schedule deviations under Treatment and Information Only, each relative to Control, with 95% confidence intervals from standard errors clustered at the experimental-day level; Treatment lowers the hourly cost by about 139 local-currency units, about 28% of the Control mean, and Information Only by about 86 units. *Panel (B)* translates the Treatment effect into an annual benefit in US dollars and reports an illustrative sensitivity distribution that jointly varies that effect, the balancing-price spread, and the exchange rate; the solid line marks the point estimate of about $20,600 and the dashed lines the 5th and 95th percentiles (about $12,000 and $34,500). This range reflects estimation and parameter uncertainty under the stated assumptions and is illustrative, not a statistical confidence interval. The hourly cost is priced using the settlement rules in equation (2) and annualized using the seasonal assumptions in Section 5.3.

Table 1: Effects of Autonomous Control on Schedule Deviations

	(1)	(2)	(3)	(4)
Treatment	−0.548***	−0.542***	−0.541***	−0.537***
	(0.130)	(0.127)	(0.127)	(0.125)
Information Only	−0.192	−0.191	−0.206	−0.207
	(0.144)	(0.145)	(0.135)	(0.138)
Treatment − Information Only	−0.356	−0.351	−0.335	−0.330
Wald p (Treatment = Information Only)	0.005	0.007	0.004	0.007
Effect (% of Control mean)	−26.1	−25.8	−25.8	−25.6
Day-of-week FE	Yes	Yes	Yes	Yes
Hour-of-day FE	Yes	Yes	Yes	Yes
Day-ahead price control	No	Yes	No	Yes
Schedule-difficulty controls	No	No	Yes	Yes
Observations	2,088	2,088	2,088	2,088
Control mean (MWh/hr)	2.099	2.099	2.099	2.099

Notes: The unit of observation is the plant-hour; the outcome is the hourly absolute schedule deviation $D_{h,t}$ (MWh/hr) between realized and required generation, defined in Section 3. The sample is the 2,088 experimental hours (696 per arm). Each column is an OLS regression of the outcome on Treatment and Information-Only indicators, with Control the omitted category; Treatment is autonomous closed-loop operation, Information Only adds an automated water-level update to otherwise-manual operation, and Control is business-as-usual manual operation. All columns include hour-of-day and day-of-week fixed effects. The price control is the predetermined day-ahead price; the schedule-difficulty controls are the mean and largest hour-to-hour change in scheduled generation. Standard errors clustered at the experimental-day level (87 clusters) are in parentheses. Column (4) is the fully saturated specification. The Treatment − Information Only row is the difference in coefficients and the Wald p tests their equality. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

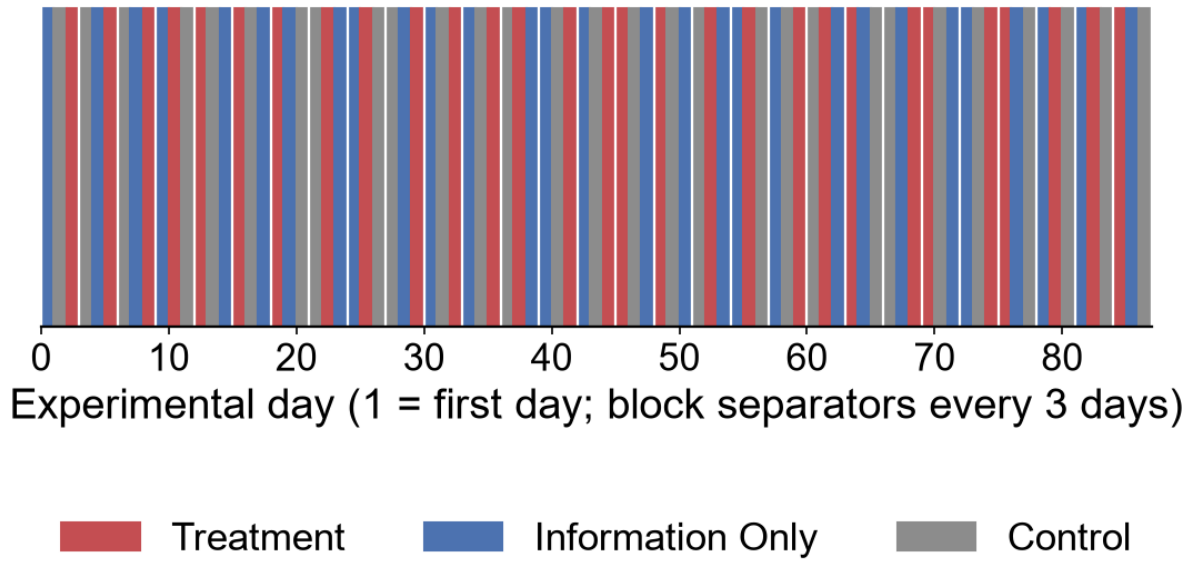
Table 2: Autonomous Control Accelerates the Corrective Response

		Coefficient (SE) vs. Control			
Outcome	Control mean	Treatment	Information Only	T−IO	p (T−IO)
A. Response time (minutes)					
Response time: onset to first action	12.74	-7.264^{***} (0.330)	−0.476 (0.440)	−6.788	< 0.001
B. Remaining adjustment once action begins (minutes)					
First action to return below threshold	1.46	0.008 (0.182)	−0.002 (0.194)	0.010	0.954

Notes: The unit of observation is the experimental day; each row is a day-level outcome for a stage of the corrective response cycle regressed on Treatment and Information-Only indicators (Control omitted) with day-of-week fixed effects on $N = 87$ experimental days; heteroskedasticity-robust standard errors are in parentheses. Each measure is defined for deviation episodes that are followed by a corrective action, and is summarized to one observation per experimental day by taking the median across that day's episodes; the second measure additionally requires the deviation to return below the threshold within the day. "Control mean" is the Control-arm mean of the outcome; the "Treatment" and "Information Only" columns are the estimated coefficients (SE) on those arm indicators; "T−IO" is the Treatment-minus-Information-Only difference in coefficients and "p (T−IO)" is the p-value testing Treatment = Information Only. "Response time" is the elapsed time from deviation onset, defined by the common 3.6 MWh/hr threshold, to the first corrective action. "First action to return below threshold" is the elapsed time from that action until the absolute deviation falls below 1.8 MWh/hr. Corrective actions are timestamped using the same operational recording system and time base in all three arms (Appendix B.2), which also details the construction of these measures. Figure A.5 reports how quickly deviations are brought back below the threshold across all deviation episodes. Appendix B.2 shows that the response-time result remains large under a conservative treatment of episodes that return below the threshold before corrective action begins. Treatment reduces the time to the first corrective action by 7.26 minutes (57% of the Control mean), while Information Only produces only a small, statistically insignificant change; the Treatment−Information Only gap is 6.79 minutes ($p < 0.001$). The time from the first corrective action until the deviation returns below the threshold is essentially unchanged across arms. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

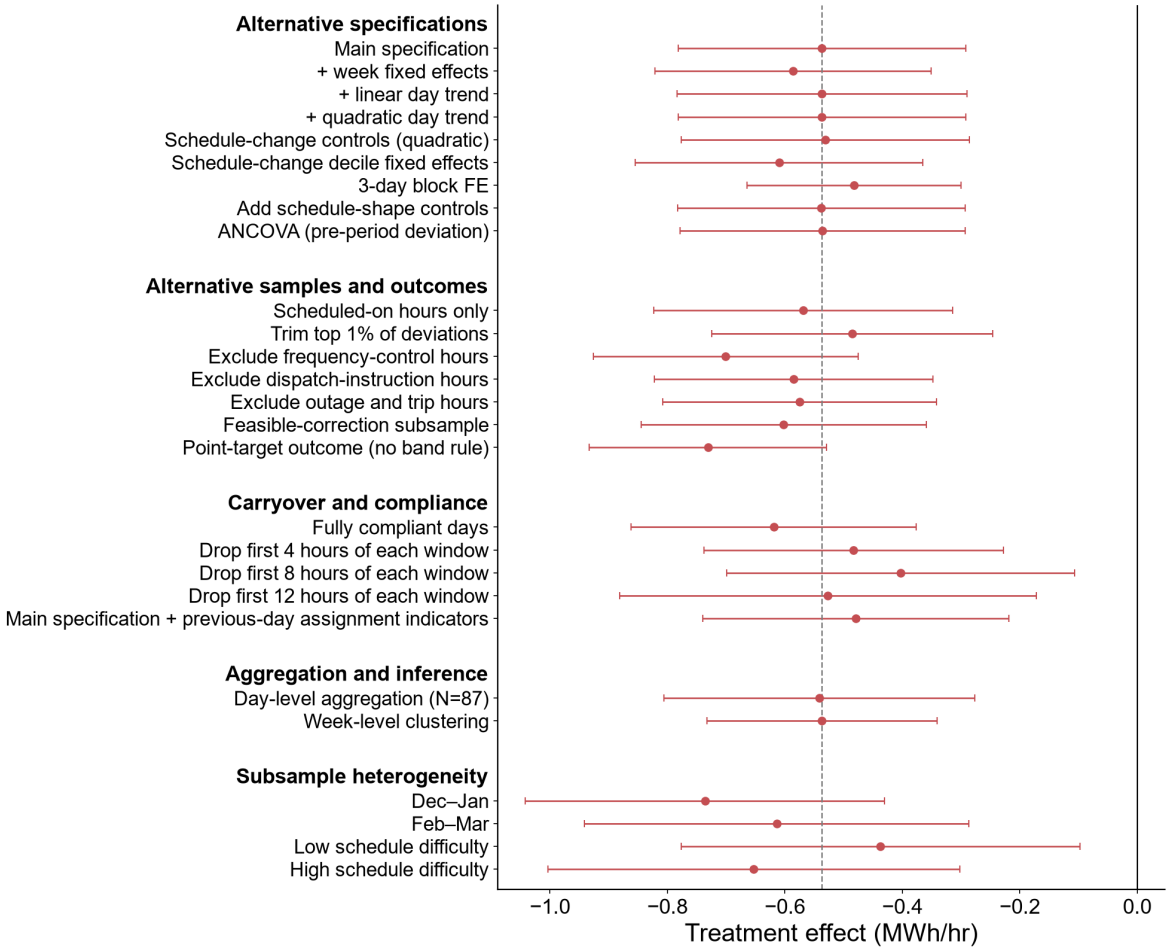
A Appendix: Figures and Tables

Figure A.1: Realized Operating-Mode Assignment Sequence



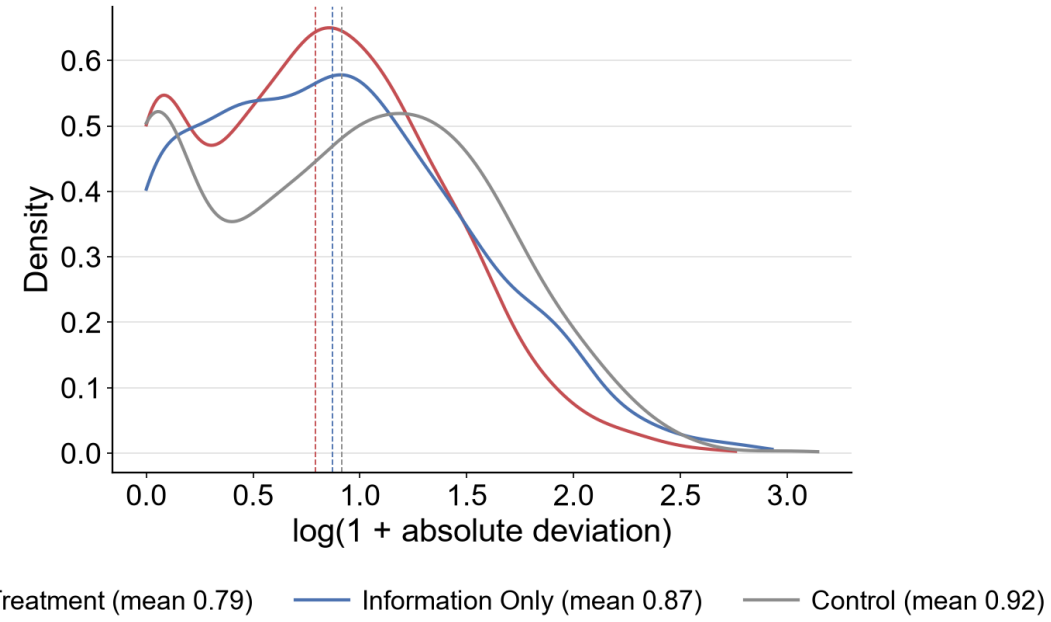
Notes: The figure reports the realized sequence of arm assignments over the 87 experimental days and shows a balanced, drift-free design consistent with the permuted-block protocol. Each tile is one experimental day, numbered 1 through 87 in chronological order and colored by assigned arm. Thin separators mark the three-day blocks of the design. Under the permuted-block protocol every three-day block contains all three arms; the realized sequence satisfies this in all 29 blocks. Assignment is well interleaved across the study window, with no temporal drift: the arm totals are balanced at 29 days each and the longest run of consecutive same-arm days is two. Section 2.3 describes the randomization protocol.

Figure A.2: Stability of the Treatment Effect Across Specifications



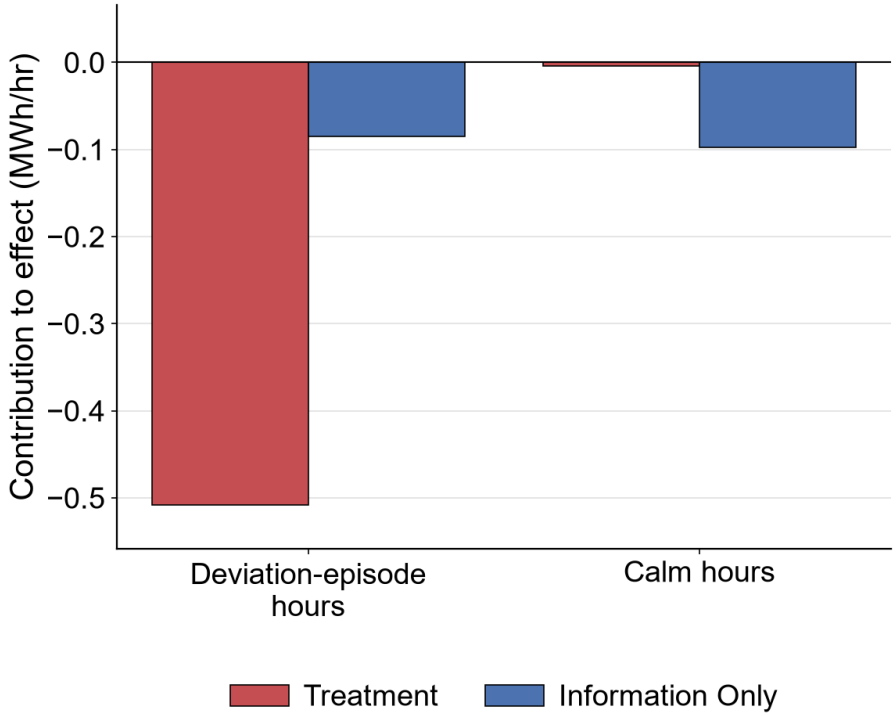
Notes: The figure reports the Treatment coefficient across a wide range of alternative specifications and shows that the effect is negative in every one. Each point is the Treatment coefficient on the absolute deviation (MWh/hr), with its 95% confidence interval, from a separate estimation, and the vertical dashed line marks the fully saturated specification estimate. The specifications, labeled on the vertical axis, start from the main model (equation (1)) and vary one design choice at a time, grouped into alternative specifications, alternative samples and outcomes, carryover and compliance checks, and aggregation and inference. Hourly specifications cluster standard errors at the experimental-day level unless the label states otherwise (the row labeled *Week-level clustering* uses week clustering). The bottom group, *Subsample heterogeneity*, adds the Treatment effect estimated separately within the two sample-period subsamples (December–January and February–March) and the low- and high-schedule-difficulty subsamples, each with experimental-day-clustered standard errors. The carryover-check row *Main specification + previous-day assignment indicators* adds indicators for Treatment and Information-Only assignment on the previous experimental day to the fully saturated specification of equation (1), estimated on the 86 days that have a preceding experimental day; the plotted point is the contemporaneous Treatment coefficient from that regression.

Figure A.3: Treatment Shifts the Deviation Distribution Toward Zero



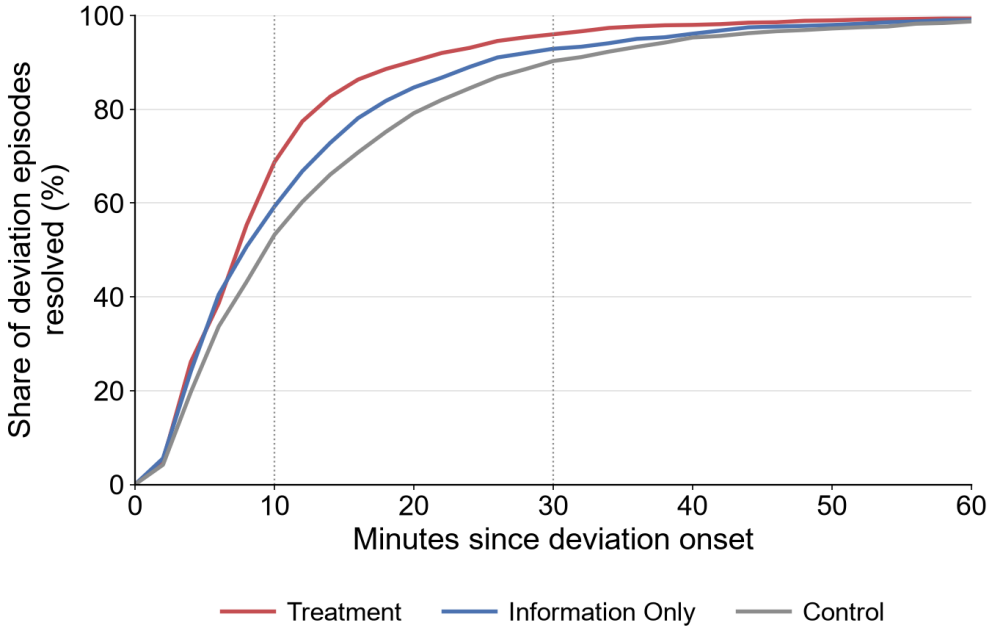
Notes: The figure reports the full distribution of log(1 + absolute deviation) by arm. The Treatment distribution is shifted toward smaller deviations relative to Control and Information Only, with the largest contribution to the mean effect arising in the upper tail. Kernel density estimates plot the log(1 + absolute deviation) by randomized operating mode ($N = 2{,}088$ hourly observations; 696 per arm across 87 experimental days). The outcome plotted is $\log(1 + D_{h,t})$, where $D_{h,t}$ is the hourly absolute schedule deviation defined in Section 3. The log transformation compresses the right tail of the deviation distribution and allows differences across arms to be compared across the full support rather than being dominated by a small number of large misses. Dashed vertical lines indicate the group mean for each arm: 0.79 for Treatment, 0.87 for Information Only, and 0.92 for Control.

Figure A.4: The Treatment–Control Gap Is Concentrated in Deviation-Episode Hours



Notes: The figure descriptively splits the raw Treatment–Control and Information Only–Control mean differences into the portions associated with active deviation-episode hours and calm hours (N = 2,088 hourly observations, 696 per arm across 87 experimental days). Deviation-episode hours are hours that fall within an active deviation episode, defined by the absolute deviation exceeding the same onset threshold (3.6 MWh/hr) in every arm and continuing until it falls below 1.8 MWh/hr; calm hours are the remaining hours, when the plant operates close to target. For Treatment, the two components reconcile to the raw Treatment–Control gap of about −0.51 MWh/hr (the covariate-adjusted estimate is −0.54; Table 1), with essentially all of the raw gap concentrated in deviation-episode hours. Bars plot Treatment (red) and Information Only (blue), each relative to Control.

Figure A.5: Deviations Resolve Faster Under Autonomous Control



Notes: The figure plots the share of deviation episodes brought back below the 1.8 MWh/hr threshold within a given number of minutes after onset, by operating mode, over all deviation episodes ($N = 3{,}726$). Deviation onset occurs when the absolute deviation first crosses the common 3.6 MWh/hr threshold. Within ten minutes of onset, 53.2% of deviations are brought back below the threshold under Control, 59.3% under Information Only, and 68.7% under Treatment; within thirty minutes the shares are 90.3%, 92.9%, and 96.0%. Treatment is significantly faster than Control across the full distribution (log-rank test, design-based randomization-inference $p < 0.001$, permutation at the experimental-day level), as is Information Only to a smaller degree ($p = 0.013$). The curves are Kaplan–Meier functions (one minus the survival function).

Table A.1: Experimental Arms and the Corrective Response Process

Corrective-response stage	**Treatment**	**Information Only**	**Control**
Existing plant information	Integrated continuously by the system	Same as Control	Available through existing operating systems
Additional automated update	No separate operator update; the system already uses broader real-time information	Automated, centralized reservoir water-level update	No
Scope of information integration	System integrates plant, schedule, and grid information continuously	Operators assemble existing information, aided by the added update	Operators assemble information from existing sources
Detects the deviation	System	Operator	Operator
Interprets the operating state	System, for routine cases	Operator	Operator
Determines the correction	System	Operator	Operator
Issues the command	System	Operator	Operator
Reassesses the plant	System, continuously	Operator	Operator
Issues subsequent corrections	System	Operator	Operator
Human oversight and intervention	Monitoring, recall to manual operation, and exceptional cases	Full responsibility	Full responsibility
Physical plant and actuator	Common to all arms	Common to all arms	Common to all arms

Notes: The table sets the three randomized operating modes side by side along the stages of the corrective response process. Control is the plant's standard manual regime. Information Only is manual operation plus an automated, centralized water-level update delivered to the operators; water level is already observable under Control, so the arm changes the *delivery* of one salient variable rather than the execution of operating commands. Treatment activates the closed-loop system, which detects deviations, determines routine corrections, and issues the commands itself, with operators present to monitor it and to handle exceptional cases. The physical actuator and the plant's generation response are common to all three arms. The design supports three pairwise comparisons: Information Only minus Control isolates the effect of the automated reservoir-information update under manual execution; Treatment minus Control is the total effect of the autonomous loop relative to standard manual operation; and Treatment minus Information Only is the incremental effect of the autonomous loop over an information-enhanced manual regime. Because Information Only does not reproduce the full information set the system acts on, the last contrast measures the value of the full loop over that channel, not of execution alone (Section 2.2).

Table A.2: Diagnostic Tests on Outcomes That Should Not Respond to Treatment

Outcome	Treatment	Information Only	p (Treatment)	Controls	N
A. Day-ahead scheduled generation (MWh/hr)	−0.453	−0.047	0.644	No	2,088
	(0.976)	(1.002)			
B. Pre-experiment deviation, 4-week window (MWh/hr)	0.010	0.011	0.814	Yes	2,088
	(0.043)	(0.043)			
C. Pre-experiment deviation, 8-week window (MWh/hr)	0.004	0.004	0.859	Yes	2,088
	(0.024)	(0.024)			
D. Next-day ($t+1$) schedule outcomes (experimental-day level; coefficient (SE))					
Scheduled generation (MWh/hr)	−0.722 (1.072)	−0.953 (1.109)	0.502	No	86
Schedule coefficient of variation	0.155** (0.061)	0.080 (0.064)	0.014	No	86
Mean hour-to-hour change (MWh/hr)	0.468 (1.082)	0.618 (0.912)	0.667	No	86
Schedule-difficulty index	6.529 (7.242)	2.989 (8.005)	0.370	No	86
Number of large hour-to-hour changes	−0.535 (0.597)	0.702 (0.521)	0.373	No	86
Hours revised in the dispatch schedule	−0.093 (0.369)	−0.387 (0.372)	0.802	No	86
Size of the dispatch revision (MWh)	0.571 (0.864)	−0.592 (0.640)	0.511	No	86
Any dispatch revision (indicator)	−0.077 (0.085)	−0.135 (0.096)	0.365	No	86

Notes: The unit of observation is the plant-hour in Panels A–C ($N = 2{,}088$) and the experimental day in Panel D ($N = 86$). Each row regresses an outcome that should not respond to the intervention on the Treatment and Information-Only indicators (Control omitted). Row A regresses the day-ahead scheduled generation, which is finalized at the day-ahead stage, before the assignment is revealed, and cannot respond to the intervention; Rows B and C use the pre-experiment absolute-deviation measures defined in Table A.3, constructed over four- and eight-week windows ending thirteen weeks before each experimental day. Rows B and C include the schedule-difficulty operating controls; Row A does not. Panel D reports eight next-day schedule outcomes, testing whether the scheduling desk changes the following day's schedule after observing the current day's operating mode; each regression holds the day-$(t+1)$ own arm fixed. Standard errors clustered at the experimental-day level are in parentheses (heteroskedasticity-robust for the day-level Panel D), and p (Treatment) is the p-value on the Treatment coefficient. No coefficient in Panels A–C is statistically distinguishable from zero, consistent with balance in predetermined outcomes and with the day-ahead schedule being fixed before treatment assignment. Of these eight outcomes, only the schedule coefficient of variation shows a nominally significant response ($p = 0.014$; randomization-inference $p = 0.010$). Given that eight next-day outcomes are examined, this isolated result should be interpreted cautiously. The other seven coefficients all have $p > 0.35$ on both the cluster and randomization-inference bases. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table A.3: Covariate Balance Across Experimental Arms

	Mean			ANOVA	Std. diff.	
	Treatment	Information Only	Control	p-value	T−C	IO−C
A. Hydrological conditions						
Water level at day start (m)	583.16	583.19	583.16	0.934	−0.013	0.078
Mean daily natural inflow (m^3/s)	25.03	25.26	26.47	0.761	−0.182	−0.153
Reservoir inflow variability	0.173	0.166	0.178	0.844	−0.070	−0.155
Share of high-inflow days	0.517	0.483	0.483	0.957	0.069	0.000
B. Market conditions (set in the day-ahead market)						
Day-ahead electricity price (local currency/MWh)	2,303.0	2,374.1	2,408.2	0.838	−0.154	−0.050
C. Generation schedule (submitted at the day-ahead stage)						
Scheduled generation (MWh/hr)	28.22	28.65	28.98	0.761	−0.196	−0.086
Mean hour-to-hour change in scheduled generation (MWh/hr)	13.16	13.69	13.16	0.802	0.001	0.153
Largest hour-to-hour change in scheduled generation (MWh/hr)	83.26	88.09	81.80	0.643	0.056	0.238
Schedule coefficient of variation	1.054	1.035	1.036	0.934	0.084	−0.002
Schedule variance (MWh^2)	915.90	886.96	956.39	0.785	−0.107	−0.184
D. Pre-experiment performance						
Pre-experiment deviation, 4-week window (MWh/hr)	1.844	1.847	1.832	0.942	0.071	0.086
Pre-experiment deviation, 8-week window (MWh/hr)	1.844	1.845	1.842	0.991	0.027	0.034
E. Staffing (predetermined roster)						
On-site staffing (persons)	3.03	3.45	3.45	0.346	−0.333	0.000
Mean operator age (years)	35.19	35.63	35.34	0.761	−0.066	0.127

Notes: The unit of observation is the experimental day; each row reports the arm-specific mean of a predetermined, day-level covariate, the p-value from a one-way analysis of variance (ANOVA) across the three arms, and the standardized differences of the Treatment (T) and Information Only (IO) arms relative to Control (C) (arm mean minus Control mean, divided by the pooled day-level standard deviation). The sample is $N = 29$ experimental days per arm (87 days total). Predetermined conditions are broadly comparable across arms: the largest standardized difference is 0.33 SD, for on-site staffing (Treatment below Control), and every ANOVA p-value is at least 0.346. Controlling for on-site staffing leaves the main Treatment estimate essentially unchanged (Section 5.4). Pre-experiment absolute deviation (Panel D) is the mean hourly absolute schedule deviation over the four (respectively eight) weeks ending thirteen weeks before each experimental day, computed from plant data for the twelve months preceding the experiment and assigned to every hour of that day; the thirteen-week offset preserves the day of week and keeps the measure strictly pre-experiment.

Table A.4: The Treatment Effect and Day-Level Persistence

	Coefficient (SE)	p	% of control
Panel A. Interaction with day-level persistence (per SD)			
Treatment × persistence	−0.059 (0.095)	0.535	−2.8
Information Only × persistence	0.068 (0.104)	0.518	3.2
Panel B. Effect at mean persistence			
Treatment	-0.526^{***} (0.115)	< 0.001	−25.0
Information Only	−0.187 (0.123)	0.134	−8.9
Panel C. Validation of the persistence measure (Control + Information Only days)			
corr(persistence, time to return below threshold)	0.599		

Notes: The unit of observation is the plant-hour; the outcome is the hourly absolute schedule deviation $D_{h,t}$ (MWh/hr), defined in Section 3, N = 2,088. "Persistence" is a predetermined, arm-invariant day-level index of hydrological persistence (the standardized log of daily inflow variability); its link to within-day deviation persistence is validated on Control and Information-Only days in Panel C, where it is positively correlated with time to return below threshold, defined as the time from onset until the deviation falls below the 1.8 MWh/hr threshold and summarized by the median across deviation episodes within each experimental day ($r = 0.599$; 95% confidence interval $[0.40, 0.74]$ over the 58 Control and Information-Only days, $p < 0.001$). Panel A interacts each arm indicator with persistence (per standard deviation) in the fully saturated specification, and Panel B reports the arm effects evaluated at mean persistence. Coefficients are on the Treatment and Information-Only indicators (Control omitted), with standard errors clustered at the experimental-day level (87 clusters) in parentheses and the % of control column expressing the coefficient relative to the Control mean. We find no evidence that the Treatment effect rises with persistence: the interaction is small and not significant (-0.059, $p = 0.535$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.5: Decomposition of the Treatment Effect by Direction of Deviation

Outcome	Treatment	Information Only	Control mean	% of control	N
A. Over-generation (MWh/hr)	-0.348^{***}	-0.043	1.172	-29.7	2,088
	(0.083)	(0.107)			
B. Under-generation (MWh/hr)	-0.189^{*}	-0.164	0.927	-20.4	2,088
	(0.111)	(0.113)			
Test: Under-gen. − over-gen. (stacked)	0.159				4,176
	(0.151)				

Notes: The unit of observation is the plant-hour and the effect is decomposed by the direction of the hourly schedule deviation, estimated with the fully saturated specification (N = 2,088). Row A is over-generation (realized above schedule) and Row B is under-generation (realized below schedule). In frequency-control hours, the directional components use the same permitted-band adjustment as the primary outcome, so the over- and under-generation components sum to $D_{h,t}$. Coefficients are on the Treatment and Information-Only indicators (Control omitted); "Control mean" is the Control mean and "% of control" the Treatment coefficient relative to it. Standard errors clustered at the experimental-day level (87 clusters) are in parentheses. The point estimate is larger for over-generation (−0.348 MWh/hr, a 29.7% reduction) than for under-generation (−0.189 MWh/hr), but the difference between the two directions is not statistically significant: the final row is a stacked test that interacts the arm indicators, the controls, and the fixed effects with the direction of the deviation, so it equals the difference between the Row B and Row A Treatment coefficients (0.159, SE 0.151, $p = 0.295$), which does not reach significance. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.6: Robustness to Alternative Outcome Specifications

Outcome	Treatment	Information Only	Control mean	N	Effect (% of mean)
A. log(1 + absolute deviation)	-0.133^{***}	-0.048	0.916	2,088	n/a
	(0.048)	(0.049)			
B. Pr(extreme scheduling miss)	-0.080^{***}	-0.024	0.132	2,088	-60.6
	(0.016)	(0.017)			
C. Pr(sustained high-deviation episode)	-0.102^{***}	-0.023	0.177	2,088	-57.6
	(0.026)	(0.029)			

Notes: The unit of observation is the plant-hour. Each row re-estimates the fully saturated specification on a different transformation of the schedule-deviation outcome: the log-transformed absolute deviation, $\log(1 + D_{h,t})$, and two binary tail outcomes. An extreme scheduling miss is an hour whose absolute deviation exceeds the 90th percentile of the pooled deviation distribution; a sustained high-deviation episode is an hour that is part of two or more consecutive hours whose absolute deviation exceeds the 75th percentile of the pooled deviation distribution. Units differ across rows, so the Control mean is reported per row. Each cell is the coefficient on the Treatment or Information-Only indicator (Control omitted), with standard errors clustered at the experimental-day level (87 clusters) in parentheses; $N = 2{,}088$ hourly observations throughout. The Effect column expresses the Treatment coefficient as a percentage of that row's Control mean, except for the log outcome (row A), whose coefficient is already approximately a proportional effect and for which a ratio to the log-scale mean is not meaningful ("n/a"). $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table A.7: The Value of the Schedule-Conformance Effect

		Coefficient (SE) vs. Control			
Outcome	Control mean	Treatment	Information Only	T−IO	p (T−IO)
Settlement cost (local currency/hr)	499.2	-138.985^{***}	-85.862^{**}	−53.123	0.155
		(35.539)	(38.968)		
Cost, over-generation	333.1	-109.759^{***}	−36.745	−73.014	0.029
		(28.972)	(37.615)		
Cost, under-generation	166.1	−29.226	-49.117^{**}	19.891	0.368
		(25.667)	(22.515)		
Additional schedule-deviation penalty	91.6	-35.972^{***}	-23.792^{***}	−12.180	0.073
		(7.577)	(8.392)		

Notes: The unit of observation is the plant-hour; each row is a separate OLS regression under the fully saturated specification ($N = 2{,}088$), pricing the schedule deviation at imbalance-settlement costs (local currency per hour). "Control mean" is the Control-arm mean; the "Treatment" and "Information Only" columns are the estimated coefficients (SE) on those arm indicators, with Control the omitted category and standard errors clustered at the experimental-day level (87 clusters) in parentheses. "T−IO" is the Treatment-minus-Information-Only difference in coefficients and "p (T−IO)" is the p-value testing Treatment = Information Only. Treatment lowers the total settlement cost by 27.8%, a reduction concentrated in the over-generation cost (a 33.0% reduction, with a significant Treatment−Information Only gap, $p = 0.029$), while the under-generation cost is not measurably changed. The additional schedule-deviation penalty falls by 39.3% ($p < 0.001$) and is already included within the directional cost measures rather than added to them separately, so the directional and penalty coefficients are not additive. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table A.8: Sensitivity of the Economic-Value and Break-Even Estimates

Scenario	Annual benefit (USD/plant-year)	Break-even investment (USD/plant)
A. Annualization window (discounting held at 10%, 10 years)		
Non-spill season, 6 months	16,900	103,800
Non-spill season, 8 months (baseline)	20,600	126,400
Non-spill season, 10 months	24,200	149,000
Excluding the under-generation component	14,700	90,300
B. Discounting (annual benefit held at the baseline, about $20,600)		
7% real rate, 15-year life	20,600	187,400
10% real rate, 10-year life (baseline)	20,600	126,400
15% real rate, 5-year life	20,600	69,000

Notes: The non-spill season is the part of the year in which the reservoir sits below its spill threshold, so that water conserved by avoiding over-generation can be stored and used to generate electricity later; during the spring spill season the reservoir may already hold more water than it can store, so conserved water has little additional value (Section 5.3). The table varies the three assumptions that carry the annual-benefit and break-even calculation of Section 5.3, one block at a time, around the baseline (an eight-month non-spill window, a 10% real discount rate, and a ten-year asset life). Panel A varies the length of the non-spill season over which the over-generation benefit accrues, and reports a row that drops the under-generation component entirely, since its underlying monetary coefficient is not individually significant (Table A.7). Panel B varies the discount rate and asset life in paired scenarios, holding the annual benefit at its baseline value (about $20,600); the break-even investment is the annual benefit multiplied by the corresponding annuity factor. The baseline cell (eight months; 10%, ten years) reproduces the headline annual benefit of about $20,600 and a break-even investment of about $126,000. These deterministic scenarios complement the Monte Carlo exercise behind Figure 3. That exercise draws 20,000 realizations, sampling three inputs independently: the Treatment coefficient on the hourly settlement cost from a normal distribution centered at its point estimate (-138.985 local-currency units per hour) with its day-clustered standard error (35.539); the balancing-price spread from a uniform distribution on $[0, 0.10]$, entering as a multiplicative scale on the hourly cost wedge; and the exchange rate from a normal distribution with mean 43.6 and standard deviation 0.9 local-currency units per US dollar. The over- and under-generation components are held at their estimated split rather than sampled separately, and the eight-month non-spill window, the discount rate, and the asset life are held fixed, so their uncertainty is not propagated. The reported 5th and 95th percentiles (about $12,000 and $34,500) are the corresponding percentiles of the simulated annualized gross-benefit distribution, which has a median of about $22,000, slightly above the point estimate because the balancing-price spread is drawn uniformly on $[0, 0.10]$, averaging above the market-average spread at which the point estimate prices deviations, and enters the benefit multiplicatively. These percentiles describe the illustrative sensitivity distribution generated by the stated assumptions and should not be interpreted as a conventional statistical confidence interval. The break-even investment is a deterministic annuity calculation on the point estimate.

Table A.9: Carryover Across Operating Days: Lagged-Assignment Specifications

	(1) Baseline controls, DoW + hour FE	(2) + persistence lags, three-day block FE
Contemporaneous Treatment	-0.360^{**} (0.150)	-0.551^{***} (0.119)
Cumulative Treatment lags	0.736 (0.335)	−0.266 (0.489)
Cluster p	0.031	0.588
Diagnostic RI p	0.227	0.807
Cumulative Information-Only lags	0.244 (0.362)	−0.957 (0.441)
Symmetry p (Treatment = Information Only)	0.272	0.150

Notes: The unit of observation is the plant-hour (N = 2,016 across 84 days once the three daily lags are taken). Each column adds indicators for Treatment and Information-Only assignment on each of the previous three experimental days (lags 1–3) to a specification: column (1) is the raw lag pattern using only the baseline day-of-week and hour-of-day controls, and column (2) accounts for persistent operating conditions and the block design by adding hydrological-persistence lags and three-day randomization-block fixed effects. "Contemporaneous Treatment" is the contemporaneous coefficient (standard error in parentheses); "Cumulative Treatment lags" is the sum of the three lagged Treatment coefficients (clustered standard error in parentheses), with its cluster and diagnostic RI p-values reported in the two rows beneath it. "Cumulative Information-Only lags" is the analogous cumulative sum for Information Only, and the final row is the p-value for equality of the cumulative Treatment and Information-Only lag sums. In column (1) the cumulative lag sum is positive with a cluster p of 0.031 but a diagnostic RI p of 0.227. The diagnostic RI p-values come from block-preserving randomization under the joint sharp null of no contemporaneous or lagged assignment effects. They illustrate the null variation generated by the permuted-block design and should not be interpreted as tests of zero carryover conditional on a non-zero contemporaneous Treatment effect. Carryover-specific inference is therefore based primarily on the lag coefficients in column (2), which leave the contemporaneous Treatment effect unrestricted: there the contemporaneous effect returns to −0.551, close to the main estimate, and the cumulative lag sum is −0.266, statistically indistinguishable from zero (cluster $p = 0.588$). The Information-Only cumulative lags are statistically indistinguishable from the Treatment lags in both specifications (symmetry $p = 0.272$ in column (1) and $p = 0.150$ in column (2)); the clustered standard errors of these Information-Only cumulative sums, computed from the same covariance matrix as the Treatment sums, are shown in parentheses. Stars on the contemporaneous Treatment coefficients denote two-sided significance under experimental-day clustering; inference on the cumulative Treatment lag sums is read from the cluster and diagnostic RI p-values reported beneath them rather than from stars. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10: Alternative Inference Procedures

Procedure	Treatment	Information Only	Treatment − Info. Only
Full-specification coefficient	-0.537^{***}	-0.207	-0.330^{***}
	(0.125)	(0.138)	(0.119)
Clustered standard error, experimental day	0.125	0.138	0.119
Clustered standard error, week	0.100	0.115	0.121
Randomization inference, three-day design blocks, p-value	0.001	0.123	0.005
Unrestricted permutation, p-value	0.001	0.200	0.011

Notes: Each column applies alternative inference procedures to an arm coefficient (or the arm contrast) from the fully saturated specification (the outcome is the hourly absolute schedule deviation in MWh/hr; $N = 2{,}088$ plant-hours). The first two columns report the Treatment and Information-Only coefficients relative to Control; the third reports the direct Treatment−Information Only contrast. The statistical significance of the Treatment effect does not depend on the inference procedure. Standard errors are for the fully saturated specification and are clustered by experimental day unless stated; the week-clustered standard error groups the same hours by calendar week. With approximately 13 week-level clusters, the week-clustered standard error should be read with caution; the day-clustered and randomization-inference results are primary. Randomization inference re-randomizes arm assignment (under the design's three-day blocks and, separately, as an unrestricted permutation) and reports the share of re-randomizations with a coefficient or contrast at least as large in magnitude as the observed one; all randomization-inference and permutation p-values use 2,000 draws. The Treatment−Information Only contrast is significant under both the block-preserving randomization inference ($p = 0.005$) and the unrestricted permutation ($p = 0.011$), consistent with the Wald test ($p = 0.007$), while the Information-Only effect relative to Control is not distinguishable from zero. Stars on the full-specification coefficient denote its two-sided significance under experimental-day clustering. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table A.11: Multiple-Hypothesis Adjustment Across the Secondary-Outcome Family

Outcome	Unadjusted p	Romano–Wolf p
Over-generation	< 0.001	0.001^{***}
Under-generation	0.092	0.108
Price-weighted deviation	< 0.001	0.001^{***}
log(1 + absolute deviation)	0.006	0.008^{***}
Deviation, % of capacity	< 0.001	0.001^{***}
Pr(extreme miss)	< 0.001	0.001^{***}
Pr(sustained episode)	< 0.001	0.001^{***}
Squared deviation	< 0.001	0.001^{***}

Notes: The unit of observation is the plant-hour ($N = 2{,}088$). The secondary-outcome family comprises the eight outcomes listed. Each row reports p-values for the Treatment effect from the fully saturated specification, with experimental-day-clustered standard errors. The over- and under-generation coefficients appear in Table A.5 and the log and tail-outcome coefficients in Table A.6; the remaining family members are summarized here by their p-values only. The two columns report, across the family, the unadjusted p-value and the Romano–Wolf stepdown p-value; significance stars are shown on the adjusted Romano–Wolf column, from which significance is read. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$. Under the adjustment the other seven outcomes remain significant while the weaker under-generation contrast does not survive the Romano–Wolf correction.

Table A.12: Additional Design and Plant-Operation Checks

	Control mean	Treatment	Information Only	Unadj. p (Treat.)	RW p (Treat.)	N
Panel A. Next-day reservoir state and hydrological conditions (day-t assignment on day-(t+1) start-of-day state; day level)						
Reservoir level at start of t+1 (m)	583.16	−0.002 (0.100)	0.038 (0.094)	0.97		86
Natural inflow on t+1 (m^3/s)	26.03	−0.098 (2.333)	−1.985 (1.762)	0.96		86
Inflow variability on t+1	0.175	−0.008 (0.026)	−0.009 (0.025)	0.73		86
Panel B. Staffing robustness (outcome: hourly absolute deviation; plant-hour level)						
Main specification (both controls)	2.099	−0.537*** (0.125)	−0.207 (0.138)	< 0.001		2,088
+ on-site staffing	2.099	−0.515*** (0.128)	−0.209 (0.137)	< 0.001		2,088
+ staffing and operator age	2.099	−0.516*** (0.130)	−0.214 (0.138)	< 0.001		2,088
Panel C. Other plant outcomes (plant-hour level)						
Hour-to-hour change in units online	0.009	0.002 (0.006)	−0.000 (0.005)	0.73	0.92	2,001
Outage or trip active (0/1)	0.046	0.054 (0.051)	−0.003 (0.029)	0.30	0.65	2,088
Capacity utilization (%)	20.68	−0.487 (0.716)	0.005 (0.755)	0.50	0.65	2,088

Notes: The table collects three design and plant-operation checks discussed in Section 5.4. *Panel A* regresses the day-(t+1) start-of-day state on day-t Treatment and Information-Only assignment (Control omitted), holding day-(t+1)'s own arm fixed, at the day level over the 86 day-to-day transitions; the Control mean is the mean over transitions following a Control day. Reservoir level at the start of t+1 is a potentially inherited plant state, whereas natural inflow and inflow variability are exogenous hydrological conditions on the following day. The reported unadjusted p-value is a block-preserving randomization-inference p-value for the Treatment coefficient; no Treatment or Information-Only coefficient is statistically significant. *Panel B* adds predetermined staffing controls to the fully saturated hourly specification (Table 1, column 4), with standard errors clustered by experimental day; "on-site staffing" is the day's on-site staff count and "operator age" the mean operator age. The Treatment coefficient equals −25.6%, −24.5%, and −24.6% of the Control mean across the three rows. *Panel C* regresses each plant outcome on the arm indicators with day-of-week and hour-of-day fixed effects, clustering by experimental day; the hour-to-hour change in the number of units online is measured within the operating day (the first hour of each day is dropped), and the outage-or-trip indicator equals one when a unit outage or trip is active. The "Unadj. p" column reports the unadjusted Treatment p-value; the "RW p" column reports the Romano–Wolf-adjusted Treatment p-value across the three Panel C outcomes and is reported for that family only. A joint test that Treatment has no effect across the three Panel C outcomes does not reject ($F(3, 86) = 0.52$, $p = 0.67$). Standard errors are in parentheses. $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

B Appendix: Additional Institutional Details

Figure B.1: The Hydropower Production Setting

(A) Hydropower generating facility



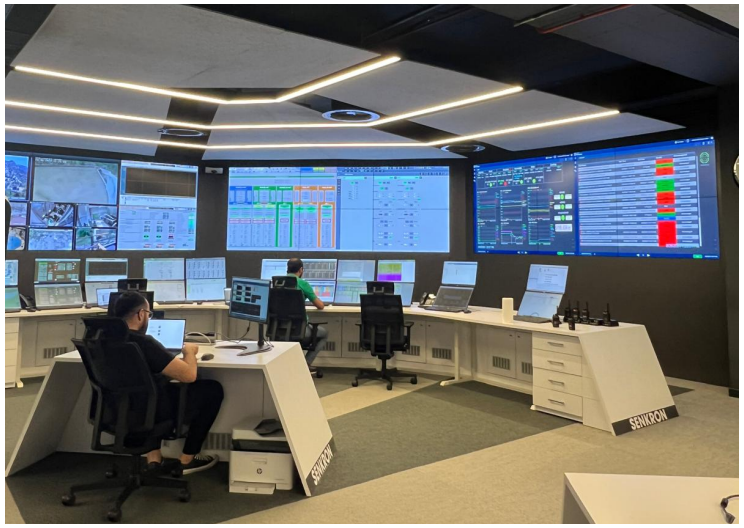
(B) Reservoir and hydropower infrastructure



Notes: Panel (A) shows the hydropower plant used in the experiment, including the generating facility and associated transmission infrastructure. Panel (B) shows the reservoir and water-management infrastructure associated with hydropower generation. The experiment changes the operating regime while leaving the plant's physical production technology unchanged.

Figure B.2: The Operating and Control Environment

(A) Operating control room



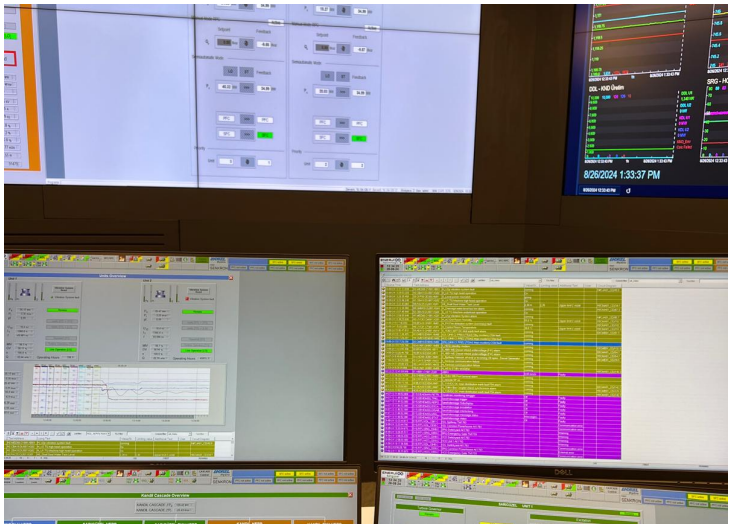
(B) Operators monitoring plant operations



(C) Control-room monitoring environment



(D) Plant monitoring and control interfaces



Notes: The figure provides context on the operating environment in which the experiment took place. Panels (A)–(C) show the control-room setting in which operators monitor electricity generation, plant conditions, and other operating information. Panel (D) provides a closer view of the operational interfaces used to monitor plant states and generating equipment. Under standard manual operation, operators use these systems to assess conditions, determine corrective responses, and enter operating commands. Under autonomous control, routine corrective decisions and their execution are delegated to the autonomous system while operators remain present for monitoring and exceptional cases.

B.1 The autonomous operating technology in detail

This appendix provides additional detail on the autonomous technology described in Section 2.2. Figures B.1 and B.2 provide visual context for the physical production environment and the operating setting in which the experiment takes place. We explain what the system observes, how it determines routine corrections, how those corrections are executed, and how human oversight is maintained. The distinction between Treatment and Information Only is important for interpreting the experiment: Treatment changes the full information-to-action process, whereas Information Only changes the delivery of one operating input while leaving decisions and execution with operators.

The wider technology combines three components: a rule-based closed-loop controller, a separate analytics layer used for monitoring and prediction, and continuous human oversight. The experiment evaluates the closed-loop operating function. The description below is based on the technical documentation and operational information provided for the experiment.

B.1.1 SCADA and closed-loop control

At the base of the system is the plant's supervisory control and data acquisition (SCADA) infrastructure. Sensors and programmable controllers continuously report information on generation, reservoir conditions, equipment status, and other operating variables, while actuators execute commands sent to the generating units.

Under manual operation, operators observe this information through the plant's existing operating systems and enter commands themselves. These commands include starting or stopping a generating unit and changing its output setpoint.

Treatment adds a closed-loop controller on top of this infrastructure. The controller continuously compares realized generation with required generation, determines whether a routine correction is needed, selects the appropriate command, and sends that command directly to the plant. It then observes the updated operating state and can issue further corrections when required.

The key distinction from manual operation is therefore not the physical equipment being controlled, which remains unchanged, but how operating information is converted into decisions and actions. Under Control, an operator must identify the deviation, determine the response, and enter the command. Under Treatment, the system performs this routine sequence directly. Operators remain present throughout and retain the ability to return the plant to manual operation, and the technology automates routine, high-frequency corrections rather than transferring all operational responsibility to software.

The control logic used for this routine operating response is rule-based rather than learning-based. It applies predetermined operating rules to the current state of the plant and therefore does not require a learning or training period during the experiment. In hours in which the plant provides frequency-control services, the permitted operating band is taken into account when performance is measured, as described in Section 3.

B.1.2 Information used by the controller

The controller operates on a continuously updated representation of the plant's operating state. Its inputs include realized generation, reservoir conditions, equipment and unit status, plant availability and operating constraints, the planned generation schedule, and instructions from the grid operator. It uses these inputs to compare realized and required generation and determine whether a routine correction is needed.

The relevant feature for the experiment is that these inputs are integrated continuously rather than requiring an operator to retrieve and combine them before deciding how to respond.

Information Only changes a much narrower part of this process. Operators receive an automated and centralized reservoir water-level update, but they continue to monitor the remaining operating information and remain responsible for deciding and executing any correction. Information Only therefore does not reproduce the full information environment available to the autonomous controller. The interpretation of the Treatment–Information Only comparison follows Section 2.2.

B.1.3 The analytics layer

The wider technology also contains a separate analytics layer used for monitoring, forecasting, and other predictive applications. This layer draws on plant and operating data but is distinct from the closed-loop controller evaluated in the experiment.

The routine operating response studied here is rule-based: the controller observes the current operating state, applies predetermined decision rules, and executes the corresponding command. We therefore do not interpret the Treatment effect as the causal effect of machine learning itself, nor can the experiment separately identify the contribution of analytics applications used elsewhere in the broader technology.

B.1.4 The operating sequence

Under Control and Information Only, the same operating problem remains, but the decision and execution stages are manual. An operator must observe the relevant change, interpret the available information, decide how to respond, and enter the command. Table A.1 summarizes these differences across the three experimental arms.

The ability to reassess conditions and issue further corrections is a feature of the autonomous system by design. We do not measure the complete sequence of later corrective actions in enough detail to identify whether Treatment changes the number, size, or sequencing of subsequent adjustments. We therefore do not use these dimensions as separate mechanisms in the empirical analysis.

B.2 Response-cycle variable definitions and cross-arm comparability

Table 2 combines operational records with high-frequency generation data to measure different stages of the corrective response. Corrective-action timestamps in all three arms come from the same operational command log, with a common schema, common time base, and one-second precision, and with machine-issued and manually entered actions distinguished by a recorded

source field. The comparison of response times therefore does not reflect differences in how actions are recorded across operating modes. Alerts, by contrast, are not defined identically across arms: the autonomous system generates and timestamps its own alerts under Treatment, whereas alerts under manual operation depend on operator entry. We therefore base the main mechanism results on events whose starting point is defined in the same way across all three arms.

A deviation episode begins when the absolute deviation first exceeds 3.6 MWh/hr and ends when it subsequently falls below 1.8 MWh/hr. The onset threshold is the plant's own alert threshold, the deviation magnitude that raises an operating alert, and the resolution threshold is set at half the onset; these are arm-neutral analysis conventions rather than settlement or engineering tolerances. The deviation series used to reconstruct episodes and to build the event-time profile in Figure 2 is sampled at one-minute resolution, with each minute recording realized minus required generation in megawatts; the onset and resolution thresholds are stated in MWh/hr to match the hourly primary outcome, using the identity that one MWh per hourly settlement interval equals an average output of one megawatt over the hour. The same thresholds are applied mechanically under Control, Information Only, and Treatment, providing a common definition of deviation onset and return below the threshold across arms. Deviation episodes are constructed within experimental-day boundaries and do not carry across the 08:00 assignment boundary.

Response time is the elapsed time from the common 3.6 MWh/hr deviation-onset threshold to the first corrective action. Because the starting event is defined identically in all three arms, this is our primary measure of how quickly corrective action begins. The Control mean of the day-level response-time measure reported in Table 2 is 12.74 minutes; Treatment reduces this measure by approximately 7.26 minutes, or 57%, while Information Only produces only a small and statistically insignificant change. Under Control there are approximately 42 deviation episodes per day. Some of these return below the 1.8 MWh/hr threshold before a corrective action occurs, and in those cases there is no response time to measure; response time is therefore defined for the subset of episodes that are followed by a corrective action.

To verify that this feature does not account for the estimated difference in response time, we conduct a conservative exercise that assigns these episodes a hypothetical response time equal to the time at which the deviation returns below the threshold. The share of episodes followed by a corrective action is about 13 percentage points higher under Treatment than Control (randomization-inference $p < 0.001$), while the number of episodes per day is not significantly different across arms (randomization-inference $p = 0.13$). Under this exercise, the Treatment-minus-Control effect remains about -5.5 minutes (randomization-inference $p < 0.001$). The response-time result therefore remains large under a conservative treatment of episodes that return below the threshold before any corrective action occurs.

The episode-based results also do not depend on the 3.6 and 1.8 MWh/hr onset and resolution thresholds. Reconstructing episodes over onset thresholds from 2.4 to 4.8 MWh/hr, with the resolution threshold at half the onset, the Treatment-minus-Control gap in the share of deviations brought back below the threshold within ten minutes ranges from 12.2 to 16.2 percentage points (randomization-inference $p < 0.001$ at every threshold), and the raw episode-level Treatment-

minus-Control gap in response time ranges from −5.3 to −6.4 minutes. This range is a raw across-episode mean difference on the reconstructed-episode sample rather than the day-level, covariate- and fixed-effect-adjusted estimate reported in Table 2; at the baseline 3.6 MWh/hr threshold it equals −6.4 minutes, consistent with, though a different estimand from, the −7.26-minute headline.

Share of deviations brought back below the threshold within ten minutes measures whether the absolute deviation falls below 1.8 MWh/hr within ten minutes of onset. It is constructed for all deviation episodes and provides a complementary measure of how quickly deviations are resolved. We construct the corresponding *share within thirty minutes* in the same way using a thirty-minute horizon. Both shares, computed over all deviation episodes, are reported in Figure A.5.

The second timing measure is the elapsed time from the first corrective action until the deviation returns below the 1.8 MWh/hr threshold. This measure asks a different question: once corrective action has begun, does the remaining adjustment occur more quickly? By construction, this interval is defined only once a corrective action has begun. Table 2 shows essentially no difference across operating modes. The timing advantage of Treatment therefore appears primarily before the first corrective action begins.

For estimation, response time is summarized to one value per experimental day by taking the median across that day's deviation episodes followed by corrective action. The first-action-to-return measure is constructed analogously among episodes followed by corrective action that also return below the threshold within the experimental day. These day-level measures match the level of randomization used in Table 2; the shares brought back below the threshold are summarized over all deviation episodes in Figure A.5.

The headline mechanism results in Section 5.2 therefore rest on two complementary measures with a common, arm-neutral starting point. Treatment substantially reduces the response time from deviation onset to the first corrective action, and it increases the share of deviations brought back below the threshold within ten minutes. Together with the event-time evidence in Figure 2, these results are consistent with an important part of the Treatment advantage emerging early in the corrective response.