Conservation Biology

Exploring mechanisms and impacts of an incentive-based conservation program with evidence from a Randomized Control Trial

Journal:	*Conservation Biology*
Manuscript ID	19-836.R1
Wiley - Manuscript type:	Registered Report
Keywords:	pre-analysis plan, pre registration, randomised control trial, impact evaluation, theory of change, payment for ecosystem services, payment for environmental services
Abstract:	Conservation science needs more high-quality impact evaluations, especially ones which explore the mechanisms for success or failure. Randomized Control Trials (RCTs) provide particularly robust evidence of the effectiveness of interventions (although they have been criticized as reductionist and unable to provide insights into mechanisms) but there have been very few such experiments investigating conservation at the landscape scale. We explored the impact of Watershared, an incentive-based conservation program in the Bolivian Andes, using one of the few RCTs of landscape-scale conservation in existence. There is strong interest in such incentive-based conservation approaches as some argue that they can avoid the negative social impacts sometimes associated with protected areas. We focused on social and environmental outcomes, using responses from a household survey in 129 communities randomly allocated to control or treatment. We controlled for incomplete program uptake by combining standard RCT analysis with matching methods, and investigated mechanisms by exploring intermediate and ultimate outcomes according to the underlying theory of change. Previous analyses, focusing on single biophysical outcomes, showed that over its first five years Watershared did not slow deforestation or improve water quality at the landscape scale. Here we show that it has influenced some intermediate outcomes (including targeted production systems and perceptions of the condition of forest), while having no impact or unexpected impact on other outcomes. By publishing this study as a Registered Report we bring an unusual degree of transparency in conservation research. We suggest that pre-registration of analysis, ideally combined with peer review, is particularly beneficial with complex analyses involving multiple outcomes as it avoids the temptation for cherry picking and reduces publication bias against negative results. This paper also demonstrates how Randomized Control Trials can give insights into the pathways of impact, as well as whether an intervention has impact.

Emma Wiik1, Julia P G Jones1*, Edwin Pynegar1,2, Patrick Bottazzi1,3, Nigel Asquith2, James Gibbons1, Andreas Kontoleon4

1: School of Natural Sciences, Deniol Road, Bangor University, LL57 2UW, UK
2: Natura Foundation Bolivia, Santa Cruz de la Sierra, Bolivia
3: Institute of Geography, University of Bern, Switzerland
4: Department of Land Economy, University of Cambridge, UK
***corresponding author Julia.jones@bangor.ac.uk**

Registered Report (stage 2)

Exploring mechanisms and impacts of an incentive-based conservation program with evidence from a Randomized Control Trial

Abstract

Conservation science needs more high-quality impact evaluations, especially ones which explore the mechanisms for success or failure. Randomized Control Trials (RCTs) provide particularly robust evidence of the effectiveness of interventions (although they have been criticized as reductionist and unable to provide insights into mechanisms) but there have been very few such experiments investigating conservation at the landscape scale. We explored the impact of Watershared, an incentive-based conservation program in the Bolivian Andes, using one of the few RCTs of landscape-scale conservation in existence. There is strong interest in such incentive-based conservation approaches as some argue that they can avoid the negative social impacts sometimes associated with protected areas. We focused on social and environmental outcomes, using responses from a household survey in 129 communities randomly allocated to control or treatment. We controlled for incomplete program uptake by combining standard RCT analysis with matching methods, and investigated mechanisms by exploring intermediate and ultimate outcomes according to the underlying theory of change. Previous analyses, focusing on single biophysical outcomes, showed that over its first five years Watershared did not slow deforestation or improve water quality at the landscape scale. Here we show that it has influenced some intermediate outcomes (including targeted production systems and perceptions of the condition of forest), while having no impact or unexpected impact on other outcomes. By publishing this study as a Registered Report we bring an unusual degree of transparency in conservation research. We suggest that pre-registration of analysis, ideally combined with peer review, is particularly beneficial with complex analyses involving multiple outcomes as it avoids the temptation for cherry picking and reduces publication bias against negative results. This paper also demonstrates how Randomized Control Trials can give insights into the pathways of impact, as well as whether an intervention has impact.

Introduction

There is considerable interest in using positive incentives to encourage sustainable land management, conserve forests, and protect biodiversity. Those promoting these incentive-based conservation approaches, which include payments for ecosystem services (PES) (Jack et al. 2008), suggest they can both effectively deliver environmental outcomes and result in better social outcomes than strict protected areas (Sims & Alix-Garcia 2017). Synthesis of the existing evidence base suggests PES-type interventions have, if anything, only a modest impact on environmental outcomes, and impacts on social outcomes are even more uncertain (Liu & Kontoleon, 2018; Samii et al. 2014). More and better quality evaluations are needed, especially those which can cast light on the mechanisms by which outcomes are, or are not, delivered (Miteva et al. 2012; Börner et al. 2016, 2017).

Randomized control trials (RCT) randomly allocate experimental units to treatment and control groups and are therefore often considered to provide particularly robust evidence of the effectiveness of interventions (Ferraro 2009). However, in the context of conservation policies, RCTs are rare (Pynegar et al. 2019). To our knowledge there have been two RCTs of incentive-based conservation interventions implemented at scale. Jayachandran et al. (2017) showed that carbon payments slowed deforestation rates in Uganda. The RCT in Bolivia of the Watershared intervention (Bottazzi et al. 2018) has been used to evaluate the impact of incentivizing farmers to keep cattle out of riparian forest and reduce deforestation on water quality (Pynegar et al. 2018), deforestation rates (Wiik et al. 2019), and environmental values (Grillos et al. 2019). A third landscape-scale RCT in conservation explored the impact of unconditional livelihood payments on deforestation rates in Sierra Leone (Wilebore et al. 2019).

Evaluation of such socioecological interventions is inherently complex because whether or not the incentives and associated social processes will produce the desired land-use change is uncertain and, even if achieved, these land use changes may (or may not) result in the desired social and environmental ultimate outcome. Impacts may also differ between strata of society (Daw et al., 2016) and take time to materialize. There is interest in other disciplines, such as public health, in bringing lessons from qualitative impact evaluation into Randomized Control Trial analysis (Bonell et al. 2012). In qualitative impact evaluation, the focus is on building and validating a theory of change (which identifies the mechanisms by which the intervention delivers intermediate and ultimate outcomes of interest; White 2009) rather than a narrow focus on ultimate outcomes. The existing published papers which use an RCT to evaluate the impact of landscape-scale conservation interventions mostly report ultimate environmental outcomes of the

intervention (e.g. deforestation rates) but say little about social outcomes (and how these might differ between different groups), and the causal linkages between the intervention and intermediate and ultimate outcomes.

The Bolivian organization Natura Bolivia began to develop the incentive-based conservation program now known as Watershared in 2003 (Asquith & Vargas 2007). Watershared aims to establish a reciprocal relationship between environmental service users (municipal governments, and water cooperatives) and services providers (upstream farmers and cattle-owners) by using in-kind incentives for forest protection and exclusion of cattle from riparian forest to protect biodiversity and improve downstream water quality (Bottazzi et al. 2018). As of 2016, Watershared had 210,000 hectares (4500 households) under conservation agreements (Asquith 2016). We use the Watershared Randomized Control Trial as a rare opportunity to fully analyze the impacts of an incentive-based conservation program. The RCT includes 129 communities randomly allocated to treatment (households were offered Watershared agreements) or control (households were not offered agreements). Using a large household survey conducted at baseline (in 2010) and endline (in 2015/16) we explore both intermediate outcomes (e.g. perceived importance of forest, livelihood changes such as cattle exclusion from riparian forest) and indicators of ultimate outcomes (e.g. perceived forest condition, incidence and frequency of diarrhea). We use the theory of change underpinning the intervention to structure the evaluation. This paper is submitted as a registered report (Parker et al. 2019).

Methods

Watershared Randomized Control Trial

In 2010, Natura decided to roll out Watershared in a new protected area (Area Natural de Manejo Integrado Río Grande y Valles Cruceños) as a randomized control trial (Fig. 1) to evaluate the impact of the intervention on deforestation rates, the quality and quantity of water available for local communities, environmental values, and local livelihoods. They selected the 129 communities in the five main municipalities overlapping the protected area, and these were randomly allocated to control (conservation agreements not offered) or treatment (agreements offered) subsequent to blocking by municipality, community size, and cattle numbers. Consent to conduct the trial was granted by municipal mayors on the understanding that the program would subsequently be implemented in all communities. The experiment was not blinded because participants unavoidably knew whether they belonged to a

control or treatment community. In 2016 the experiment ended, and agreements were offered in control communities.

The Watershared intervention operates through combining incentives with environmental education; a key feature of the intervention is promoting the message that watershed protection is in everyone's interest (Bottazzi et al 2018). Natura gave an environmental education presentation to all treatment and control communities prior to recruiting treatment participants, so the randomization primarily tested the effect of the incentives. Reinforcement of the education messages would have occurred more strongly in treatment communities, where there were multiple visits to offer and monitor the conservation agreements from 2011 to 2015.

Watershared agreements

There were three levels of Watershared agreements. Level 1 and 2 agreements applied to forested land within 100m of a stream or waterway while Level 3 agreements applied to any forested land (details in Bottazzi et al. [2018]). In all three levels, land clearance or timber extraction were not permitted. In addition, cattle had to be excluded from land under level 1 agreements (while level 2 required working towards removing cattle). The value of incentive packages ranged from the equivalent of US$1/ha/year to US$10/ha/year, and farmers with level 1 agreements received an additional 100 US$ worth of in-kind incentives at signing. Transportation costs of the materials to communities were covered by the program. Agreements were for an initial 3 years, were renewable, and were offered in treatment communities twice per year. Program technicians monitored level 1 and level 2 land annually by walking transects across the parcels to verify compliance; level 3 agreements were monitored using remote sensing (forest cover). Where blatant noncompliance was detected, the materials that farmers had been given were removed and redistributed to the community. As with many incentive-based conservation interventions, not all conservation funded with Watershared agreements is additional (i.e. some would have happened anyway in the absence of the scheme; a common issue with PES-type programs [Ezzine-de-Blas et al. 2016]). Bottazzi et al. (2018) estimated that a maximum of 30% of agreements to exclude cattle and 14% to avoid deforestation appear to be additional.

Watershared Household survey

The household survey was a structured questionnaire with > 100 questions (Bottazzi et al. 2017). The baseline household survey was carried out by Natura in 2010 and the endline survey by Natura and Bangor University staff

between October 2015 and June 2016. The aim was to deliver the baseline household survey to all households in all communities (Bottazzi et al. 2017). This was mostly achieved; while the baseline reached 2623 households, only 57 previously unsurveyed households were found in the endline survey (Supporting Information 1). However, the endline was incomplete (S1) because 8 communities did not have any households with data from both baseline and endline surveys; these were excluded from the analysis (Fig. 1a).

Out of all households surveyed in both baseline and endline in treatment communities (n = 970), 456 households took up *Watershared* agreements and 514 did not (i.e. 47% uptake). The allocation to control and treatment was not perfect as 32 out of 702 households in control communities had agreements (Fig. 1b); however, 28 of these agreements were in land owned in treatment communities. Uptake percentages varied across communities from 0% to 100%, which in part reflects high percent uptake in a few small communities (Fig. 1b). Uptake of the program was influenced by barriers to entry (Grillos 2017), individual motivations (Bottazzi et al. 2018), and whether or not households were available to attend the meetings in which the program was presented (Wiik et al. 2019).

The consent form used in both baseline and endline surveys is archived alongside the data (Bottazzi et al. 2017). The endline survey was assessed under the Bangor University Research Ethics Framework. Natura were involved in the research (and paid the enumerators), which is a potential conflict of interest because they are also the implementers of the Watershared program that this work evaluates. However, the independent Bangor University team trained the enumerators, designed the survey, managed and cleaned the data, and conducted the analysis.

Selection of outcome variables

There are a large number of potential outcome variables from the survey which could be explored. We systematically selected outcome variables for analysis based on there being a clear hypothesized mechanism linking to Watershared objectives (S2), based on the program's underlying theory of change (see Fig. 2). The outcome variables selected include intermediate outcomes seen to contribute to the attainment of Watershared ultimate outcomes (e.g. number of water intakes protected from cattle, perception that forest delivers benefits) and self-reported indicators of the ultimate Watershared outcomes (e.g. diarrheal disease in children, perception in forest condition). In total, we identified 11 main outcome variables of interest (some of which have more than one indicator) (Fig. 2).

One outcome had already been evaluated. Pynegar (2018) found no effect of the intervention as implemented on diarrheal incidence and frequency. We accept this finding and do not reanalyze. However, we conduct a secondary analysis exploring the impact of the intervention on the subset of households that report having an individual water intake, which households plausibly have more control over.

Four outcomes were analyzed for a subset of households (Fig. 2) rather than the full dataset. Diarrhea frequency and incidence among children was only analyzed for households that have their own drinking water intake and have children. Hectares of irrigated land was only analyzed for those who reported having access to irrigation in both baseline and endline surveys. The extent and method by which water intakes are protected was only analyzed for those who reported protecting their water intake in the endline survey (there was not a baseline question on this variable). Hectares of improved grazing land was only analyzed for those who said they owned cattle in baseline and endline surveys.

Data analyses

Our data analyses focused on testing, within the theory-of-change framework, the individual hypotheses within the 11 outcome categories of survey questions (Fig. 2). Within each category, we identified one main analysis where we would expect to see a change driven by the intervention if successful, and in some cases, also secondary analyses where changes may either be premature to detect, or indicative more of a detail within a process than an overarching mechanism or success of Watershared (Fig. 2). In our analyses, we followed a hierarchy by which the main analysis within a category was given most weight in evaluating the program (e.g. whether intakes are protected more or less in treatment communities, is more important than when intakes were protected). The results from all analyses were evaluated against the theory-of-change logic. Where results conflict with this logic, we evaluated the strength of evidence based on robustness checks. If results were robust, this casts doubt on the theory of change.

We tested our hypotheses using two analytical approaches; one estimated the average treatment effect (Glennerster & Takavarasha, 2013) of the program as it was rolled out, and the other estimated the program effect specifically on those who participated (Fig. 3). The first, As-Randomized analysis, compared outcomes in all households in treatment communities with all households in control communities, regardless of uptake. The second, As-Treated analysis, compared outcomes in Treated households (households in treatment communities who participated, regardless of which incentives they selected [Supporting Information]) with statistically matched

Control households (matched Control = households in control communities likely to have participated, had they had the opportunity, excluding those who signed agreements). The distinction between the As-Randomized and As-Treated analyses is important due to the incomplete uptake of Watershared. For example, overall impacts of interventions may be low not because the intervention lacks efficacy but instead because of low levels of uptake or poor implementation and compliance (Glennerster & Takavarasha 2013).

Some of our perception-based variables represent an observation of community-wide change and so blur the distinction between As-Randomized and As-Treated analyses (e.g., a perception of how the community is managing its forest can be the same for a participating and non-participating household). In these cases, the difference between the As-Randomized and As-Treated analyses tests whether participation in the program changes how a person perceives their environment.

Before stage 1 review of this registered report we completed three phases of data preprocessing (Fig. 3). Phase I involved choosing variables for use in matching (in the As-Treated analyses) and for use as control variables in the final outcome regressions (both analyses). We selected variables that we hypothesized to influence both uptake and outcomes of the program. Candidate variables were considered based on previous work exploring the uptake of the Watershared intervention (Grillos 2017; Bottazzi et al. 2018; Wiik et al. 2019). We avoided variables with a lot of missing data (S4). The final set included variables capturing community cohesion, wealth, education, and predisposing environmental attitudes (Table 1). Baseline data for an outcome, where available, were used as control variables in outcome regressions as per some difference-in-differences analyses, but not as matching variables to avoid regression to the mean (Daw & Hatfield 2018) (S4). In phase II (S5), we developed propensity score models, based on the variables selected, to predict selection bias for households in control communities based on modeled participation in the program among households in treatment communities. In phase III (S6), we used the selected variables and the propensity scores (a primary and secondary version) to match Treated households with the best available counterfactuals from the control households through a genetic matching algorithm. We used the R packages Matching (Sekhon 2011) to perform the matching, cobalt (Greifer 2019) to evaluate balance visually, mgcv for regressions (Wood 2011, 2017), and ggplot2 for plots (Wickham 2016).

The final two phases (outcome regressions; phase IV, and robustness checks; phase V) were carried out after stage I review was complete. Since we tested many outcomes, there was an increased probability of encountering at least

one false positive (finding a significant impact on an outcome when there is, in fact, none). We therefore applied the Benjamini Hochberg (1995) method to control the false detection rate (FDR) at a level of 0.05, ranking p values based on the p value from the primary analysis within outcome categories (Fig. 2; also see S8 for full description of the multiple testing procedures). These methods were reviewed as part of our stage 1 plan.

The regressions used for hypothesis testing (as opposed to robustness checks) were those that included the primary propensity score (S5, S6) and, in the case of matched analyses, the regressions run on the least restrictive caliper while still attaining adequate balance (a caliper limits the difference between any one pair of observations to within a given standard deviation, meaning that Treated observations deemed too different from any one Control were discarded) (S5). Regressions including the secondary propensity score and additional matching outputs were used as robustness checks (S6, S7) as per recommended best practice (Ho et al. 2007). For example, we would not expect robust results to be changed by using a slightly different set of Control observations, or a subset of Treated households (where a caliper results in losing Treated households).

In the As-Randomized analyses (Phase IV), the outcome was regressed on the experimental group (control or treatment) plus control variables, including the baseline data for an outcome where available. Our control variables included those used in matching to control for non-independence of observations (Wan 2019), add precision to our effect estimate, correct for remaining biases (Ho et al. 2007; Hill 2008; Streiner 2015), and allow evaluation of heterogeneous treatment effects based on variable interactions (Ferraro & Hanauer 2014). All regressions were undertaken using generalized additive models (GAMs) (Wood 2017); families were fitted to the response expectation (e.g. the binomial family with a logit link for binary outcomes; S8).

As-Treated regressions were similar except for being undertaken on the matched Control and Treated subsets of data as per the matching protocol. The protocol resulted in four possible datasets: the combinations of 1) matching with and without a caliper; and 2) matching with two versions of the propensity score (S6). For the outcomes that were analyzed with the full data set (Fig. 2), we tested all four datasets in four regressions. For the outcomes only appropriate to explore with a smaller subset of data (e.g. those who own cattle, or have children, Fig. 2), we ran only two regressions because applying a caliper resulted in losing too many Treated observations (S6).

To explore the extent to which the intervention may benefit socio-economic groups differently and our expectation that some outcomes may be more feasible to achieve for some households than others, we explored a number of outcome interactions based on education and wealth indicators (Table 1). We also included an interaction between perception of water quantity or quality and the experimental group (control or treatment or matched Control or Treated) to examine whether the program had different impacts on those who are more influenced by these issues (Table 1).

Deviations from pre-registration

We opted to undertake all outcome analysis using truncated values of highly skewed predictors, contrary to what was stated in the Supplementary Information of Stage 1. This was because we felt this added an unnecessary complication (testing whether outliers were biasing our estimates for each of 98 individual models).

Results

Checks suggest that results are robust as there are no inconsistencies in the direction of effects for any models (SI 9). Robustness checks also confirmed the significance (or lack of) of the main analysis for As-Randomized analyses. There is slightly less agreement in the significance for As-Treated results (Fig. 4). This may be because power is reduced in As-Treated results due to lower sample sizes.

When presenting results, we talk both about Treated households and Treatment households. Treated households are those in treatment communities which signed Watershared agreement. They are always compared against a counterfactual of households in control communities matched on socio-economic predictors of uptake of Watershared agreements. These results are those from the As-Treated analysis. Treatment households are all those in treatment communities. They are always compared against households in control communities (without matching). These results derive from the As-Randomized analysis.

For some outcomes there were significant treatment effects in the direction hypothesized (Fig. 2, 5, Fig. SI 9). Treated households and Treatment households had significantly more small fruit trees (a mean of 50 and 25 respectively), and more fruit trees in production (mean of 100 and 150, respectively), than their counterfactuals. Treated and Treatment households were also more likely to perceive positive trends over the last 5 years in water quantity and forest condition. They also were more likely to perceive that the wider community care more about the forest (Fig. 2, 5, Fig. SI 9). The intervention may also have had an effect of increasing the area of improved grazing

land; while the results of the main model were not significant following p value correction, the models used in the robustness checks did show a significant effect (Fig. 2, 4, Fig. SI 9).

We did not find evidence of a treatment effect on whether or not a household perceives gaining benefits from forest (Fig 2, 5, Fig. SI 9). However, given that the vast majority (over 90% in all groups) of respondents perceived benefits at baseline, there was little scope for increase. Nor were there treatment effects on irrigation access or irrigated land extent or perceptions of changes in water quality over time.

There was no convincing treatment effect (i.e. effects were not significant after p value correction) on whether intakes were protected from cattle, or the strength of protection of main water intakes from cattle access (Fig.2, 4). However, our analyses on the nature of intake protection suggest that Treated and Treatment households were more likely to use barbed wire than traditional methods to protect intakes, and to have protected intakes more recently than their counterfactuals.

For some outcomes there were significant treatment effects in the opposite direction to our hypotheses. At endline, control respondents were about 20% more likely to be members of water committees than Treated or Treatment households (Fig. 2, 5). There was weaker evidence of a treatment effect against hypothesis for outcomes associated with diarrhea. The frequency of diarrhea was higher in treatment groups in both analyses (although the effect was not significant in robustness checks in the As-Treated analysis; Fig. 2, 4). There was also some evidence that incidence of diarrhea was higher in the Treatment group (although this was not significant after p-value correction and the effect was not seen in the As-Treated analysis). This result may be an artefact of subsample bias or a lack of power in this subgroup. The diarrhea analysis was conducted on a small subset of the data (only those households with children and their own water intake). In the As-Randomized analysis, only 7 incidents of diarrhea were reported in the control group (N = 61); this was further reduced after matching in the As-Treated analyses. It followed that in some models there was perfect separation.

Discussion

Data analysis involves multiple decisions as researchers seek to reveal the truth from complex, often messy, data (Fraser et al. 2018). Studies revealing the lack of reproducibility in fields such as pre-clinical medicine (Freedman et al. 2015) and psychology (Open Science Collaboration 2015) have resulted in much needed scrutiny of how these

decisions are vulnerable to confusion, or even corruption. While conservation science has so far avoided a scandal of reproducibility, a recent study of researchers in ecology and evolution (Fraser et al. 2018) revealed a worrying prevalence of cherry picking (failing to report results which are not significant), or reporting unexpected findings as if they were hypothesized from the start (Hypothesizing After Results Are Known; HARKing). Pre-registration of analysis can avoid these problems as long as the study is adequately powered to detect differences of interest. Submitting planned research for peer review as a registered report goes one step further and also reduces publication bias (Parker et al. 2019). The multiple outcomes available for analysis from the Watershared RCT were inevitably vulnerable to cherry-picking and HARKing. By publishing this as a registered report, we reduced both the temptation to use, and the impression we may have used, questionable research practices to tell a better story. Ideally, of course, pre-registration should precede data collection. Data collection for this RCT began in 2010 and was complete in 2015, before pre-registration was widely advocated. However we submitted Stage 1 before looking at any outcome variables meaning the study was accepted in principle based on the introduction and methods alone. This study is one of the first registered reports in conservation science.

While large-scale RCTs of interventions are receiving increasing attention (for example the 2019 economics Nobel prize was awarded to Kremer, Banerjee and Duflo for their experimental work on alleviating poverty), they remain rare in conservation (Pynegar et al. 2019). Our paper is the first we know of which uses a Randomized Control Trial to look at outcomes from across the theory of change to give insights into the mechanism by which a conservation intervention works, or does not work. Watershared ultimate aims are to reduce the rate of forest clearance and degradation, improve livelihoods, and improve water quality and quantity. Our previous analyses of the intervention, looking simply at biophysical measures of ultimate outcomes, revealed minimal impact on deforestation (Wiik et al. 2019) and water quality (Pynegar et al. 2018). Those analyses alone say little about why the intervention may not have resulted in a change in those outcomes, or whether measurable impact might be detected given time. Looking closely at intermediate outcomes, as we do in this paper, provides valuable insights to answer such questions about mechanisms.

Watershared aims to conserve forest by increasing the awareness of the benefits forests provide and increasing farmers' investment in improved grazing (reducing forest grazing) and alternatives to cattle ranching (such as fruit production). Over 90% of respondents already perceived benefits from forests, so it is unsurprising that the

intervention did not increase this. The intervention appears to have increased the area of improved grazing, and also significantly increased fruit tree production (although this was not yet apparent in market values, which we predicted owing to lags in fruit tree maturation). Watershared also provided cement and irrigation tubing to increase irrigated agriculture. However, due to their relatively high cost, they were less popular than barbed wire and fruit trees (and field observations suggest these materials were often used to improve drinking water systems). It is therefore unsurprising that the program had no significant impact on irrigation capacity. Previous remote-sensing analysis of forest area showed no landscape-scale impact of Watershared on deforestation (Wiik et al. 2019). However, our results suggest that the intervention is having an impact on some relevant intermediate outcomes. The program's theory of change may thus be correct, but it is perhaps still too early to detect ultimate impacts.

Watershared aims to improve water quality by encouraging people to keep cattle out of rivers by providing barbed wire, and materials to build cattle drinking troughs. While there was no evidence of the intervention increasing the number of water intakes protected from cattle nor cattle drinking points separated from rivers ($p < 0.05$), Treated and Treatment households were more likely to use barbed wire to protect water intakes and to have done this more recently, suggesting that more intakes might have been protected at baseline in control communities (we lack data on this; however it would be surprising to invest in protecting intakes already protected). Regardless of a potential baseline imbalance, it is clear that water quality-related outcomes did not materialize. The lower membership of water committees in treatment than control groups may have been because households perceived issues with water quality had been dealt with by the intervention (but this deserves further investigation).

It is interesting that the As-Treated and As-Randomized analyses gave quite similar results. This suggests that identified effects of Watershared were felt by the wider population in treated communities and not just those who entered agreements. This is not surprising given that several outcomes either related to outcomes independent of individual actors (such as perceptions of the wider environment), or related to shared resources (such as water intakes).

One of the key challenges in conservation impact evaluations is dealing with spillovers (Baylis et al. 2016). When benefits of a program flow from treatment to control communities (through biophysical or social processes), the measured difference in outcomes of interest between the groups is reduced, making an impacts of the intervention harder to detect. Accepting the risk of such spillover is inherent to any study such as ours, which treats communities

within a continuous social-ecological system as randomization units; however, as spillovers make it harder to detect an intervention effect, we believe that our identification of significant effects is conservative.

Overall, we show that the Watershared intervention has changed land use practices and environmental perceptions. Following the theory of change, it seems plausible that some ultimate outcomes may yet materialise. However the impact of the intervention would likely have been enhanced with spatial targeting (Pynegar et al. 2018), increased technical support, and higher additionality (Bottazzi et al. 2018).

Given the importance of improving the effectiveness of conservation interventions, especially those which aim to deliver better social outcomes alongside environmental benefits (Sims & Alix-Garcia 2017), more robust evaluations are sorely needed (Snilsveit et al. 2019). While RCTs certainly are not practical or desirable in every situation and have well understood limitations (Deaton & Cartwright 2018), we show that the criticism that RCTs are inherently reductionist and cannot give insights into mechanisms is unjustified. By using the Watershared RCT to explore outcomes from across the intervention's theory of change we have provided understanding of what is, and is not, changing on the ground because of the intervention. Such an analysis is inevitably complex. Pre-registration (ideally alongside a peer review commitment to publish whether the results are positive or negative), is particularly important in such circumstances. We hope that pre-registration becomes the norm in conservation science, as it is increasingly so in other applied disciplines.

Supporting Information

Household survey coverage (Appendix S1), outcome selection rationale (Appendix S2), definition of Treated households (Appendix S3), matching variable selection rationale (Appendix S4), propensity score construction and selection (Appendix S5), matching protocol (Appendix S6), multiple testing adjustment (Appendix S7), outcome regression details (Appendix S8), outcome regression supplementary results (Appendix S9), and supplementary literature (Appendix S10) are available online. The authors are solely responsible for the content and functionality of these materials. Queries should be directed to the corresponding author.

Literature cited

Asquith N, Vargas MT. 2007. Fair deals for watershed services in Bolivia. IIED.

343 Asquith NM. 2016. Watershared: Adaptation, mitigation, watershed protection and economic development in Latin
344 America. Climate & Development Knowledge Network.

345 Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. 2012. Realist randomised controlled trials: A new approach to
346 evaluating complex public health interventions. Social Science & Medicine **75**:2299–2306. Pergamon. Available from
347 https://www.sciencedirect.com/science/article/abs/pii/S0277953612006399 (accessed May 7, 2019).

348 Börner J, Baylis K, Corbera E, Ezzine-de-Blas D, Ferraro PJ, Honey-Rosés J, Lapeyre R, Persson UM, Wunder S. 2016.
349 Emerging Evidence on the Effectiveness of Tropical Forest Conservation. PLOS ONE **11**:e0159152.

350 Börner J, Baylis K, Corbera E, Ezzine-de-Blas D, Honey-Rosés J, Persson UM, Wunder S. 2017. The Effectiveness of
351 Payments for Environmental Services. World Development **96**:359–374.

352 Bottazzi P et al. 2017. Baseline and endline socio-economic data from a Randomised Control Trial of the
353 Watershared intervention in the Bolivian Andes. ReShare UK Data Archive. Colchester, Essex. Available from
354 http://reshare.ukdataservice.ac.uk/852623/.

355 Bottazzi P, Wiik E, Crespo D, Jones JPG. 2018. Payment for Environmental "Self-Service": Exploring the Links Between
356 Farmers' Motivation and Additionality in a Conservation Incentive Programme in the Bolivian Andes. Ecological
357 Economics **150**:11–23.

358 Daw JR, Hatfield LA. 2018. Matching and Regression to the Mean in Difference-in-Differences Analysis. Health
359 Services Research **53**:4138–4156. John Wiley & Sons, Ltd (10.1111).

360 Daw TM et al. 2016. Elasticity in ecosystem services: Exploring the variable relationship between ecosystems and
361 human well-being. Ecology and Society **21**:art11. The Resilience Alliance.

362 Deaton A, Cartwright N. 2018. Understanding and misunderstanding randomized controlled trials. Social Science and
363 Medicine **210**:2–21.

364 Ezzine-de-Blas D, Wunder S, Ruiz-Pérez M, Moreno-Sanchez R del P, Nikolakis W, Wilson P. 2016. Global Patterns in
365 the Implementation of Payments for Environmental Services. PLOS ONE **11**:e0149847. Earthscan at Routledge.
366 Available from http://dx.plos.org/10.1371/journal.pone.0149847 (accessed February 23, 2018).

Ferraro PJ. 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation **2009**:75–84. John Wiley & Sons, Ltd.

Ferraro PJ, Hanauer MM. 2014. Advances in Measuring the Environmental and Social Impacts of Environmental Programs. Annual Review of Environment and Resources **39**:495–517.

Fortnam M, Brown K, Chaigneau T, Crona B, Daw TM, Gonçalves D, Hicks C, Revmatas M, Sandbrook C, Schulte-Herbruggen B. 2019. The Gendered Nature of Ecosystem Services. Ecological Economics **159**:312–325. Elsevier.

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. 2018. Questionable research practices in ecology and evolution. PLOS ONE **13**:e0200303. Public Library of Science. Available from https://dx.plos.org/10.1371/journal.pone.0200303 (accessed November 21, 2019).

Freedman LP, Cockburn IM, Simcoe TS. 2015. The Economics of Reproducibility in Preclinical Research. PLOS Biology **13**:e1002165. Public Library of Science. Available from https://dx.plos.org/10.1371/journal.pbio.1002165 (accessed November 21, 2019).

Glennerster R, Takavarasha K. 2013. Running Randomized Evaluations: A Practical Guide. Princeton University Press, Princeton, USA.

Greifer N. 2019. cobalt: Covariate Balance Tables and Plots.

Grillos T. 2017. Economic vs non-material incentives for participation in an in-kind payments for ecosystem services program in Bolivia. Ecological Economics.

Grillos T, Bottazzi P, Crespo D, Asquith N, Jones JPG. 2019. In-kind conservation payments crowd in environmental values and increase support for government intervention: A randomized trial in Bolivia. Ecological Economics **166**:106404.

Hill J. 2008. Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin,Statistics in Medicine. Statistics in Medicine **27**:2055–2061.

Ho DE, Imai K, King G, Stuart EA, Ho DE, Imai K, King G, Stuart EA. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis **15**:199–236. Cambridge University Press.

Jack BK, Kousky C, Sims KRE. 2008. Designing payments for ecosystem services: Lessons from previous experience with incentive-based mechanisms. Proceedings of the National Academy of Sciences **105**:9465–9470. Available from http://www.pnas.org/content/105/28/9465.abstract.

Jayachandran S, de Laat J, Lambin EF, Stanton CY, Audy R, Thomas NE. 2017. Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. Science **357**:267–273.

Liu Z, Kontoleon A. 2018. Meta-Analysis of Livelihood Impacts of Payments for Environmental Services Programmes in Developing Countries. Ecological Economics **149**:48–61.

Miteva DA, Pattanayak SK, Ferraro PJ. 2012. Evaluation of biodiversity policy instruments: What works and what doesn't? Oxford Review of Economic Policy **28**:69–92.

Open Science Collaboration OS. 2015. Estimating the reproducibility of psychological science. Science **349**:aac4716–aac4716. American Association for the Advancement of Science. Available from http://www.ncbi.nlm.nih.gov/pubmed/26315443 (accessed November 21, 2019).

Parker T, Fraser H, Nakagawa S. 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology **33**:cobi.13342. John Wiley & Sons, Ltd (10.1111). Available from https://onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13342 (accessed August 12, 2019).

Pynegar E. 2018. The use of Randomised Control Trials in evaluating conservation interventions: the case of Watershared in the Bolivian Andes. Bangor University.

Pynegar EL, Gibbons JM, Asquith NM, Jones JPG. 2019. What role should Randomised Control Trials play in providing the evidence base underpinning conservation? Oryx **6**.

Pynegar EL, Jones JPG, Gibbons JM, Asquith NM. 2018. The effectiveness of Payments for Ecosystem Services at delivering improvements in water quality: lessons for experiments at the landscape scale. PeerJ **6**:e5753. PeerJ Inc. Available from https://peerj.com/articles/5753 (accessed November 4, 2018).

Samii C, Lisiecki M, Kulkarni P, Paler L, Chavis L. 2014. Effects of Payment for Environmental Services (PES) on Deforestation and Poverty in Low and Middle Income Countries: A Systematic Review. Campbell Systematic Reviews **10**.

Sekhon JS. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. Journal of Statistical Software **42**:1–52.

Sims KRE, Alix-Garcia JM. 2017. Parks versus PES: Evaluating direct and incentive-based land conservation in Mexico. Journal of Environmental Economics and Management **86**:8–28.

Snilsveit B, Stevenson J, Langer L, Da N, Zafeer S, Promise R, Polanin NJ, Shemilt I, Eyers J, Ferraro PJ. 2019. Systematic Review 44 Incentives for climate mitigation in the land use sector-the effects of payment for environmental services (PES) on environmental and socio-economic outcomes in low-and middle-income countries A mixed-method systematic review. Available from https://doi.org/10.23846/SR00044 (accessed December 9, 2019).

Streiner DL. 2015. Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. The American Journal of Clinical Nutrition **102**:721–728.

Wan F. 2019. Matched or unmatched analyses with propensity-score–matched data? Statistics in Medicine **38**:289–300. John Wiley & Sons, Ltd.

White H. 2009. Theory-based impact evaluation: principles and practice. Journal of Development Effectiveness **1**:271–284. Taylor & Francis . Available from http://www.tandfonline.com/doi/abs/10.1080/19439340903114628 (accessed April 24, 2019).

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

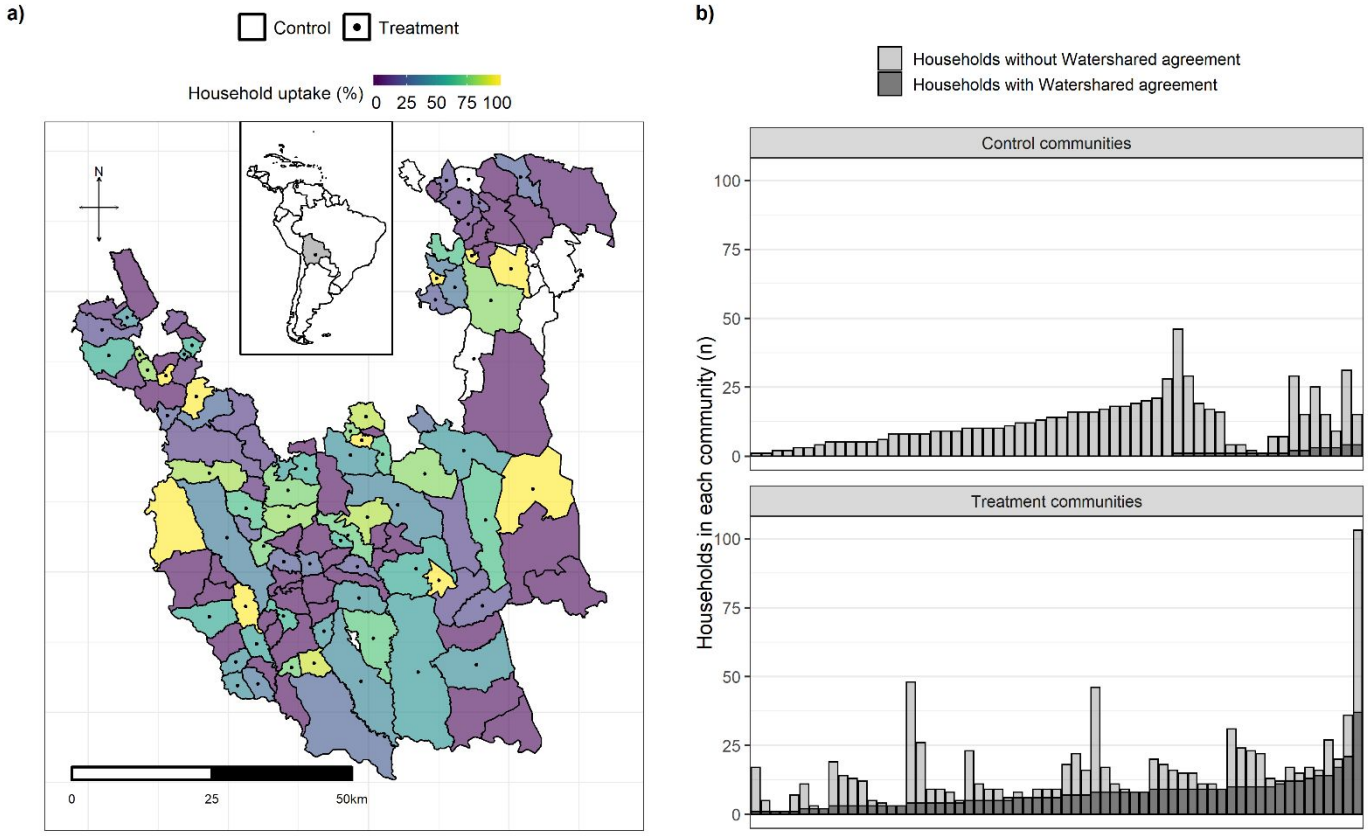
Wiik E, d'Annunzio R, Pynegar E, Crespo D, Asquith N, Jones JPG. 2019. Experimental evaluation of the impact of a payment for environmental services program on deforestation. Conservation Science and Practice:e8. John Wiley & Sons, Ltd.

Wilebore B, Voors M, Bulte E, Coomes D, Kontoleon A. 2019. Unconditional Transfers and Tropical Forest Conservation. Evidence from a Randomized Control Trial in Sierra Leone. American Journal of Agricultural Economics.

439 Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric
440 generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**:3–36.
441 Blackwell Publishing Ltd.

442 Wood SN. 2017. Generalized Additive Models: An Introduction with R, 2nd edition. CRC Press.

443



444

445 **Figure 1:** a: Locations of the 64 control communities (Watershared agreements not offered) and 65 treatment
446 communities (Watershared agreements offered) within the Area Natural de Manejo Integrado Río Grande y Valles
447 Cruceños protected area. White communities are those for which there are no households with both baseline and
448 endline data; omitted from our analysis. b: The distribution of the number of households per community (and the
449 number which took up Watershared agreements) in control (top) and treatment (bottom) communities, ordered by
450 number of households with agreements.

451

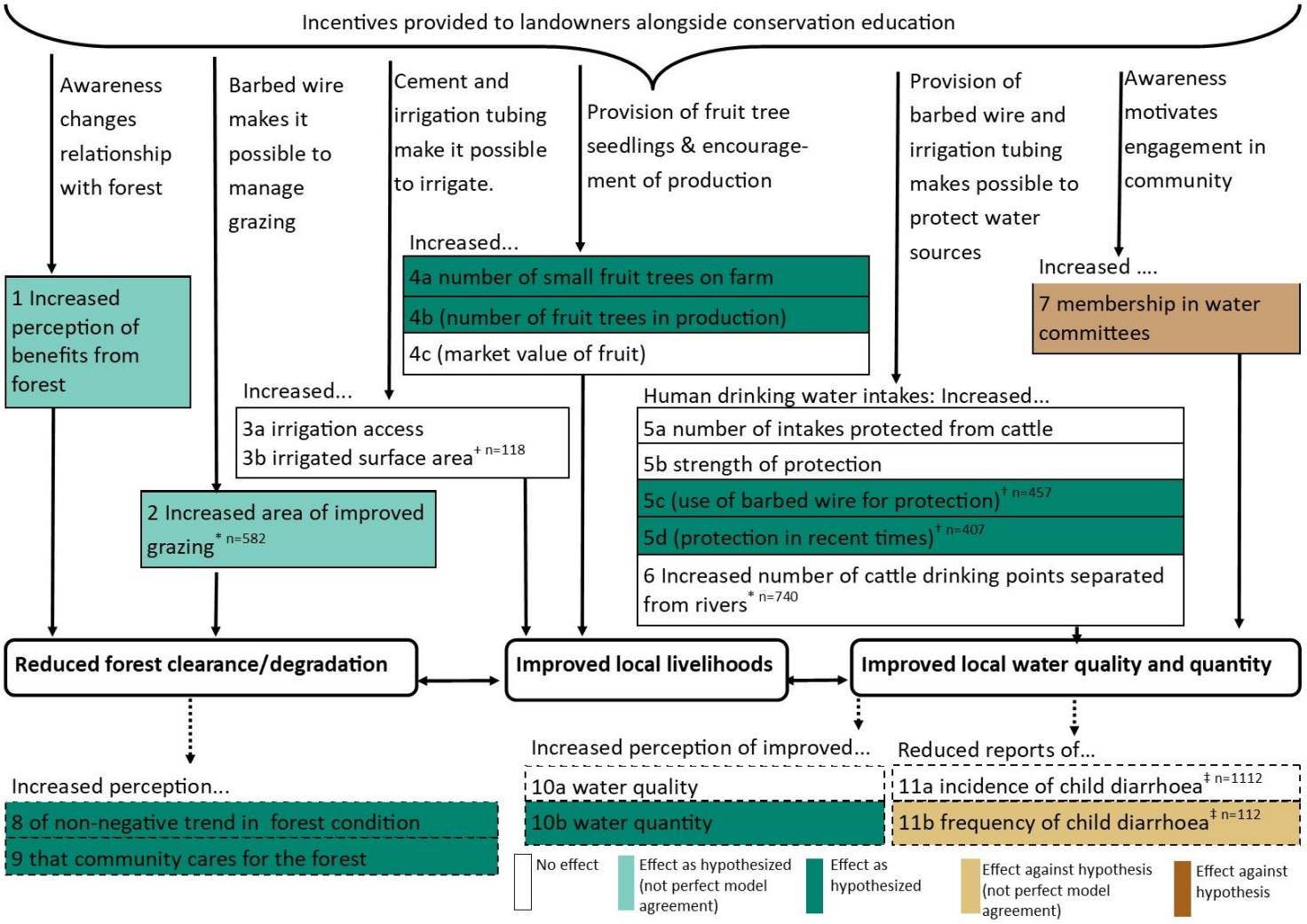
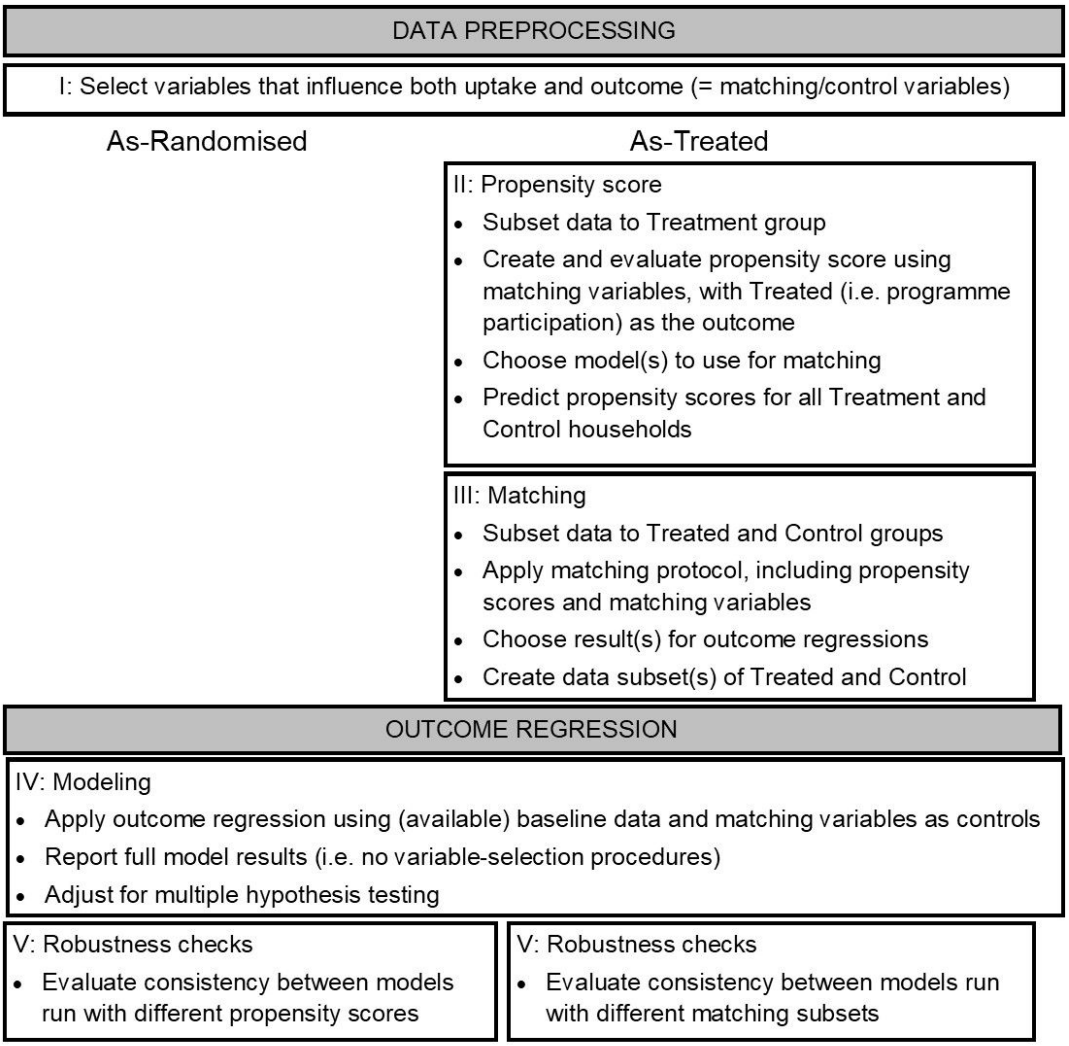


Figure 2: The simplified theory of change linking the Watershared intervention with intermediate outcomes (square boxes), ultimate outcomes (rounded boxes) and indicators of these ultimate outcomes (square boxes with dashed lines). The hypothesized direction of the effect of the intervention is indicated for each outcome tested in our analysis. Some analyses are only relevant for a sub-set of data: *: Households who own cattle; +: Households who have irrigation access; †: Households reporting protected drinking water intakes; ‡: Households with children and personal water intake. Brackets indicate outcomes for which we expect limited impact (e.g. the number of fruit trees in production may not yet be affected as they will not yet have had time to reach maturity). The colors show the results of the regression analyses (for As-Treated models only, Fig. SI 9.3 shows the same results from the As-Randomized models) with green indicating results which support our hypothesis and browns indicating results against our hypotheses. Less saturated colors show outcomes for where there was some disagreement in the significance between the models used as robustness checks (Figure 4).



464 **Figure 3:** Outline of methods workflow for the As-Randomized models and As-Treated models. Phases I, II, and III
465 show pre-processing undertaken for Stage 1 of this registered report (they were completed before initial peer
466 review). Stages IV and V occurred at Stage 2.

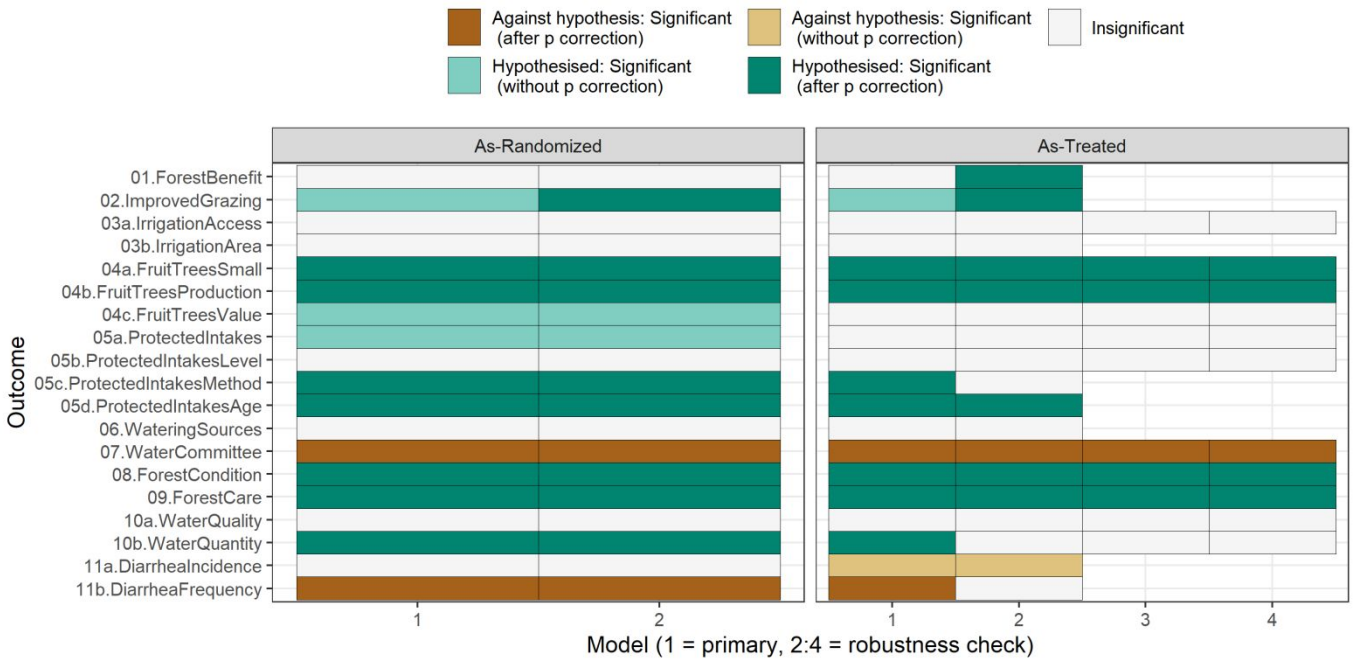


Figure 4: Robustness checks based on comparing the significance of intercept differences between control and treatment groups for the As-Randomized and As-Treated analyses for the primary model and subsidiary models (max 4 in As-Treated due to matching protocols). P-value correction uses the Benjamini-Hochberg threshold (see Methods).

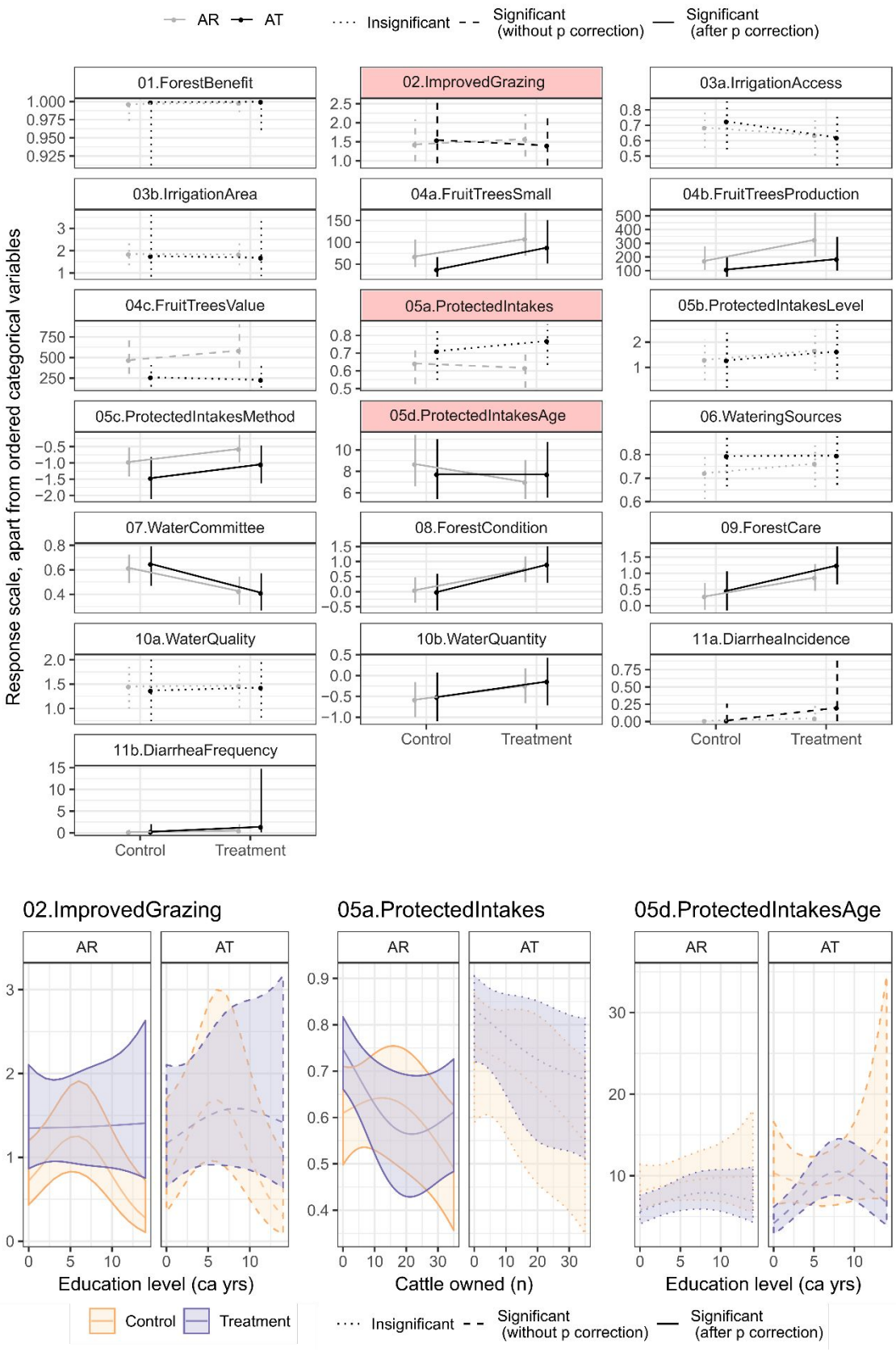


Figure 5: The differences in intercept values (line plots with confidence bars) and interaction slopes (line plots with confidence bands, where the x axis is a continuous variable) between control and treatment groups for both As-Randomized (AR) and As-Treated (AT) analysis. Only interactions where at least one analysis shows significance are

shown. The predictions are based on mean values for continuous predictors, and common values for factor predictors (e.g., perceiving benefits from forest). 95% confidence intervals relate to the entire prediction, rather than the control-treatment difference.

Table 1: Variables selected for matching variables and control variables in the final-outcome models, indicating which will be interacted with the experimental group (treatment or control, or treated or matched control) in outcome regressions (Stage 2) (NA = Variables not interacted).

Variable	Category	Mechanism	Outcomes for which variable is interacted with experimental group
Community work frequency (n/yr)	Community cohesion	Likely to be related to motivation to participate and adhere to agreements due to social norms	NA
Generations in a community (n)	Community cohesion	Likely to be related to level of engagement in the community and also ability to participate and follow through with agreements	NA
Land owned (ha)	Wealth	Likely to be related to ability to afford to invest time and effort in conservation	Water committee membership; Diarrhea; Irrigation implementation
Forest ownership (binary)	Wealth	Likely to be related to owning eligible land and being able to afford to invest time and effort in conservation	NA
Cattle owned (n)	Wealth	Likely to be related to ability to afford to invest time and effort in conservation	Cattle and human drinking water management; Improved grazing
Number of rooms in home	Wealth	Likely to be related to ability to afford to invest time and effort in conservation	NA

Education level (approx. yrs)	Education	Likely to be related to capacity to engage with the conservation program	Cattle and human drinking water management; Improved grazing; Diarrhea; Irrigation implementation; Water committee membership
Perceived benefits from forest (binary)	Environmental attitudes	Related to motivation to engage with conservation	NA
Perceived problems in water quality, quantity (binary)	Environmental attitudes	Related to motivation to engage with conservation	Human and cattle drinking water management