The effectiveness of Payments for Ecosystem Services at delivering improvements in water quality: lessons for experiments at the landscape scale

Edwin L Pynegar [1, 2] , Julia P G Jones [Corresp., 1] , James M Gibbons [1] , Nigel M Asquith [3, 4]

[1] School of Environment, Natural Resources and Geography, Bangor University, Bangor, United Kingdom

[2] School of Life Sciences, University of Sussex, Falmer, United Kingdom

[3] Harvard Forest, Harvard University, Petersham, United States

[4] Kennedy School of Government, Harvard University, Cambridge, United States

Corresponding Author: Julia P G Jones
Email address: julia.jones@bangor.ac.uk

Background. Randomised Control Trials (RCTs) are used in impact evaluation in a range of fields. However, despite calls for their greater use in environmental management, their use to evaluate landscape scale interventions remains rare. Payments for Ecosystem Services (PES) incentivise land users to manage land to provide environmental benefits. We present the first RCT evaluation of a PES program aiming to improve water quality. *Watershared* is a PES program which incentivises landowners to avoid deforestation and exclude cattle from riparian forests. Using this unusual landscape-scale experiment we explore the efficacy of *Watershared* at improving water quality, and draw lessons for future RCT evaluations of landscape-scale environmental management interventions. **Methods.** One hundred and twenty-nine communities in the Bolivian Andes were randomly allocated to treatment (offered *Watershared* agreements) or control (not offered agreements) following baseline data collection (including *Escherichia coli* contamination in most communities) in 2010. We collected end-line data in 2015. Using our end-line data, we explored the extent to which variables associated with the intervention (e.g. cattle exclusion, absence of faeces) predict water quality locally. We then investigated the efficacy of the intervention at improving water quality at the landscape scale using the RCT. This analysis was done in two ways; for the subset of communities for which we have both baseline and end-line data from identical locations we used difference-in-differences (matching on baseline water quality), for all sites we compared control and treatment at end-line controlling for selected predictors of water quality. **Results.** The presence of cattle faeces in water adversely affected water quality suggesting excluding cattle has a positive impact on water quality locally. However, both the matched difference-in-differences analysis and the comparison between treatment and control communities at end-line suggested *Watershared* was not effective at reducing *E. coli* contamination at the

landscape scale. Uptake of *Watershared* agreements was very low and the most important land from a water quality perspective (land around water intakes) was seldom enrolled.

Discussion. Although excluding cattle may have a positive local impact on water quality, higher uptake and better targeting would be required to achieve a significant impact on the quality of water consumed in the communities. Although RCTs potentially have an important role to play in building the evidence base for approaches such as PES, they are far from straightforward to implement. In this case, the randomised trial was not central to concluding that *Watershared* had not produced a landscape scale impact. We suggest that this RCT provides valuable lessons for future use of randomised experiments to evaluate landscape-scale environmental management interventions.

1 *The effectiveness of Payments for Ecosystem Services at delivering improvements in water*

2 *quality: lessons for experiments at the landscape scale*

4 *Author Names and Affiliations*

5 Edwin L. Pynegar[a, b], Julia P. G. Jones[a], James M. Gibbons[a], and Nigel M. Asquith[c, d]

6 [a] School of Environment, Natural Resources and Geography, Bangor University, Bangor, UK.

7 [b] School of Life Sciences, University of Sussex, Falmer, UK.

8 [c] Harvard Forest, Harvard University, Petersham, USA.

9 [d] Harvard Kennedy School, Harvard University, Cambridge, USA.

10 *Corresponding author*

11 Julia P G Jones, julia.jones@bangor.ac.uk

12   **Abstract**

13   **Background.** Randomised Control Trials (RCTs) are used in impact evaluation in a range of

14   fields. However, despite calls for their greater use in environmental management, their use to

15   evaluate landscape scale interventions remains rare. Payments for Ecosystem Services (PES)

16   incentivise land users to manage land to provide environmental benefits. We present the first RCT

17   evaluation of a PES programme aiming to improve water quality. *Watershared* is a PES

18   programme which incentivises landowners to avoid deforestation and exclude cattle from riparian

19   forests. Using this unusual landscape-scale experiment we explore the efficacy of *Watershared* at

20   improving water quality, and draw lessons for future RCT evaluations of landscape-scale

21   environmental management interventions.

22   **Methods.** One hundred and twenty-nine communities in the Bolivian Andes were randomly

23   allocated to treatment (offered *Watershared* agreements) or control (not offered agreements)

24   following baseline data collection (including *Escherichia coli* contamination in most communities)

25   in 2010. We collected end-line data in 2015. Using our end-line data, we explored the extent to

26   which variables associated with the intervention (e.g. cattle exclusion, absence of faeces) predict

27   water quality locally. We then investigated the efficacy of the intervention at improving water

28   quality at the landscape scale using the RCT. This analysis was done in two ways; for the subset

29   of communities for which we have both baseline and end-line data from identical locations we

30   used difference-in-differences (matching on baseline water quality), for all sites we compared

31   control and treatment at end-line controlling for selected predictors of water quality.

32   **Results.** The presence of cattle faeces in water adversely affected water quality suggesting

33   excluding cattle has a positive impact on water quality locally. However, both the matched

34   difference-in-differences analysis and the comparison between treatment and control communities

35   at end-line suggested *Watershared* was not effective at reducing *E. coli* contamination at the

36   landscape scale. Uptake of *Watershared* agreements was very low and the most important land

37   from a water quality perspective (land around water intakes) was seldom enrolled.

38   **Discussion.** Although excluding cattle may have a positive local impact on water quality, higher

39   uptake and better targeting would be required to achieve a significant impact on the quality of

40   water consumed in the communities. Although RCTs potentially have an important role to play in

41   building the evidence base for approaches such as PES, they are far from straightforward to

42   implement. In this case, the randomised trial was not central to concluding that *Watershared* had

43   not produced a landscape scale impact. Nevertheless we suggest that this RCT provides valuable

44   lessons for future use of randomised experiments to evaluate landscape-scale environmental

45   management interventions.

46   **Introduction**

47   Whether an intervention is effective at delivering the outcomes expected is a key question for

48   evidence-based policy making (e.g. White, 2013). This question is highly pertinent in the field of

49   conservation and environmental management as awareness grows of the amount of money that has

50   been spent on interventions with limited understanding of their effectiveness (Ferraro &

51   Pattanayak, 2006; Bowler et al., 2012; Waeber et al., 2016; Salzman et al., 2018). As a result,

52   conservation effectiveness is being widely discussed in the academic literature (Baylis et al., 2016;

53   Börner et al., 2016; Sutherland & Wordley, 2017), the policy literature (Puri et al., 2016; Duchelle,

54   Wunder & Martius, 2018), conservation journalism (Dasgupta, Gaworecki & Burivalova, 2018)

55   and the mainstream media (Mooney, 2016). Randomised Control Trials (RCTs), in which

56   experimental units are randomly allocated to treatment or control groups, allow the creation of

57   robust counterfactuals from which to infer what would have happened in the absence of the

58   intervention (e.g. Rubin 1974). RCTs are widely used in many areas of public policy including

59   medicine, education, and development economics (Glennerster & Takavarasha, 2013; Council of

60   Economic Advisers, 2014). Although small-scale RCTs have been a mainstay of applied

61   ecological experiments for decades, there are very few examples of RCTs of large-scale

62   environmental management interventions, and there have been calls for their increased use

63   (Greenstone & Gayer, 2009; Samii et al., 2014; Baylis et al., 2016; Börner et al., 2016, 2017).

64   Payments for Ecosystem Services (or Payments for Environmental Services – the terms are largely

65   interchangeable [Wunder, 2015]) translate external, non-market values of the environment into

66   financial incentives for local actors to provide environmental services. The focus of many PES

67   programs in Latin America (Martin-Ortega, Ojea & Roux, 2013; Grima et al., 2016), and to a lesser

68   extent in Asia and Africa, e.g. (Calvet-Mir et al., 2015) is the increase or maintenance of supply

69   of good quality water. At least 1.8 billion people still rely on drinking water sources contaminated

70   with faecal matter (Bain et al., 2014a). Where sources lack adequate physical or chemical treatment

71   the quality of drinking water is influenced by land use and ecosystem management around and

72   upstream of those water sources. Hence provision of clean water can be considered as an ecosystem
73   service or as a precursor to multiple ecosystem services benefiting society (Keeler et al., 2012).
74   Despite the increasing number of PES programs in operation, there are very few robust evaluations
75   of the extent to which they deliver the outcomes they seek to supply (Pattanayak, Wunder &
76   Ferraro, 2010; Miteva, Pattanayak & Ferraro, 2012; Naeem et al., 2015; Baylis et al., 2016; Börner
77   et al., 2017; Ferraro, 2017; Salzman et al., 2018). We know of only a single published Randomised
78   Control Trial of a PES (evaluating the impact of a program in Uganda on deforestation;
79   Jayachandran et al., 2017) and none evaluating the impact of PES on water quality.

80   Gastrointestinal illnesses caused by consumption of contaminated water are a major cause of
81   mortality and morbidity in the developing world (Prüss-Ustün et al., 2014). *Escherichia coli* is a
82   bacterium that lives only in the guts of warm-blooded animals (Leclerc et al., 2001). While some
83   strains of *E. coli* are pathogenic, the majority are not but are useful indicators of faecal
84   contamination and the presence of other pathogens (Ashbolt, Grabow & Snozzi, 2001). Sources of
85   faecal contamination may include faulty sewerage systems and leaking septic tanks (Richards et
86   al., 2016), open defecation (Spears, Ghosh & Cumming, 2013), or the presence of wildlife (Ahmed
87   et al., 2012). However, a major source of contamination is the presence of domestic livestock,
88   particularly free-roaming cattle (Crane et al., 1983). Therefore, cattle exclusion has been practiced
89   as a means of reducing faecal contamination of watercourses. In the UK, for example, the Good
90   Agricultural and Environmental Conditions standard 1 requires farmers in receipt of certain
91   subsidies to maintain buffer strips and refrain from spreading manure within areas close to water
92   bodies (Anonymous, 2016). There is evidence of such actions being effective at significantly
93   reducing *E. coli* concentration and other faecal contamination of water supplies (Sunohara et al.,
94   2012). However, many uncertainties remain about the extent to which these interventions,
95   incentivised via a PES program, can deliver consistent benefits in water quality at the landscape
96   scale.

97   The Bolivian non-governmental organization *Fundación Natura Bolivia* (*Natura*) began using in-
98   kind incentives to encourage conservation in the Andean region of Bolivia in 2003. Their program,
99   now known as *Watershared,* aims to slow forest loss and protect the quality of water available to
100 communities through providing modest development support in exchange for avoiding
101 deforestation and excluding livestock from riparian forest (Bottazzi et al., 2018). Although *Natura*

102   does not characterise *Watershared* as PES (Asquith, 2016), the program meets the most widely

103   used PES definition (Wunder, 2015): "voluntary transactions between service users and service

104   providers that are conditional on agreed rules of natural resource management for generating

105   offsite services". As of 2016, 210,000 hectares of forest owned by 4,500 households were under

106   *Watershared* conservation agreements (Asquith, 2016).

107   Given the growing interest in evaluating the effectiveness of different conservation approaches,

108   *Natura* established a Randomised Control Trial (RCT) to evaluate *Watershared*. One hundred and

109   twenty-nine communities were randomly allocated to control (not offered *Watershared*

110   agreements) or treatment groups (offered agreements). We use this unique setup to investigate the

111   effectiveness of the intervention at delivering improvements in microbial water quality. We

112   address three interconnected questions: 1) Do the features of *Watershared* agreements (e.g. cattle

113   exclusion, absence of faeces) have a measurable impact on water quality at a site, accounting for

114   other predictors? 2) Did the implementation of *Watershared* in treatment communities result in an

115   improvement in water quality relative to control communities? 3) What lessons does the

116   *Watershared* RCT evaluation offer for the wider use of experiments to evaluate the impact of

117   conservation interventions at the landscape scale?

118 **Materials and Methods**

119 *Context and RCT design*

120 This article focuses on the *Watershared* intervention in the Río Grande Valles Cruceños Natural

121 Integrated Management Area (Spanish acronym *ANMI Río Grande-Valles Cruceños*), a protected area

122 of 7339km² in the Andean region of the Santa Cruz Department in eastern Bolivia (Fig. 1). Forests

123 in this area are perceived locally as contributing to providing high-quality water for human

124 consumption and irrigation, despite the mixed scientific evidence on this topic (Bruijnzeel, 2004;

125 Ponette-González et al., 2015). Gastrointestinal illnesses are endemic; for example, in 2015 the

126 health centre of Moro Moro, a community of approximately 800 people, treated 236 cases of

127 diarrhoea (information from *Servicio Nacional Integral de Salud, Centro de Salud Moro Moro*,

128 obtained 4th April 2016). Faecal contamination from cattle is widely considered an important

129 contributor to the high prevalence of these diseases as the traditional farming system involves

130 cattle grazing freely within the forests from where most communities take their water. While some

131 communities have rudimentary sedimentation and filtration systems, these are of limited

132 effectiveness and often become clogged with sediment after each rainfall event. Chlorination or

133 other chemical treatment is rare.

134 In 2010, 129 communities within the *Río Grande-Valles Cruceños* protected area were selected

135 for inclusion in a Randomised Control Trial (Fig. 1). Consent to randomization was granted by

136 community leaders on the understanding that the intervention would subsequently be implemented

137 in all communities (this general roll-out was conducted in 2016). Communities were randomly

138 allocated to control (64 of these communities in which conservation agreements were not offered)

139 or treatment (65 communities in which agreements were offered) groups following stratification

140 based on municipality, community size, and estimated cattle density. The RCT was not blinded as

141 participants unavoidably knew whether they belonged to a treatment or control community.

142 However, in order to avoid observer bias effects during data collection, those conducting water

143 quality monitoring did not know which communities belonged to the treatment or control group.

144 Individuals belonging to treatment communities were offered the chance to conserve land

145 belonging to them under *Watershared* agreements (see Bottazzi et al., [2018] for more detail on

146 the *Watershared* program) and received education on the importance of cattle exclusion and forest

147 conservation for the maintenance of water quality and quantity. Individuals belonging to control

148 communities received the environmental education only. *Natura* offered landowners in treatment

149 communities three-year conservation agreements to conserve upstream forest and exclude cattle

150 from riparian forest in return for in-kind incentives such as fruit trees, barbed wire, or irrigation

151 materials. Participants could enrol their land in one of three kinds of agreements (for details see

152 Table S1). In this paper we only consider level 1 agreements, in which landowners were offered

153 $10/hectare/year in-kind equivalent (plus the equivalent of $100 regardless of the size of the area

154 enrolled) in return for conserving forested land within 100m of a watercourse and excluding cattle

155 from these areas. Landowners were offered the opportunity to enrol their land twice per year,

156 beginning in August 2011. Compliance monitoring and distribution of the in-kind compensations

157 was conducted yearly. A recent analysis suggests that 31% of the area of level 1 agreements

158 resulted in additional conservation (i.e. cattle were kept out of land which otherwise they would

159 have been allowed in; Bottazzi et al., 2018).

160 ***Sampling strategy***

161 Our analyses are based upon two rounds of monitoring of the quality of water intended for human

162 consumption. A baseline was taken between February and July of 2010 by the NGO *Natura* before

163 the sites were allocated to control or treatment groups. The allocation to control or treatment was

164 not stratified by measured water quality at the sites, and the baseline data was not otherwise used

165 until our team started work on the project in 2013. A more detailed end-line monitoring round was

166 undertaken by our team from Bangor University in collaboration with Natura between March and

167 May of 2015, i.e. following completion of the first signed agreements. In the end-line we had more

168 stringent protocols and also measured a number of additional potential indicators of water quality.

169 The communities within the RCT are small (maximum number of households is 123; Bottazzi et

170 al., 2017) with diverse water supply systems. Some have a single water intake, others multiple

171 intakes and in a few cases no functional intake at all (community members take water directly

172 from streams or other water bodies). Resource and logistical constraints meant that not all intakes

173 and taps could be sampled and so the tap supplying the community's school, along with the intake

174 supplying that tap, were taken as sampling sites based on the assumption that these would have

175 the greatest importance for health (Fig. 2). In cases where the community had no school, we

176 monitored at the intake which supplied the greatest number of households and a representative tap

177 fed by that intake. In the cases in which the community had no functional water system at all, we

178   took a sample in the water body where the greatest number of households collected their water.

179   Thus most communities had two site measurements (intake and tap), whereas a few (those lacking

180   an intake) only had one, in a few communities two intakes were measured. In this paper we refer

181   to the combination of an intake and a tap (or the location where households collected water where

182   an intake is not present) as a water system.

183   In 2010 the randomisation process assigned 129 communities to control or treatment group.

184   Independently technicians from *Natura* monitored water quality in 120 of these communities. In

185   the 2015 end-line we monitored 118 communities that were part of the RCT plus an additional six

186   communities not in the RCT. Two of the RCT sites were excluded from the end-line analysis as

187   their water system was supplied from rainfall collected from roofs and therefore cattle could not

188   affect water quality. Water quality was measured at each site (see below). In 2015 we also recorded

189   other characteristics of the site which may predict water quality. Table S2 summarises the sample

190   size for each analysis.

191   The *Natura* technicians who took water measurements in 2010 recorded the location with a

192   handheld GPS. However when we returned in 2015 it was not always possible to confirm that we

193   were at the same water intake (GPS readings in forest can be inaccurate). In addition, between

194   baseline data collection in 2010 and our team's visit in 2015, the location of both water intakes

195   and the main tap serving the community had changed in a number of sites. As a result, for only 83

196   sites in 47 communities were we 100% confident that had measured water quality at the same

197   location between 2010 and 2015 (Table S2). The fact that end-line data from the same location

198   was not available from all sites is somewhat analogous to the common problem of attrition due to

199   loss to follow up in clinical trials (Jüni, Altman & Egger, 2001). However as we still took an end-

200   line measure (although as we could not be sure the location was identical and therefore the data

201   were not included in the difference-in-differences analysis) we do not use the term attrition in this

202   paper. It is also important to note that a few of the water intakes could not be unambiguously

203   assigned to control or treatment (for example there are two cases in which one intake supplied two

204   communities, one of which was treatment and the other control); these have been excluded from

205   the RCT analysis.

206   ***Water quality monitoring***

207   The principal metric recorded was *Escherichia coli* colony forming unit concentration (CFUs) in
208   water samples. *E. coli* concentration, along with that of other non-*E. coli* bacteria belonging to the
209   coliform group, was quantified using the Coliscan Easygel method (Micrology Labs, Goshen, IN,
210   USA). Coliscan Easygel allows enumeration of coliforms as after incubation *E. coli* colonies
211   appear purple, blue-purple or dark blue due to metabolism of both beta-galactosidase and beta-
212   glucuronidase. Other non-*E. coli* coliforms are pink based upon metabolism of beta-galactosidase
213   only. Colonies of a blue-green or sky blue color (metabolism of beta-glucuronidase only) and white
214   colonies were not counted (Micrology Labs, 2016). The Easygel method (which uses only 5ml of
215   water per sample) does not comply with the World Health Organization's 100ml standard for
216   coliform monitoring but studies have shown that it is reasonably robust and not susceptible to false
217   negatives (Chuang, Trottier & Murcott, 2011). This method had been selected by *Natura* due to
218   the logistical challenges with using alternative methods such as membrane filtration in the remote
219   and low-resource context of the study area (many sites are reachable only with difficult drives and
220   long walks). When we modified the protocol in 2015 (to overcome some of the limitations of the
221   2010 protocol), we elected to retain the method for the same reason and to ensure data were
222   comparable as possible.

223   In 2010, one sample was placed into sterile Coliscan Easygel sampling flasks (35ml) taking care
224   to avoid any external contamination. Up to two days later (but normally on the same day) the
225   *Natura* team then inoculated Easygel Petri dishes using 5ml of the water from each flask. After
226   solidification the Petri dishes were sealed and incubated at ambient temperature for 48 hours, after
227   which numbers of *E. coli* and other non-*E. coli* coliform CFUs were counted.

228   In 2015, four separate samples were taken using sterile Coliscan Easygel sampling flasks (35ml
229   each) and placed on ice within 1 hour of sampling. Within 6 hours of sampling (although generally
230   within 4) we produced Easygel Petri dishes using 5ml of water from each flask as inoculum. After
231   solidification we sealed the Petri dishes and incubated them for 24 hours at 35-37°C in a portable
232   incubator (NQ28 model, Darwin Chambers, St Louis, MO, USA). In locations where no mains
233   electricity was available we maintained a constant incubation temperature through use of a 12V
234   vehicle power supply or supply from a car battery. After incubation we counted *E. coli* and other
235   non-*E. coli* coliform CFUs.

236   In 2015 we also measured in each site a number of physico-chemical parameters of water:

237   temperature, dissolved oxygen in mg/l and per cent of saturation value, pH, salinity and

238   conductivity in each site with an HQ40d portable multi-parameter meter and IntelliCAL LDO101,

239   PHC101 and CDC401 rugged probes respectively (HACH Company, Loveland, CO, USA). We

240   measured turbidity in formazin attenuation units through the use of a DR/850 colorimeter (HACH

241   Company, Loveland, CO, USA). Additionally, at the intake sites, we recorded other variables that

242   may predict *E. coli* concentration, including the presence or absence of cattle (judged based upon

243   presence of faeces, hoof prints, or cattle paths recently used) and the presence or absence of cattle

244   faeces in the riparian forest, in the water, or on banks. Some were recorded at the intake itself and

245   others along a 10m transect upstream (uphill in the case of intakes in springs) of the intake. Details

246   of all monitored variables are available in Table S3.

247   We used *Natura's* community database to determine which intakes supplied treatment or control

248   communities (we did not have this information when conducting field sampling to avoid any

249   observer bias effects). We used GIS software (ArcGIS 10.2, ESRI, Redlands, CA, USA) and

250   *Natura*'s shapefiles to calculate the percentage of eligible land in each community which was

251   enrolled in *Watershared* agreements (we compared this percentage between treatment and control

252   communities), and which monitored intakes fell within land enrolled in *Watershared* agreements.

253   We used data held by *Natura* to confirm which of these intakes were in agreements which had

254   been compliant with the agreements according to Natura's observations. In a number of the earliest

255   sites monitored during 2015 we accidentally disturbed the sediment in the water intake while

256   taking samples; sites in which this happened were recorded as such.

Balance achieved in the allocation of control and treatment

258   At baseline *Natura* collected a number of variables at the community level. Two of these (mean

259   number of cows per person in the community, and number of households in the community; both

260   blocked into two groups) were used in the stratification to allocate communities to control or

261   treatment. We explored the balance achieved at baseline for measured variables which may affect

262   the outcome of interest (*E. coli* contamination at end-line) using standardized mean differences

263   between treatment and control estimated in the R cobalt package (Greifer, 2018). Included

264   variables were those used in the stratification, time from community centre to hospital (a proxy for

265   remoteness), turbidity and the baseline measure of *E. coli* water contamination. Mean differences

266   were standardized using pooled standard deviations (Fig. 3). Looking at all 120 communities

267   monitored at baseline, variables appear quite well balanced (all close to or <0.25 pooled standard

268   deviation units).

269   However, there is less balance between control and treatment at baseline when just the sites for

270   which we have comparable data for baseline and end-line (i.e. excluding all sites for which we

271   were not 100% sure that water quality was measured at the same location). When the baseline

272   conditions for these 83 sites in 47 communities are compared, it is clear an imbalance has been

273   introduced in a number of variables. In particular, the remaining control sites had substantially

274   higher *E. coli* contamination) at baseline and were in more remote communities.

275   ***Statistical analysis***

276   We used generalized linear mixed model (GLMMs) to investigate whether features of

277   *Watershared* agreements (e.g. cattle exclusion, absence of faeces) have a measurable impact on

278   water quality at a site accounting for other predictors. For this we used the much richer 2015 data.

279   We used the *glmmADMB* package in R (Fournier et al., 2012; R Development Core Team, 2014;

280   Skaug et al., 2016) to produce GLMMs predicting *E. coli* concentrations, specifying a negative

281   binomial error structure and log-link. We included the water system identifier throughout as a

282   random effect, as measurement at an intake and then a tap supplied by that intake represents

283   repeated measures of the same water system. The unit of analysis is therefore the water system.

284   We used model selection based upon comparisons of the Akaike's Information Criterion (AIC)

285   and compared relative goodness of models through Akaike weighting. We then determined 95%

286   confidence intervals for predictors in the principal model of interest. For this analysis we used all

287   sites monitored which had a complete set of predictors, with the exception of a single site where

288   the community collect water from a river with a catchment size of 9768km^2, meaning this site is

289   qualitatively different from all other sites (water systems N=124). For predictor variable selection

290   we first removed closely correlated predictors and then classified variables we considered likely

291   to be important in predicting *E. coli* concentration (Table 1). We associated predictor data relating

292   to intake features with both intakes and their respective associated taps. We then produced

293   GLMMs for all purely biophysical traits of sites, while also including all 2-way and 3-way

294   interactions between temperature, pH, and salinity (Table S4 a). To determine intervention

295   effectiveness, we then added features that related directly to the intervention (cattle access, whether

296   the site was in a level 1 *Watershared* agreement, and faeces presence) and again conducted model
297   selection based on AIC minimization (See Table S4 b).

298   We used two approaches to explore whether the implementation of *Watershared* in treatment
299   communities resulted in an improvement in water quality relative to control communities. In the
300   first approach (evaluating the difference-in-differences from 2010 and 2015 between treatment
301   and control sites), we only included the subset of sites where the sample locations remained the
302   same between 2010 and 2015 and the intake is unambiguously associated with a treatment or
303   control community (site N=83, communities N=47). While we use the term difference-in-
304   differences analysis to refer to this analysis, we note that we do not use the standard difference-in-
305   differences estimator, but take a panel approach and include the baseline outcome value as a
306   covariate (Gelman & Hill, 2007). This approach is considered more appropriate especially if
307   autocorrelation is low and has greater power (McKenzie, 2012). As the balance check suggested
308   that the reduction in sites had resulted in a lack of balance in pre-existing contamination we
309   performed a matched analysis. Sites were matched on the baseline *E. coli* measure using genetic
310   matching in the R MatchIt package (Ho et al., 2011). Using the weighting from this matched data
311   we then estimated a differences-in-differences GLMM with *E. coli* concentration in 2015 as the
312   response variable and site treatment status (whether a site is in a control or treatment community)
313   and 2010 *E. coli* concentration as potential predictors. Retention of the 2010 *E. coli* count
314   controlled for the different pre-existing levels of contamination within the matched data set. We
315   also included an interaction term between 2010 *E. coli* concentration and site treatment status (if
316   this interaction were a significant predictor, this would represent a significant effect of the
317   intervention on water quality). Given the different volume of water sampled in 2010 (5ml) and
318   2015 (20ml), we included an offset term of $\log_e(4)$ in each of the models to ensure equivalence
319   between 2010 and 2015 *E. coli* CFU counts. We also included water system as a random effect to
320   account for the tap and intake measures. This model was fitted using the glmmTMB package in R
321   (Brooks et al., 2017) as glmmADMB does not accommodate weights. To account for taps and
322   intakes being connected we included water system as a random effect.

323   In the second approach (evaluating at end-line only) *E. coli* concentrations in 2015 for all sites
324   (site N=228, community N=116) measured were used (regardless of whether we had baseline
325   water quality values measured at the identical location). Again the model was fitted in glmmTMB

326   with site treatment status as the predictor of interest but also including the water source (spring or

327   stream) and data collection point (tap or intake) to control for any differences in these. For both

328   analyses we excluded two sites where water was collected from roofs rather than streams or

329   springs. We accounted for some communities having more than one water system measured by

330   including water system and community as a random effect.

331   To establish whether in practice land use differs between treatment and control communities, we

332   also determined (for all intakes monitored in 2015 for which data on cattle access is available)

333   whether relative proportions of intake sites protected from cattle differed between the treatment

334   and control communities. We tested for a significant difference using a chi-squared test.

335   **Results**

336   ***Cattle faeces in water is one of the significant predictors of E. coli concentration at the local***

337   ***scale***

338   *E. coli* concentration in 2015 (water systems N=124) is significantly predicted by a number of

339   variables (Fig. 4). The details of model selection can be seen in Table S4a (for purely biophysical

340   model selection) and Table S4b for model selection including parameters relating directly to the

341   intervention. Intakes are significantly more contaminated than taps, sites associated with stream

342   intakes are significantly more contaminated than sites associated with spring intakes, and turbidity

343   and disturbance of the sediment by the research team during sampling are both also associated

344   with higher recorded contamination. In terms of variables directly connected to the intervention,

345   the presence of cattle faeces in or close to the water is a significant predictor of contamination.

346   Although faeces presence in the wider forest shows a positive trend, it is not significant at 95% CI.

347   Details of the model can be found in Table S4c.

348   ***The intervention had no significant effect on E. coli concentration at the landscape scale***

349   We analyze the RCT in two ways. In the first (difference-in-differences; Fig. 5) we used weights

350   derived from genetic matching the communities based on the *E. coli* count in 2010 and only the

351   sub-sample of communities where we are certain water sampling locations are the same in 2010

352   and 2015 (sites N=83, communities N=47). The weighting strongly downweighted 2 communities

353   (w=0.19), moderately upweighted 9 (w=1.19), strongly upweighted 1 (w=1.5) and did not change

354   the relative weighting of the remaining communities. We include *E. coli* count in 2010 as a

355 predictor to control for sites with naturally higher contamination (e.g. streams rather than springs).

356 Results from a negative binomial GLM show no significant effect of a site being in a control or

357 treatment community on *E. coli* concentration in 2015. This is because there is no significant

358 interaction between RCT status and *E. coli* concentration in 2010, meaning that the rate of change

359 in *E. coli* concentration between 2010 and 2015 is not significantly different in sites associated

360 with treatment or control communities.

361 In the second analysis we analyse *E. coli* at end-line only (levels in 2015) again using a negative

362 binomial GLM (sites N=228, communities N=116). As we do not have baseline *E. coli* counts for

363 all these sites we include important predictors of water quality from Fig. 4. The results show that

364 while taps had significantly lower contamination than intakes and springs lower contamination

365 than streams there was no significant effect of treatment (Fig. 6). Taken together both analyses

366 together show that, using the robust RCT design and both the baseline and end-line datasets, we

367 did not find a significant effect of the intervention on *E. coli* concentration at the landscape scale.

368 ***Uptake was low, highly variable and treatment and control communities do not differ with***

369 ***respect to protection of intakes from cattle.***

370 In treatment communities (where every household was offered the opportunity to enrol land) a low

371 proportion of land eligible to be enrolled in *Watershared* agreements was actually enrolled, and

372 this proportion was highly variable between communities (Fig 7: range from 0-18% with a median

373 uptake of 2.5%). There is no significant difference (N=129; p=0.97; chi-squared test) between the

374 number of intakes protected from cattle in control and treatment community sites. Water intakes

375 in both control and treatment communities tend to be protected from cattle (61% and 62% of

376 intakes respectively) despite only a quarter of intakes being in compliant level 1 areas (Table 2).

377 **Discussion**

378 **Did *Watershared* improve water quality?**

379 We show that presence of cattle faeces in water or on the stream banks results in higher *E. coli*

380 contamination at individual sites. This suggests that excluding cattle from water sources (one of

381 the key actions *Watershared* seeks to incentivize) can contribute to improving water quality. This

382 should perhaps not be surprising given that fresh cattle faeces can have more than $10^8 \, \text{kg}^1$ colony

383 forming units (Weaver, Entry & Graves, 2005). However, the presence of cattle faeces is only one

384   predictor of water quality. Intakes fed by streams were much more contaminated than those fed by

385   springs, as could have been predicted (Howell, Coyne & Cornelius, 1995; WaterAid, 2013).

386   Indeed, the Millennium Development Goal definition of an improved water source allows some

387   springs to be considered improved without further chemical treatment while no stream or river

388   intakes can (Bain et al., 2014b). We also found that intakes were more heavily contaminated than

389   taps. This suggests that although the sedimentation and filtration chambers in many of the water

390   systems may not always be effective, they have at least some positive effect on water quality. It is

391   unsurprising that turbidity was an important predictor as this is a well-known as a predictor of *E.

392   coli* contamination (LeChevallier, Evans & Seidler, 1981).

393   Despite finding evidence that excluding cattle from riparian forests had a significant impact on

394   water quality at the local scale, both the RCT analyses (comparing *E. coli* contamination between

395   treatment and control communities) showed no evidence of an effect at the landscape scale.

396   There is therefore an apparent paradox in that the intervention incentivised (excluding cattle) does

397   apparently improve water quality but the program has not had an impact at the landscape scale. It

398   is possible that there has not been sufficient time between implementation of the intervention and

399   subsequent evaluation for differences to become apparent. Some of the *Watershared* areas

400   included in the analysis had been enrolled in the latter part of 2014, hence in some cases only a

401   few months before the end-line monitoring was undertaken. It is well known that *E. coli* can persist

402   in freshwater sediments for long periods of time (Pachepsky & Shelton, 2011; Cho et al., 2016).

403   However, we argue that features of the way in which the PES program was implemented in the

404   area meant that it was highly likely that no difference would be detected between control and

405   treatment.

406   Firstly, a very low percentage of eligible land was enrolled in *Watershared* agreements (the median

407   uptake of enrolled land was 2.5% of eligible land in treatment communities). In addition, GIS

408   analysis has shown that much of the land which was enrolled could not influence water quality at

409   the intake or tap because it was situated in a different catchment (Pynegar 2018). Livestock-derived

410   *E. coli* can enter water intakes through a number of routes including overland flow and movement

411   of groundwater (Oliver et al., 2010), and not solely through direct deposition which is what the

412   *Watershared* agreements try to prevent. The small areas conserved at or above the intakes may

413   well have reduced faeces presence and so reduced *E. coli* concentration at these sites. However,

414 upstream or uphill of these intakes contamination may have continued to enter water bodies
415 through multiple routes. Instructively, evidence from a 26-year-old conservation area in the
416 community of La Aguada, near to our study area, shows that despite the 20% of the catchment
417 nearest to the intake being under conservation with cattle excluded (R. Rueda, pers. comm.), the
418 water remains contaminated (Pynegar 2018).

419 Secondly, the iteration of the *Watershared* intervention that we studied did not oblige, or even
420 provide extra incentives, for landowners to conserve land surrounding or in the same catchment as
421 monitored intakes. Farmers were free to enrol any land which met the criteria (forest within 100m
422 of a stream or spring). The intervention was not spatially targeted towards areas critical for
423 community water supplies, and in fact only 16 of the 68 water intakes in treatment communities
424 were located inside enrolled and compliant parcels of land. Also, many communities in the study
425 area had previously excluded cattle from water intakes independently of the *Watershared* program,
426 meaning there is no significant difference in the proportion of intakes protected from cattle
427 between treatment and control communities.

428 Finally, not all the land enrolled in *Watershared* agreements will represent additional conservation.
429 There is a large body of evidence showing that adverse participant selection in PES programs
430 means that a significant proportion of land enrolled would have been conserved in the absence of
431 the incentives (Börner et al., 2017). There has been concern about the extent to which conservation
432 funded under the *Watershared* programme will represent additional conservation since the early
433 days of the scheme (Robertson & Wunder, 2005; Asquith, Vargas & Wunder, 2008). Best
434 estimates for the *Watershared* intervention suggest that only about one third of level 1 agreements
435 have resulted in cattle being excluded from land which they would have otherwise been using
436 (Bottazzi et al., 2018).

437 Given the very low uptake of *Watershared* agreements, the lack of targeting of the land enrolled,
438 and the fact that (like in any PES programme) not all the land enrolled represents additional
439 conservation, it is perhaps not surprising that we did not detect an impact of the intervention on
440 water quality at the landscape scale.

441 **How would *Watershared* have to change to result in a significant impact on water quality?**

442   Some of the reasons why *Watershared* did not produce landscape-scale impacts on water quality
443   relate to commonly recognised issues in PES implementation. First, the link between the land use
444   incentivised (the proxy) and the ecosystem service desired is often weak and poorly understood
445   (Jack, Kousky and Sims 2008). In the case of *Watershared*, it is unclear how much land in a
446   catchment would need to be protected, where, and over what timescale, to obtain a significant
447   improvement in water quality at the landscape or even the catchment scale. Second, the marginal
448   benefits from service provision (or in this case the land use proxy for service provision) are highly
449   spatially heterogeneous. Land enrolled directly upstream of intakes will probably have an effect
450   on monitored water quality while areas under conservation elsewhere (for example below the water
451   intake) obviously cannot. In such cases, spatial targeting and differentiated payments would likely
452   increase program efficiency (Ezzine-de-Blas et al., 2016).

453   Both theory (Persson & Alpízar, 2013) and empirical research on PES programmes (Arriagada,
454   Sills, Pattanayak, & Ferraro, 2009) suggest that low levels of payments result in low uptake. It thus
455   seems likely that higher *Watershared* payments would have ensured that a higher proportion of
456   eligible land was enrolled. It is difficult to directly compare payments in *Watershared* with those
457   in related programmes both because of the payment structure (in *Watershared* participants are paid
458   an enrolling fee plus a per hectare payment) and because a dollar is worth more in some countries
459   than others. However the per-hectare payments in *Watershared* is certainly lower than other
460   payments for watershed services-type programs in Latin America: for example Mexico's PSA-H
461   program pays 27 USD/hectare/year for primary forest and 36 USD/ha/year for cloud forest
462   (Muñoz-Piña *et al.* 2008) while Costa Rica's national PES pays 45 to 163 USD/hectare/year
463   (Wunder, Engel & Pagiola 2008). The Ugandan PES program analysed in the RCT by
464   Jayachandran *et al.* (2017) paid landowners 28 USD per year per hectare of forest. Bottazzi *et al.*
465   (2018) show that those signing *Watershared* agreements have multiple motivations for doing so.
466   To ensure that areas most valuable in terms of their potential ecosystem service provision were
467   enrolled would require targeting. However such targeting would increase the complexity and
468   transaction costs of the program (Jack, Kousky & Sims, 2008) and poses potential issues in terms
469   of perceived fairness locally.

470   There is also likely to be a limit to the impact that livestock exclusion can achieve, and this will
471   depend on the extent to which faecal contamination derives from other sources such as wildlife,

472   inadequate sanitation infrastructure, spreading of manure on agricultural land, or from open

473   defecation. Those involved in promoting similar interventions should check the extent to which

474   cattle contamination is indeed the driver of microbial water quality issues in the region, perhaps

475   using genetic testing of *E. coli* (Carson et al., 2001). Context-appropriate engineering solutions,

476   such as protection of springs used for drinking water (Kremer et al., 2011), use of springs rather

477   than streams as drinking water sources, construction of filtration systems, or introduction of

478   household-level interventions (Clasen et al., 2007), may be more effective at improving water

479   quality than livestock exclusions. Such solutions however do not provide the desired co-benefits

480   of the intervention, such as forest carbon storage, biodiversity conservation, and increases in local

481   incomes. Future work may aim to combine both conservation and engineering solutions and

482   involve more direct conservation actions such as purchase or rent of particularly sensitive or

483   important catchments.

484   ***Did the RCT enable robust evaluation of the efficacy of the Watershared program?***

485   Uptake of *Watershared* agreements was very low (the median % enrolment of eligible land in

486   treatment communities was just 2.5%). There was also no targeting of the land enrolled meaning

487   that little of the land enrolled could impact water quality. Although 24% of the water intakes in

488   treatment communities were in compliant level 1 *Watershared* agreements, there is no difference

489   in the percentage of water intakes protected from cattle in treatment and control communities. This

490   is because communities protect their water intakes for many reasons and it appears that the

491   intervention was not significantly effective at increasing this level of protection. Therefore, the

492   conclusion we draw from the RCT analysis about the efficacy of *Watershared* as delivered in this

493   area at influencing water quality at the landscape scale, could equally have been drawn without a

494   large-scale experiment conducted over 5 years. Theory-based impact evaluation (mapping out the

495   causal chain from inputs to outcomes and testing the underlying assumptions; White, 2009) could

496   have been just as effective. The RCT in this case therefore added little to evaluating the impact of

497   this intervention.

498   Despite the challenges of implementing this RCT, and observation that the same conclusions about

499   programme impact could probably have been made without the RCT, we remain positive about

500   the potential value of RCTs in evaluating the impact of landscape-scale conservation. There is

501   certainly demand from policy makers and funders for high quality causal inference about the

502   efficacy of a programme and this is only going to increase. We therefore suggest that those
503   interested in quality impact evaluation of landscape-scale conservation interventions learn from
504   the experience of the *Watershared* RCT.

505   ***Lessons learnt from the experience of the Watershared RCT for future evaluations of landscape-***
506   ***scale conservation interventions***

507   *The randomisation unit needs to be carefully selected*: The choice of randomisation unit in an RCT
508   will influence spillover; the phenomenon in which treatment may affect outcomes in non-treated
509   units resulting in an erroneously low estimate of treatment effect size (Glennerster & Takavarasha,
510   2013). In the case of the *Watershared* RCT, randomization was at the level of the community.
511   However for evaluating the impact of the program on water quality, a more appropriate
512   randomization unit would have been the catchment above the intake supplying water to a
513   community. The selection of communities as the randomisation unit in this RCT was problematic
514   because catchments above treatment community water intakes could fall partially within
515   neighbouring treatment communities (or vice versa); potentially resulting in spillover of effects
516   between treatment and control communities (Pynegar 2018). However *Watershared* was not
517   designed only to impact water quality but also to reduce deforestation, improve forest biodiversity
518   and provide socio-economic benefits (Asquith, 2016). Designing an RCT to evaluate the impact
519   of an intervention on multiple outcomes is inevitably challenging (Pynegar 2018). A future RCT
520   of landscape-scale environmental management interventions should ensure that the randomisation
521   unit is selected to minimise the risk of spillover. This may mean an RCT can only effectively
522   evaluate the impact of one or two outcomes at once.

523   *There are difficult decisions to make in selecting sites at which to monitor outcomes*: One of the
524   aims of *Watershared* was to improve the quality of water consumed in communities. The decision
525   was therefore made to measure water quality at the intake and tap serving the largest number of
526   households in each community. Unfortunately, in a significant number of communities, the water
527   intake serving the community had changed over the 5 years of the intervention. At such sites, if
528   end-line data were to be taken from the water system serving the majority of the community, the
529   end-line and baseline data were inevitably not at the same location. These sites thus had to be
530   excluded from the difference-in-differences analysis. By conducting the analysis in two ways (our
531   difference-in-differences analysis for the sub-set of communities where the intake was comparable

532 across the period and using end-line data only from the full set of communities) we were able to

533 use as much of the data as possible.

534 *An RCT is not appropriate unless the intervention is well developed*: The substantial investment

535 of time and resources in an RCT means that it is only appropriate when implementers are confident

536 that they have an intervention whose functioning is reasonably well developed (Pattanayak, 2009;

537 Cartwright, 2010). At earlier stages of developing an intervention, formative rather than

538 summative impact evaluation processes might be more appropriate (Rossi, Lipsey & Freeman,

539 2004). The implementers of *Watershared* had experimented with versions of the program before

540 implementing the RCT. However in the event, uptake of *Watershared* was so low that an RCT was

541 not an appropriate method of evaluating its impact on water quality at a landscape scale. Precise

542 levels of uptake cannot be known until the experiment is already set up and the intervention

543 offered, but pilot work (or analysis of the GIS data showing the location of enrolled agreements)

544 might have revealed this low uptake and efforts been made to increase it (by raising the payments

545 offered).

546 *Effort should be made to ensure balance between treatment and control in outcomes of interest:*

547 In the *Watershared* RCT, the allocation to treatment and control achieved quite good balance in

548 the variables likely to influence *E. coli* contamination (although when only the subset of sites for

549 which there is a comparable end-line is considered a significant imbalance in baseline values for

550 *E. coli* contamination between treatment and control was introduced). Of course imperfect balance

551 is a common problem in RCTs and is not a barrier to valid inference (Senn, 2013). We were able

552 to account for the imbalance in our difference-in-differences analysis by matching on baseline *E.*

553 *coli* values at each site. However to reduce the chance of random allocation to control and

554 treatment resulting in an imbalance in important variables, baseline information on the outcome of

555 interest (in this case *E. coli* contamination) should be included in the stratification.

556 *Care needs to be taken to avoid contamination of the control:* In an ideal RCT, there is zero uptake

557 of the intervention in the control group and consistent and high uptake in the treatment community.

558 Unfortunately, in many situations the control is contaminated by information from treatment

559 participants spreading to controls (Torgerson, 2001). In the *Watershared* RCT there was non-zero

560 uptake among the control communities. In this case, the contamination of the control was due to

561 people living in treatment communities but owning land in treatment communities which they

562   chose to enrol. This was not noticed during the roll out of the program for two reasons. Firstly, the

563   community boundaries were not available until they were generated (with substantial investment)

564   by our research team in 2017. Secondly, for the technicians involved in day-to-day implementation

565   of *Watershared*, the detail of where enrolled land occurred relative to treatment and control

566   communities was not a priority. In future, we suggest that those establishing an RCT to evaluate

567   the impact of environmental management interventions at scale ensure they have sufficient

568   research capacity during the implementation phase to monitor the RCT while it is in progress so

569   issues such as this could be avoided.

570   *Blinding is unlikely to be fully possible in land-scale scale conservation interventions:* Double

571   blinding is considered best practice in RCTs so neither the researcher nor the participants know

572   who has been assigned to the treatment or control group (Glennerster & Takavarasha, 2013). In

573   the case of the *Watershared* RCT the researchers scoring the *E. coli* contamination did not know

574   whether the sample came from a treatment of control community; so blinding was achieved there.

575   However the participants clearly knew they were in a control or treatment community. Those

576   allocated to control or treatment may have different expectations or show different behaviour or

577   effort simply as a consequence of being allocated to a control or treatment group (Chassang, Padró

578   i Miquel & Snowberg, 2012). Some authors have claimed that these behavioural effects may be

579   large (Bulte et al., 2014) but they have not been extensively studied. They should be considered in

580   the design of any landscape-scale RCT.

581   **Conclusions**

582   There is global interest in PES because it is seen as an efficient way to provide environmental

583   outcomes. The effectiveness of PES in achieving its intended outcomes is fundamentally an

584   empirical question but the quality of the evidence base concerning the delivery of benefits from

585   PES is mixed. There is therefore substantial interest in robust evaluation of the effectiveness of

586   PES programs at delivering outcomes. We conclude that this particular program would require

587   much greater uptake (probably requiring higher payments) and more intensive targeting (which

588   would increase substantially the transaction costs and design complexity of the intervention) to

589   have a significant impact on water quality. However, although this paper presents the results of a

590   Randomised Control Trial (one of the very few implemented to evaluate the impacts of a

591   conservation intervention at scale), these same conclusions could have been drawn without the

592   RCT. The low uptake of the program and the lack of a difference in water intakes protected from

593   cattle between control and treatment communities make the result of the RCT (no effect of the

594   program detected) inevitable. Randomised Control Trials have the potential to contribute to

595   building the evidence base for understanding the impact of environmental management approaches

596   such as Payments for Ecosystem Services. However, as evidenced by the *Watershared* experience,

597   they are not straightforward to implement in practice. We hope that by publishing the experience

598   of the *Watershared* RCT we will encourage future landscape-scale conservation impact

599   evaluations to improve on the use of this evaluation approach in conservation.

600   **Acknowledgements**

601   We thank colleagues at Fundación Natura Bolivia for all of their assistance, contributions and

602   ideas, especially María Teresa Vargas, Tito Vidaurre, Maximo García, Dionicio Toledo, Hugo

603   Vallejos, Miler Guzmán, Antonio Daza, Denis Calderón, Huascar Azurduy, Veronica Chávez and

604   Victoria Aguilera. We are very grateful to Kelsey Jack for initial randomisation and to Dave

605   Chadwick, Prysor Williams, Patrick Bottazzi, Robert Rueda and David Crespo for valued

606   discussion. Sven Wunder and two anonymous reviewers gave very helpful feedback during the

607   review process which substantially improved both the analysis and the framing of this paper. We

608   also thank members of the communities in which we conducted monitoring for their support and

609   permission.

610   ***References***

611   Ahmed W., Hodgers L., Sidhu JPS., Toze S. 2012. Fecal indicators and zoonotic pathogens in

612   household drinking water taps fed from rainwater tanks in Southeast Queensland, Australia.

613   *Applied and Environmental Microbiology* 78:219–226. DOI: 10.1128/AEM.06554-11.

614   Anonymous. 2016.Establishment of buffer strips along watercourses. *Available at*

615   *https://www.gov.uk/guidance/guide-to-cross-compliance-in-england-2016/gaec-1-*

616   *establishment-of-buffer-strips-along-watercourses* (accessed August 3, 2016).

617   Arriagada RA., Sills EO., Pattanayak SK., Ferraro PJ. 2009. Combining qualitative and

618   quantitative methods to evaluate participation in Costa Rica's program of payments for

619   environmental services. *Journal of Sustainable Forestry* 28:343–367.

620 Ashbolt N., Grabow W., Snozzi M. 2001. Indicators of microbial water quality. In: *Water*
621   *Quality: Guidelines, Standards and Health*. 289–316. DOI: 10.4324/9781315693606.

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

624 Asquith NM., Vargas MT., Wunder S. 2008. Selling two environmental services: In-kind
625   payments for bird habitat and watershed protection in Los Negros, Bolivia. *Ecological*
626   *Economics* 65:675–684. DOI: 10.1016/J.ECOLECON.2007.12.014.

627 Bain R., Cronk R., Hossain R., Bonjour S., Onda K., Wright J., Yang H., Slaymaker T., Hunter
628   P., Prüss-Ustün A., Bartram J. 2014a. Global assessment of exposure to faecal
629   contamination through drinking water based on a systematic review. *Tropical Medicine &*
630   *International Health* 19:917–927. DOI: 10.1111/tmi.12334.

631 Bain R., Cronk R., Wright J., Yang H., Slaymaker T., Bartram J. 2014b. Fecal Contamination of
632   Drinking-Water in Low- and Middle-Income Countries: A Systematic Review and Meta-
633   Analysis. *PLoS Medicine* 11:e1001644. DOI: 10.1371/journal.pmed.1001644.

634 Baylis K., Honey-Rosés J., Börner J., Corbera E., Ezzine-de-Blas D., Ferraro PJ., Lapeyre R.,
635   Persson UM., Pfaff A., Wunder S. 2016. Mainstreaming Impact Evaluation in Nature
636   Conservation. *Conservation Letters* 9:58–64. DOI: 10.1111/conl.12180.

637 Börner J., Baylis K., Corbera E., Ezzine-de-Blas D., Ferraro PJ., Honey-Rosés J., Lapeyre R.,
638   Persson UM., Wunder S. 2016. Emerging Evidence on the Effectiveness of Tropical Forest
639   Conservation. *PLOS ONE* 11:e0159152. DOI: 10.1371/journal.pone.0159152.

640 Börner J., Baylis K., Corbera E., Ezzine-de-Blas D., Honey-Rosés J., Persson UM., Wunder S.
641   2017. The Effectiveness of Payments for Environmental Services. *World Development*
642   96:359–374. DOI: 10.1016/j.worlddev.2017.03.020.

643 Bottazzi P., Vargas Rios MT., Asquith N., Crespo D., Vidaurre T., García M., Miler Guzmán R.,
644   Gonzales L., Vallejos H., Pérez Guzman B., Avila Becerra F., Toledo D., Jones JPG. 2017.
645   Baseline and endline socio-economic data from a Randomised Control Trial of the
646   Watershared intervention in the Bolivian Andes. *ReShare UK Data Archive*. DOI:
647   10.5255/UKDA-SN-852623.

648   Bottazzi P., Wiik E., Crespo D., Jones JPG. 2018. Payment for Environmental "Self-Service":
649   Exploring the Links Between Farmers' Motivation and Additionality in a Conservation
650   Incentive Programme in the Bolivian Andes. *Ecological Economics* 150:11–23. DOI:
651   10.1016/j.ecolecon.2018.03.032.

652   Bowler DE., Buyung-Ali LM., Healey JR., Jones JPG., Knight TM., Pullin AS. 2012. Does
653   community forest Management provide global environmental benefits and improve local
654   welfare? *Frontiers in Ecology and the Environment* 10:29–36. DOI: 10.1890/110040.

655   Brooks ME., Kristensen K., van Benthem KJ., Magnusson A., Berg CW., Nielsen A., Skaug HJ.,
656   Mächler M., Bolker BM. 2017. The R journal. *The R Journal* 9:378–400.

657   Bruijnzeel LA. 2004. Hydrological functions of tropical forests: not seeing the soil for the trees?
658   *Agriculture, Ecosystems & Environment* 104:185–228. DOI:
659   http://dx.doi.org/10.1016/j.agee.2004.01.015.

660   Bulte E., Beekman G., Di Falco S., Hella J., Lei P. 2014. Behavioral Responses and the Impact
661   of New Agricultural Technologies: Evidence from a Double-blind Field Experiment in
662   Tanzania. *American Journal of Agricultural Economics* 96:813–830. DOI:
663   10.1093/ajae/aau015.

664   Calvet-Mir L., Corbera E., Martin A., Fisher J., Gross-Camp N. 2015. Payments for ecosystem
665   services in the tropics: A closer look at effectiveness and equity. *Current Opinion in
666   Environmental Sustainability* 14:150–162. DOI: 10.1016/j.cosust.2015.06.001.

667   Carson CA., Shear BL., Ellersieck MR., Asfaw A. 2001. Identification of Fecal Escherichia coli
668   from Humans and Animals by Ribotyping. *Applied and Environmental Microbiology*
669   67:1503–1507. DOI: 10.1128/AEM.67.4.1503-1507.2001.

670   Cartwright N. 2010. What are randomised controlled trials good for? *Philosophical Studies*
671   147:59–70. DOI: 10.1007/s11098-009-9450-2.

672   Chassang S., Padró i Miquel G., Snowberg E. 2012. Selective Trials: A Principal-Agent
673   Approach to Randomized Controlled Experiments. *American Economic Review* 102:1279–
674   1309. DOI: 10.1257/aer.102.4.1279.

675   Cho KH., Pachepsky YA., Oliver DM., Muirhead RW., Park Y., Quilliam RS., Shelton DR.

676   2016. Modeling fate and transport of fecally-derived microorganisms at the watershed scale:

677   State of the science and future opportunities. *Water Research* 100:38–56. DOI:

678   10.1016/j.watres.2016.04.064.

679   Chuang P., Trottier S., Murcott S. 2011. Comparison and verification of four field-based

680   microbiological tests: H2S test, Easygel®, Colilert®, Petrifilm™. *Journal of Water*

681   *Sanitation and Hygiene for Development* 1:68–85.

682   Clasen T., Cairncross S., Haller L., Bartram J., Walker D. 2007. Cost-effectiveness of water

683   quality interventions for preventing diarrhoeal disease in developing countries. *Journal of*

684   *Water and Health* 5:599. DOI: 10.2166/wh.2007.010.

685   Council of Economic Advisers. 2014. Evaluation as a tool for improving federal programs. In:

686   *Economic Report of the President, Together with the Annual Report of the Council of*

687   *Economic Advisors*. Washington DC: U.S. Government Printing Office, 269–298.

688   Crane SR., Moore JA., Grismer Jr. ME., Milner JR. 1983. Bacterial pollution from agricultural

689   sources: A review. *Transactions of the American Society of Agricultural Engineers* 26:858–

690   866.

691   Dasgupta S. 2017. Experience or evidence: How do big conservation NGOs make decisions?

692   *Mongabay*.

693   Dasgupta S., Gaworecki M., Burivalova Z. 2018. Conservation Effectiveness. *Mongabay*.

694   Duchelle A., Wunder S., Martius C. 2018. Understanding what works in forest-based mitigation.

695   *Green Climate Fund Independent Evaluation Unit*.

696   Ezzine-de-Blas D., Wunder S., Ruiz-Pérez M., Moreno-Sanchez R del P., Nikolakis W., Wilson

697   P. 2016. Global Patterns in the Implementation of Payments for Environmental Services.

698   *PLOS ONE* 11:e0149847. DOI: 10.1371/journal.pone.0149847.

699   Ferraro PJ. 2017. Are payments for ecosystem services benefiting ecosystems and people? In:

700   Kareiva P, Marvier M, Silliman B eds. *Effective Conservation Science: Data Not Dogma*.

701   New York, NY: Oxford University Press, 159–166. DOI:

702   10.1093/oso/9780198808978.001.0001.

703 Ferraro PJ., Pattanayak SK. 2006. Money for Nothing? A Call for Empirical Evaluation of
704   Biodiversity Conservation Investments. *PLoS Biol* 4:e105. DOI:
705   10.1371/journal.pbio.0040105.

706 Fournier DA., Skaug HJ., Ancheta J., Ianelli J., Magnusson A., Maunder MN., Nielsen A., Sibert
707   J. 2012. AD Model Builder: using automatic differentiation for statistical inference of
708   highly parameterized complex nonlinear models. *Optimization Methods and Software*
709   27:233–249. DOI: 10.1080/10556788.2011.597854.

710 Gelman A., Hill J. 2007. *Data analysis using regression and multilevel/hierarchical models*.
711   Cambridge University Press.

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

714 Greenstone M., Gayer T. 2009. Quasi-experimental and experimental approaches to
715   environmental economics. *Journal of Environmental Economics and Management* 57:21–
716   44. DOI: 10.1016/j.jeem.2008.02.004.

717 Greifer N. 2018. cobalt: Covariate Balance Tables and Plots. In: *R package version 3.3.0*.

718 Grima N., Singh SJ., Smetschka B., Ringhofer L. 2016. Payment for Ecosystem Services (PES)
719   in Latin America: Analysing the performance of 40 case studies. *Ecosystem Services* 17:24–
720   32. DOI: 10.1016/J.ECOSER.2015.11.010.

721 Ho DE., Imai K., King G., Stuart EA. 2011. MatchIt: Nonparametric Preprocessing for
722   Parametric Causal Inference. *JSS Journal of Statistical Software* 42.

723 Howell JM., Coyne MS., Cornelius P. 1995. Fecal Bacteria in Agricultural Waters of the
724   Bluegrass Region of Kentucky. *Journal of Environment Quality* 24:411. DOI:
725   10.2134/jeq1995.00472425002400030003x.

726 Jack BK., Kousky C., Sims KRE. 2008. Designing payments for ecosystem services: Lessons
727   from previous experience with incentive-based mechanisms. *Proceedings of the National*
728   *Academy of Sciences* 105:9465–9470. DOI: 10.1073/pnas.0705503104.

729 Jayachandran S., de Laat J., Lambin EF., Stanton CY., Audy R., Thomas NE. 2017. Cash for

730   carbon: A randomized trial of payments for ecosystem services to reduce deforestation.

731   *Science* 357:267–273. DOI: 10.1126/science.aan0568.

732   Jüni P., Altman DG., Egger M. 2001. Assessing the quality of controlled clinical trials. *BMJ* 323.

733   Keeler BL., Polasky S., Brauman KA., Johnson KA., Finlay JC., O'Neill A., Kovacs K., Dalzell

734   B. 2012. Linking water quality and well-being for improved assessment and valuation of

735   ecosystem services. *Proceedings of the National Academy of Sciences of the United States*

736   *of America* 109:18619–18624. DOI: 10.1073/pnas.1215991109.

737   Kremer M., Leino J., Miguel E., Zwane AP. 2011. Spring cleaning: Rural water impacts,

738   valuation, and property rights institutions. *Quarterly Journal of Economics* 126:145–205.

739   DOI: 10.1093/qje/qjq010.

740   LeChevallier MW., Evans TM., Seidler RJ. 1981. Effect of turbidity on chlorination efficiency

741   and bacterial persistence in drinking water. *Applied and Environmental Microbiology*

742   42:159–167.

743   Leclerc H., Mossel D a., Edberg SC., Struijk CB. 2001. Advances in the bacteriology of the

744   coliform group: their suitability as markers of microbial water safety. *Annual review of*

745   *microbiology* 55:201–234. DOI: 10.1146/annurev.micro.55.1.201.

746   Martin-Ortega J., Ojea E., Roux C. 2013. Payments for Water Ecosystem Services in Latin

747   America: A literature review and conceptual model. *Ecosystem Services* 6:122–132.

748   McKenzie D. 2012. Beyond baseline and follow-up: The case for more T in experiments.

749   *Journal of Development Economics* 99:210–221. DOI: 10.1016/J.JDEVECO.2012.01.002.

750   Micrology Labs. 2016. Detection of Waterborne Coliforms and E. coli with Coliscan Easygel.

751   Miteva DA., Pattanayak SK., Ferraro PJ. 2012. Evaluation of biodiversity policy instruments:

752   What works and what doesn't? *Oxford Review of Economic Policy* 28:69–92. DOI:

753   10.1093/oxrep/grs009.

754   Mooney C. 2016. A cheap, simple experiment just found a very effective way to slow

755   deforestation. *The Washington Post*.

756   Muñoz-Piña C., Guevara A., Torres JM., Braña J. 2008. Paying for the hydrological services of

Mexico's forests: Analysis, negotiations and results. *Ecological Economics* 65:725–736. DOI: 10.1016/j.ecolecon.2007.07.031.

Naeem S., Ingram JC., Varga A., Agardy T., Barten P., Bennett G., Bloomgarden E., Bremer LL., Burkill P., Cattau M. 2015. Get the science right when paying for nature's services. *science* 347:1206–1207.

Oliver DM., Page T., Hodgson CJ., Heathwaite AL., Chadwick DR., Fish RD., Winter M. 2010. Development and testing of a risk indexing framework to determine field-scale critical source areas of faecal bacteria on grassland. *Environmental Modelling and Software* 25:503–512. DOI: 10.1016/j.envsoft.2009.10.003.

Pachepsky YA., Shelton DR. 2011. Escherichia coli and Fecal Coliforms in Freshwater and Estuarine Sediments. *Critical Reviews in Environmental Science and Technology* 41:1067–1110. DOI: 10.1080/10643380903392718.

Pattanayak SK. 2009. *Rough Guide to Impact Evaluation of Environmental and Development Programs*. Kathmandu, Nepal.

Pattanayak SK., Wunder S., Ferraro PJ. 2010. Show me the money: Do payments supply environmental services in developing countries? *Review of Environmental Economics and Policy* 4:254–274. DOI: 10.1093/reep/req006.

Persson UM., Alpízar F. 2013. Conditional Cash Transfers and Payments for Environmental Services—A Conceptual Framework for Explaining and Judging Differences in Outcomes. *World Development* 43:124–137. DOI: 10.1016/j.worlddev.2012.10.006.

Ponette-González AG., Brauman KA., Marín-Spiotta E., Farley KA., Weathers KC., Young KR., Curran LM. 2015. Managing water services in tropical regions: From land cover proxies to hydrologic fluxes. *Ambio* 44:367–375. DOI: 10.1007/s13280-014-0578-8.

Prüss-Ustün A., Bartram J., Clasen T., Colford JM., Cumming O., Curtis V., Bonjour S., Dangour AD., De France J., Fewtrell L., Freeman MC., Gordon B., Hunter PR., Johnston RB., Mathers C., Mäusezahl D., Medlicott K., Neira M., Stocks M., Wolf J., Cairncross S. 2014. Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. *Tropical Medicine &*

785   *International Health* 19:894–905. DOI: 10.1111/tmi.12329.

786   Puri J., Nath M., Bhatia R., Glew L. 2016. *Examining the evidence base for forest conservation*
787   *interventions*. New Delhi.

788   R Development Core Team. 2014. R: A language and environment for statistical computing.

789   Richards S., Paterson E., Withers PJA., Stutter M. 2016. Septic tank discharges as multi-
790   pollutant hotspots in catchments. *Science of the Total Environment* 542:854–863. DOI:
791   10.1016/j.scitotenv.2015.10.160.

792   Robertson N., Wunder S. 2005. *Fresh Tracks in the Forest: Assessing Incipient Payments For*
793   *Enivronmental Services Initiatives in Bolivia*. Jakarta: CIFOR.

794   Rossi P., Lipsey M., Freeman H. 2004. *Evaluation: a Systematic Approach*. Thousand Oaks, CA:
795   SAGE Publications.

796   Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized
797   studies. *Journal of Educational Psychology* 66:688–701. DOI: 10.1037/h0037350.

798   Salzman J., Bennett G., Carroll N., Goldstein A., Jenkins M. 2018. The global status and trends
799   of Payments for Ecosystem Services. *Nature Sustainability* 1:136–144. DOI:
800   10.1038/s41893-018-0033-0.

801   Samii C., Lisiecki M., Kulkarni P., Paler L., Chavis L. 2014. Effects of Payment for
802   Environmental Services (PES) on Deforestation and Poverty in Low and Middle Income
803   Countries: A Systematic Review. *Campbell Systematic Reviews* 10. DOI:
804   10.4073/csr.2014.11.

805   Senn S. 2013. Seven myths of randomisation in clinical trials. *Statistics in Medicine* 32:1439–
806   1450. DOI: 10.1002/sim.5713.

807   Skaug H., Fournier D., Bolker B., Magnusson A., Nielsen A. 2014. Generalized Linear Mixed
808   Models using AD Model Builder.

809   Spears D., Ghosh A., Cumming O. 2013. Open Defecation and Childhood Stunting in India: An
810   Ecological Analysis of New Data from 112 Districts. *PLoS ONE* 8:e73784. DOI:
811   10.1371/journal.pone.0073784.

812   Sunohara MD., Topp E., Wilkes G., Gottschall N., Neumann N., Ruecker N., Jones TH., Edge
813   TA., Marti R., Lapen DR. 2012. Impact of riparian zone protection from cattle on nutrient,
814   bacteria, F-coliphage, and loading of an intermittent stream. *Journal of Environmental*
815   *Quality* 41:1301–1314. DOI: 10.2134/jeq2011.0407.

816   Sutherland WJ., Wordley CFR. 2017. Evidence complacency hampers conservation. *Nature*
817   *Ecology & Evolution* 1:1215–1216. DOI: 10.1038/s41559-017-0244-1.

818   Torgerson DJ. 2001. Contamination in trials: is cluster randomisation the answer? *BMJ* 322.

819   Waeber PO., Wilmé L., Mercier JR., Camara C., Lowry PP. 2016. How effective have thirty
820   years of internationally driven conservation and development efforts been in Madagascar?
821   *PLoS ONE* 11. DOI: 10.1371/journal.pone.0161115.

822   WaterAid. 2013. Gravity-fed Schemes. *WaterAid Technology Briefs*.

823   Weaver RW., Entry JA., Graves A. 2005. Numbers of fecal streptococci and Escherichia coli in
824   fresh and dry cattle, horse, and sheep manure. *Canadian Journal of Microbiology* 51:847–
825   851. DOI: 10.1139/w05-071.

826   White H. 2009. *Theory-Based Impact Evaluation: Principles and Practice*.

827   White H. 2013. An introduction to the use of randomised control trials to evaluate development
828   interventions. *Journal of Development Effectiveness* 5:30–49. DOI:
829   10.1080/19439342.2013.764652.

830   Wunder S. 2015. Revisiting the concept of payments for environmental services. *Ecological*
831   *Economics* 117:234–243.

832   Wunder S., Engel S., Pagiola S. 2008. Taking stock: A comparative analysis of payments for
833   environmental services programs in developed and developing countries. *Ecological*
834   *Economics* 65:834–852.

835

Figure 1

Map of the study area.

a) Locations of the 65 treatment communities (*Watershared* agreements offered) and 64 control communities (*Watershared* agreements not offered) within the Río Grande Valles Cruceños Natural Integrated Management Area (*ANMI RG-VC*). b) Location of the ANMI RG-VC protected area within Bolivia

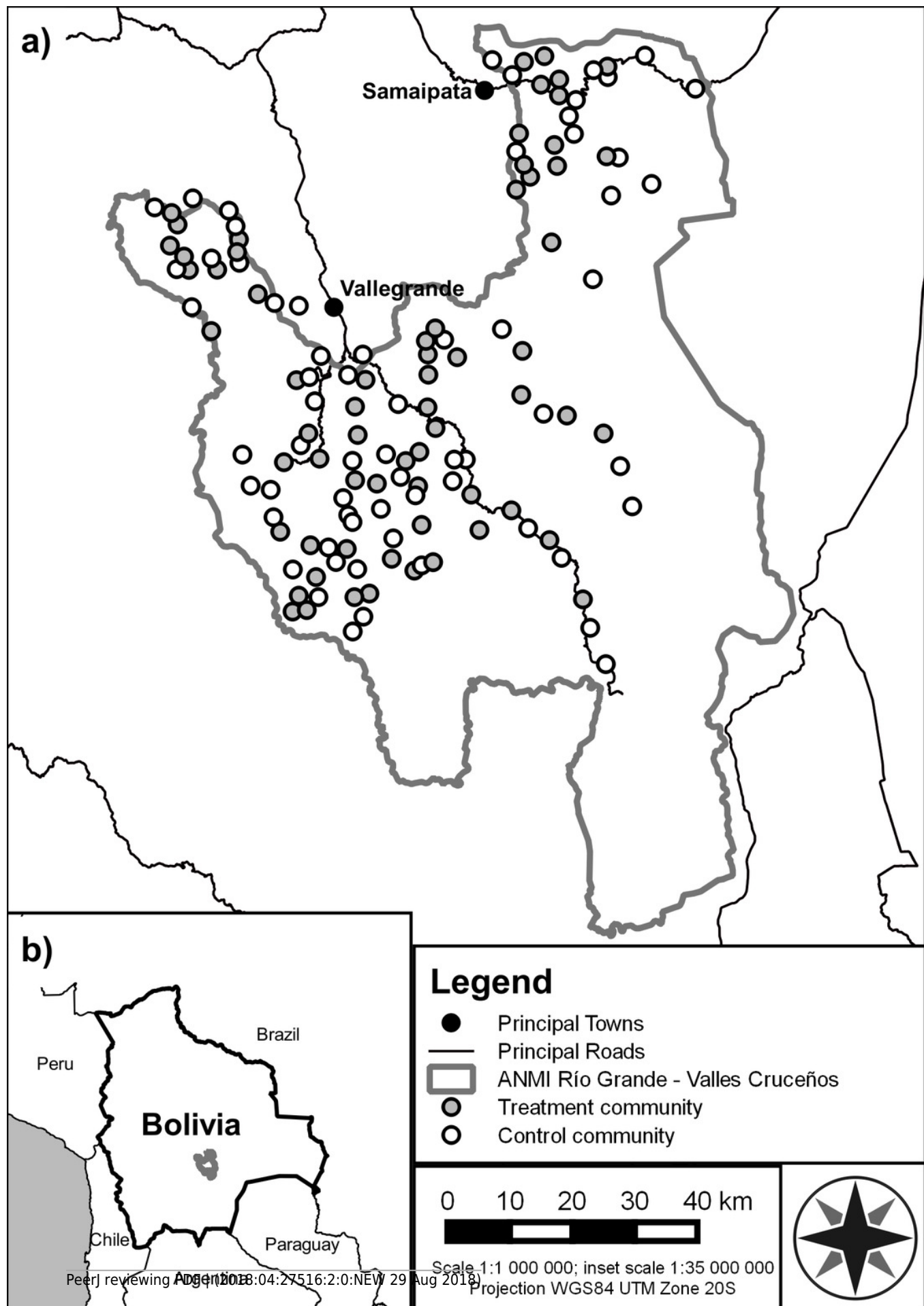
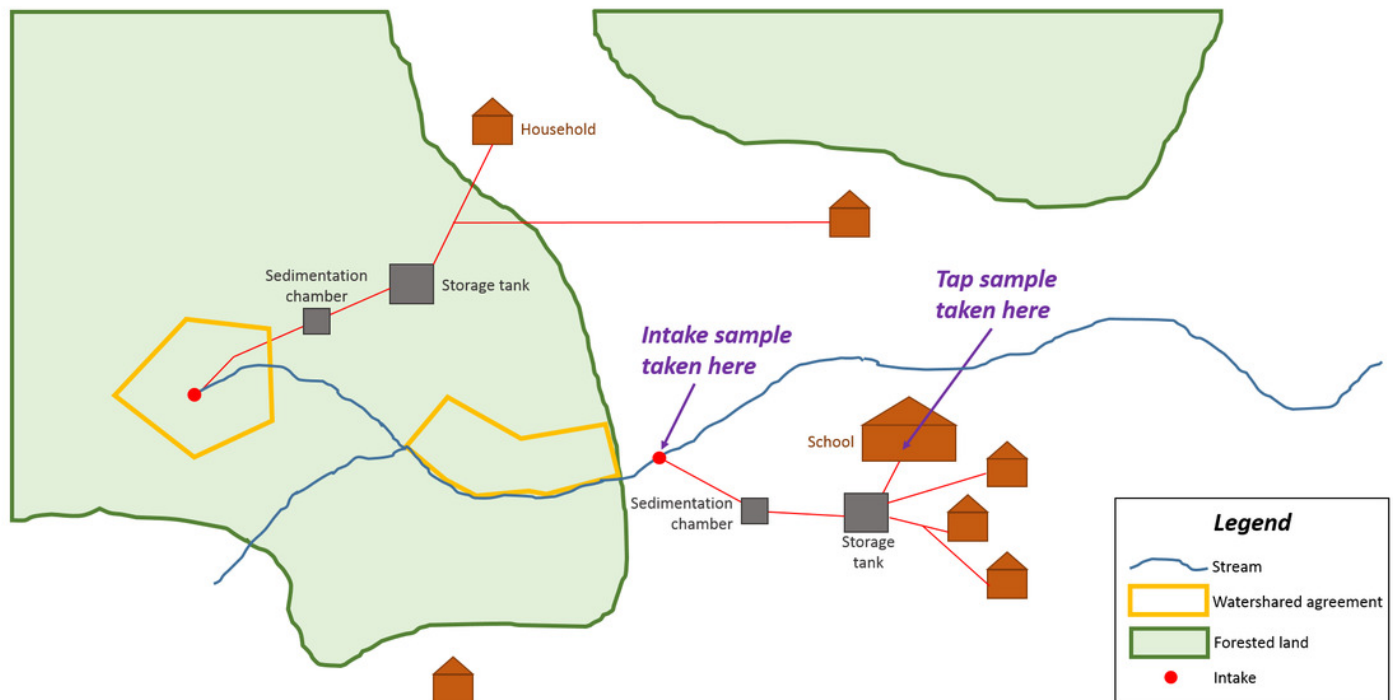



Figure 2

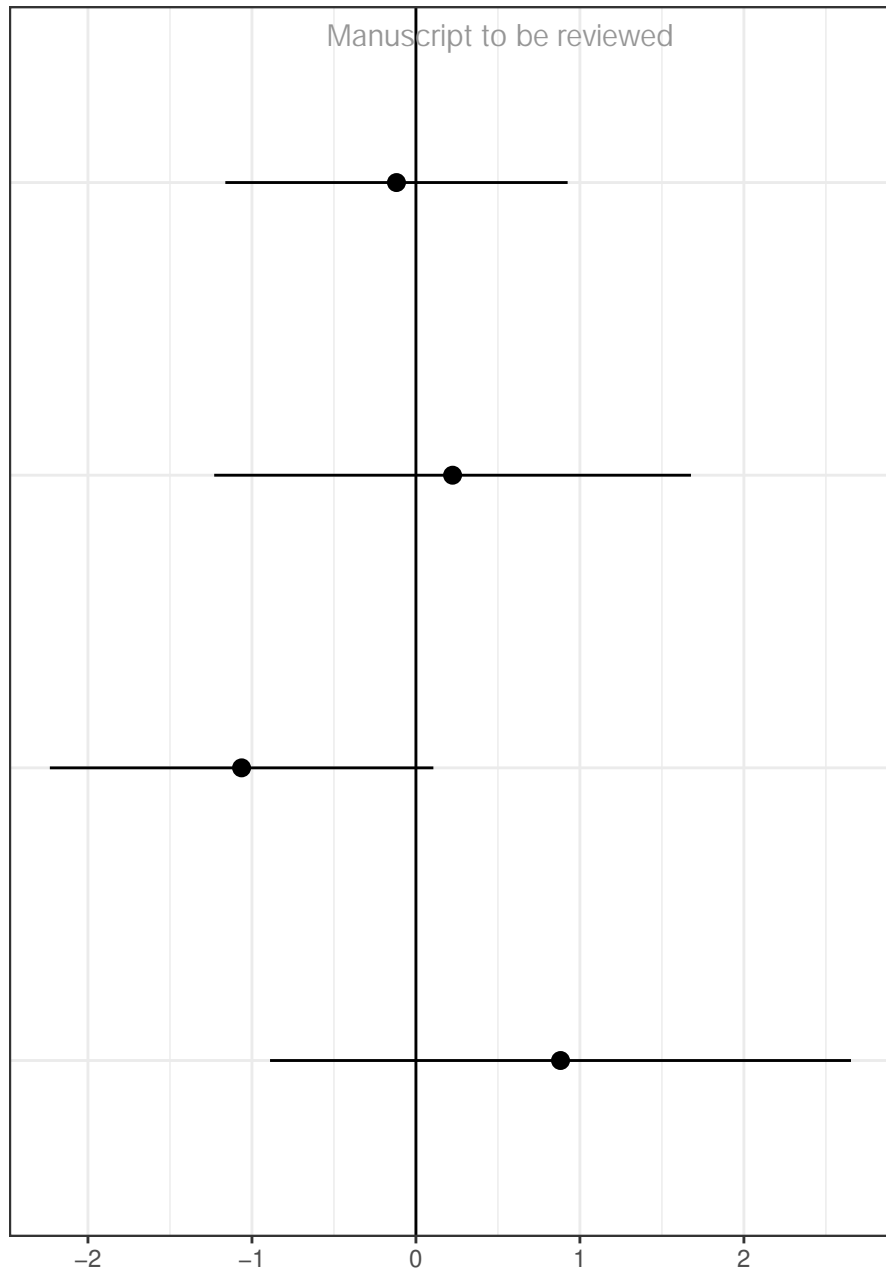
Schematic of an example community with two water intakes showing locations of intake and tap sampling sites.

Balance plot showing the standardized absolute mean differences between control and treatment communities at baseline.

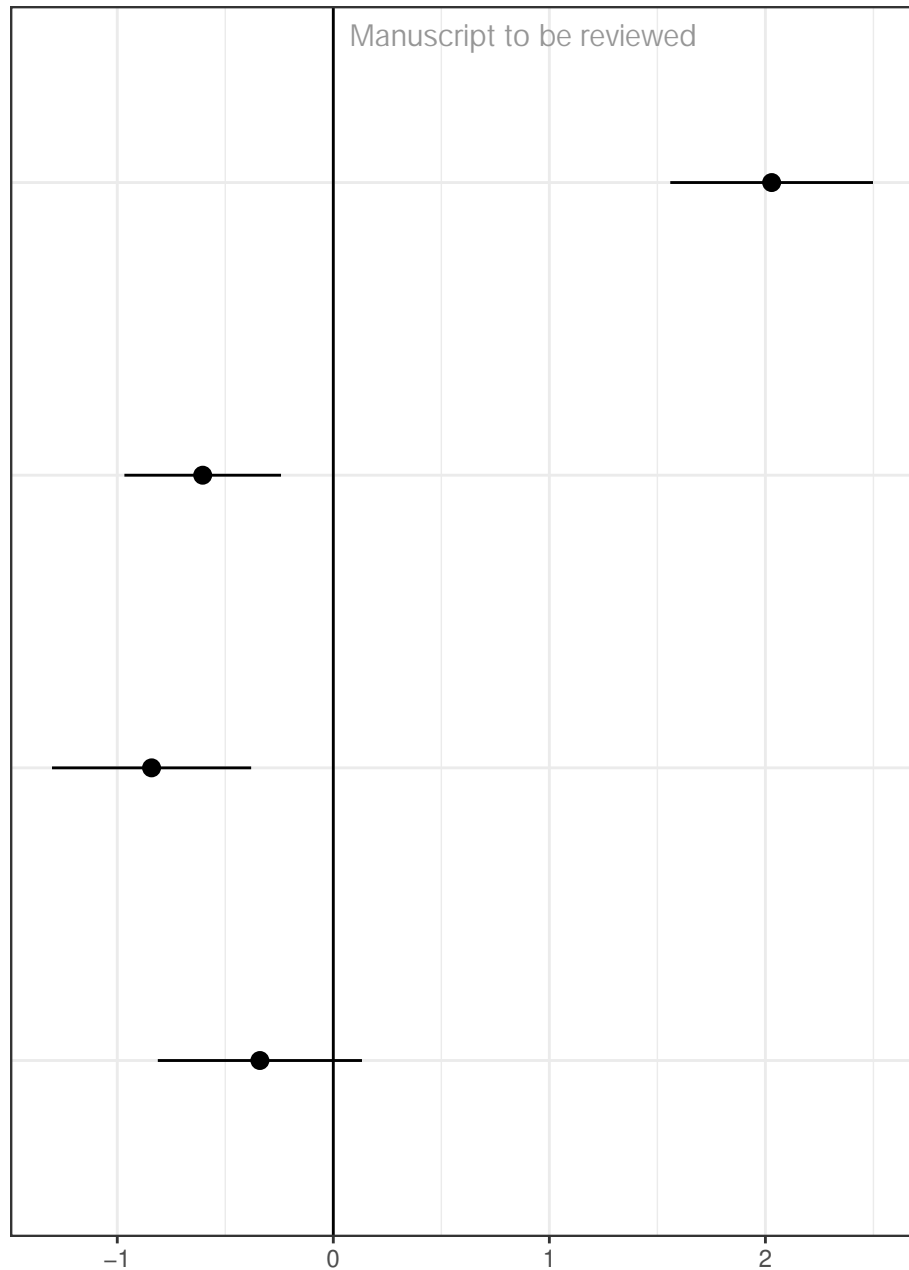
Key baseline variables likely to influence end-line microbial water quality are shown both for the 120 communities where water quality was monitored at baseline, and the subset of 47 communities for which the monitoring location remained the same between baseline and end-line data collection. The loss of sites resulted in a substantial increase in imbalance for two variables (especially in baseline *E. coli* contamination). The dashed line indicates acceptable balance (<0.25 pooled standard deviation units).

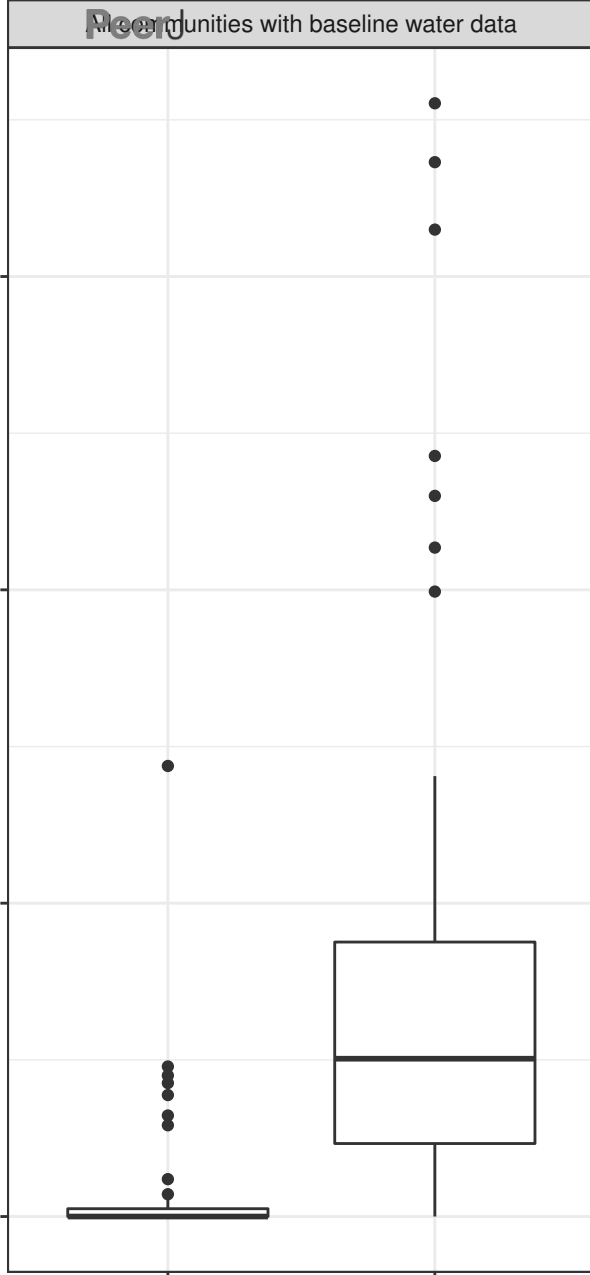
The effect of site features which predict 2015 *E. coli* concentration.

Error bars show 95% confidence intervals. This shows the results of the most highly supported GLMM (see Tables S4a, b and c for model selection tables and details of the presented model). Water systems N=124.

RCT difference-in-differences analysis matched on baseline *E. coli* contamination

This analysis includes only the subset of communities where the sampling location remained the same between baseline and end-line (communities N=47). The model shows no effect of the intervention on microbial water quality (the error bars on the coefficient 'difference in difference in *E. coli* CFU concentration between 2010 and 2015 between treatment and control sites' overlaps zero). The coefficients and confidence intervals are presented in Table S5.

RCT analysis for end-line data only from all communities for which we have end-line data (communities N=116) controlling for predictors of *E. coli* contamination.

This shows no effect of the intervention on microbial water quality (the error bars on the coefficient 'End-line effect of being a treatment community' overlaps zero). The coefficients and confidence intervals are presented in Table S6.

Box plots showing the difference between control and treatment communities in the proportion of eligible land enrolled in level 1 *Watershared* agreements.

The data is shown for all communities for which baseline water quality is available and the subset of communities for which we have a directly comparable end-line.

Variables hypothesized to be important in predicting *E. coli* concentration in 2015.

Codes are used in subsequent model selection tables (Tables S4a and S4b).

Variable	Code	Classification	Base level
Site type	ST	Intake; Tap	Intake
Intake category	IC	Stream; Spring	Stream
Sediment disturbance	SD	Undisturbed; Disturbed	Undisturbed
Intake substrate	IS	Rock only; with sand; with mud	Rock only
Cattle presence	C	Absent; Present	Absent
Agriculture presence	A	Absent; Present	Absent
Turbidity	Tu	Continuous; FAU/100	-
Temperature	T	Continuous	-
Salinity	S	Continuous	-
pH	pH	Continuous	-
Cattle access	CA	Yes; No	Yes
Faces presence	F	Absent; Present in forest; Present in water or on stream banks	Absent
Compliant level 1 *Watershared* area	ARA	None; Intake entirely within conserved area	None

Number (and proportion) of intakes visited in 2015 in compliant *Watershared* agreements and protected from cattle compared between treatment and control communities.

	Treatment community intake	Control community intake
N	68	61
Compliant level 1 *Watershared* conservation agreement (%)	16 (24%)	0 (0%)
Sites with no cattle access (%)	40 (59%)	37 (61%)