Supplementary Information part 3: Testing different forest cover relationships

R. Willem Vervoort[1], Eliana Nervi, Jimena Alonso

Abstract

This supplementary material file compares different linear and non-linear relationships for the impact of forest cover on the change in stream flow.

1. Introduction

This supplementary material is related to 'Generalising the impact of forest cover on streamflow from experimental data: it is not that simple. Vervoort et al.'

In this document we tested the following, which is outlined in the methods of the paper. The changes in forest cover contain both positive (forestation) and negative values (deforestation). In Zhang et al. [2], these changes were jointly analysed, assuming the effect on the change in flow was linear and the effect of removing forest cover was the same as an equivalent addition of forest cover. Here we test two variation of this approach, leading to three approaches

1. Joint analysis and assuming a linear relationship between the change in streamflow and the change in forest cover, following Zhang et al. [2].
2. The impact of an increase in forest cover can be different from the same fractional decrease in forest cover. The question becomes how best to analyse this. One approach would be to allow a different slope and a different intercept for the decreases relative to the increases.
3. A second approach is to test the change in forest cover as a non-linear relationship in the Generalised additive model (GAM).

2. Methods

2.1. Approach 1

To estimate how the change in streamflow is affected by the change in forest cover, while considering the effects of the other variables, we applied generalised additive modelling (GAM) [1].

The general model tested is:

---

*Corresponding author

Email addresses: `willem.vervoort@sydney.edu.au` (R. Willem Vervoort), `eliananervif@gmail.com` (Eliana Nervi), `jalonso@fing.edu.uy` (Jimena Alonso)

[1]Corresponding Author

$$\Delta Qf\% \sim \Delta\%forest\ cover+$$
$$\sum X_i + \sum s(Z_i) + \varepsilon \qquad \text{(S1)}$$

2.2. Approach 2

Deforestation is different from reforestation. This can be tested by converting all the change in forest cover data to positive values, and an additional binary column ($sign_{forestcover}$) can be included indicating whether it was a forest cover increase or decrease. In the model, the parameter for $sign_{forestcover}$ will indicate the difference in the changes in flow for increases in forest cover compared to decreases in forest cover. The disadvantage of this approach is that the relationship with forest cover becomes discontinuous at the origin (0 change in forest cover).

This results in the following model:

$$\Delta Qf\% \sim \Delta\%forest\ cover_{positive} + sign_{forestcover}+$$
$$\sum X_i + \sum s(Z_i) + \varepsilon \qquad \text{(S2)}$$

2.3. Approach 3

The relationship between the change in forest cover and change in streamflow is non-linear. The non-linear approach involves adding a smoothing function the the change in forestry. A shrinkage spline [1] is applied to allow the non-linear effect to be removed. Because a shrinkage penalty is used, this will also test the non-linear assumption and allows the variable for forest cover to be continuous. The disadvantage of this approach is that the relationship between forest cover and change in flow is less easy to interpret, as the non-linear fit in the GAM has no direct parametric form. This leads to the following model:

$$\Delta Qf\% \sim s(\Delta\%forest\ cover)+$$
$$\sum X_i + \sum s(Z_i) + \varepsilon \qquad \text{(S3)}$$

2.4. Preliminary data management

First we will read in the data

We will combine the different tables, but will keep an indicator to see where the data are from.

```
Zhang_small$From <- as.numeric(Zhang_small$From)
Zhang_small$To <- as.numeric(Zhang_small$To)
Zhang_all <- bind_rows(Zhang_large,Zhang_small) %>%
  mutate(dataset = "original Zhang et al data")
```

```
new_data <- new_data %>%
  mutate(dataset = "new data")
All_data <- bind_rows(Zhang_all, new_data)
```

2.5. Implementing the changes to the overall data

The following code implements the changes described in the Supplementary data part 1. However, many of the changes were implemented manually into the data set. These are simply the remaining changes not implemented manually.

1. removing the duplicates.

```
All_data <- All_data %>%
  mutate(`Possible duplicate` =
           ifelse(is.na(`Possible duplicate`)==T,0,`Possible duplicate`),
         `Possible duplicate` = as.numeric(`Possible duplicate`)) %>%
  filter(`Possible duplicate` != 1)
```

2. calculating the dryness

```
# calculate dryness index
All_data <- All_data %>%
  mutate(Dryness = E0/Pa_mm)
```

3. remove watershed 1 (the Amazon) from the analysis

```
All_data <- All_data %>%
  filter(`Watershed #` != 1)
```

4. remove data set 188 and 254 Kamakia and Sambret

```
All_data <- All_data %>%
  filter(`Watershed #` != 188) %>%
  filter(`Watershed #` != 254)
```

5. add a column that indicates forst loss of forest gain

```
All_data <- All_data %>%
  mutate(forest_sign = ifelse(DeltaF_perc < 0, "Forest Cover Loss",
                              "Forest Cover Gain"))
```

3. Results

3.1. Approach 1: linear relationship across all changes in forest cover

First the variable `length` is added, representing the number of years that the data represents.

```
All_data2 <- All_data %>%
  mutate(Forest_Sign = ifelse(DeltaF_perc < 0,
                              "Decrease", "Increase"),
         DeltaF_perc_pos = ifelse(DeltaF_perc < 0,
                                  -1*DeltaF_perc,
                                  DeltaF_perc))
All_data2 <- All_data2 %>%
  mutate(length = To - From,
         mid_year = From + (To - From)/2)

Forest_model_all <- gam(DeltaQf_perc ~ DeltaF_perc +
                    s(log10(Area_km2), k = 5, bs="ts") +
                    s(Dryness, k = 10, bs="ts" ) +
                     s(length, k = 35, bs="ts") +
                    Precip_data_type +  Assessment_technique +
                    Forest_type +
                    Hydrological_regime
                    , data = All_data2)
#summary(Forest_model_all)
#gam.check(Forest_model_all)
#plot(model6_all)
```

Table S1: (#tab:m_all-linear) Statistical summary for the linear terms the full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.47	15.87	-0.47	0.64
DeltaF__perc	-0.59	0.05	-10.89	0
Precip__data__typeOB	-17.87	12.91	-1.38	0.17
Precip__data__typeSG	0.2	14.82	0.01	0.99
Assessment__techniqueEA, HM	18.67	41.7	0.45	0.65
Assessment__techniqueHM	26.61	11.43	2.33	0.02
Assessment__techniquePWE	30.77	11.68	2.63	0.01
Assessment__techniquePWE, HM	15.8	42.22	0.37	0.71
Assessment__techniqueQPW	41.35	19.66	2.1	0.04
Assessment__techniqueQPW, EA	26.05	23.84	1.09	0.28
Assessment__techniqueSH	39.31	11.54	3.41	0
Forest__typeCF	-9.28	7.41	-1.25	0.21
Forest__typeMF	-6.34	7.38	-0.86	0.39
Hydrological__regimeSD	0.13	8.94	0.01	0.99

⁶⁵ This is the baseline result as it echoes Zhang et al. [2] but in this case the
⁶⁶ relationship is tested together with other variables as is discussed in the main
⁶⁷ paper.

⁶⁸ For this model, the focus can be on the linear variables. The variable
⁶⁹ `DeltaF_perc` clearly explains the variation in the change in flow with a p-value
⁷⁰ of 0. The overall model explains 50 % of the variation in the data. For compar-
⁷¹ ison across models with different structures and different variables, the Aikaike
⁷² Information Criteria (AIC) can be used. The AIC of the model is 3166.6801428.

⁷³ *3.2. The change in forest cover converted to positive values*

```
Forest_model_all2 <- gam(DeltaQf_perc ~ DeltaF_perc_pos + Forest_Sign +
                    s(log10(Area_km2), k = 5, bs="ts") +
                    s(Dryness, k = 10, bs="ts" ) +
                     s(length, k = 35, bs="ts") +
                    Precip_data_type +  Assessment_technique +
                    Forest_type +
                    Hydrological_regime
                    , data = All_data2)
#summary(Forest_model_all2)
```

Table S2: (#tab:m_all2-linear) Statistical summary for the lin-
ear terms the alternative model where the change is forestry is all
positive

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	10.91	19.47	0.56	0.58
DeltaF__perc__pos	0.29	0.1	2.97	0
Forest__SignIncrease	-52.26	7.62	-6.86	0
Precip__data__typeOB	-15.75	14.22	-1.11	0.27
Precip__data__typeSG	-5.2	16.81	-0.31	0.76
Assessment__techniqueEA, HM	14.21	45.49	0.31	0.76
Assessment__techniqueHM	27	13.31	2.03	0.04
Assessment__techniquePWE	28.19	13.57	2.08	0.04
Assessment__techniquePWE, HM	16.36	45.51	0.36	0.72
Assessment__techniqueQPW	28.63	21.65	1.32	0.19
Assessment__techniqueQPW, EA	35.76	25.74	1.39	0.17
Assessment__techniqueSH	33.52	12.94	2.59	0.01
Forest__typeCF	-12.47	8.06	-1.55	0.12
Forest__typeMF	-5.1	8.18	-0.62	0.53

	Estimate	Std. Error	t value	Pr(>|t|)
Hydrological_regimeSD	3.45	9.76	0.35	0.72

⁷⁴ Again the focus can be on the linear variables. In this case the variable
⁷⁵ `DeltaF_perc_pos` also is important to explain the variation in the change in
⁷⁶ flow with a p-value of 0. However, the results also show that in this case,
⁷⁷ the deforestation (`Forest_Sign` < 0) has a different effect than reforestation
⁷⁸ (`Forest_Sign` > 0). In fact, in absolute terms deforestration creates a -5226
⁷⁹ % greater effect on streamflow than reforestation (and of course deforestation
⁸⁰ creates and increase in flow and reforestation a decrease).

⁸¹ However, overall this model explains less of the overall variation in the data
⁸² with a deviance explained of 0.46 and an AIC of 3212.700798.

⁸³ *3.3. Assuming the change in forest cover is non-linear*

⁸⁴ The difference between reforestation and deforestation might suggest that
⁸⁵ the relationship between the change in forest cover and change in stream flow
⁸⁶ is non-linear. This is tested in the model where the change in forest cover is
⁸⁷ considered non-linear. The number of knots was decided based on checking the
⁸⁸ edf value relative to the apparent knots k' wiht the procedure `gam.check()` in
⁸⁹ the package `mgcv` [1].

```
Forest_model_nonlin <- gam(DeltaQf_perc ~ s(DeltaF_perc, k=10, bs = "ts") +
                    s(log10(Area_km2), k = 5, bs="ts") +
                    s(Dryness, k = 10, bs="ts" ) +
                     s(length, k = 35, bs="ts") +
                    Precip_data_type +  Assessment_technique +
                    Forest_type +
                    Hydrological_regime
                    , data = All_data2)
#summary(Forest_model_nonlin)
```

Table S3: (#tab:m_nonlin-smooth) Statistical summary for the smooth terms for the model that considers the change in forest cover to be a non-linear variable

|	edf	Ref.df	F	p-value
s(DeltaF_perc)	1.11	9	12.72	0
s(log10(Area_km2))	0.76	4	0.8	0.03
s(Dryness)	4.7	9	2.2	0
s(length)	4.45	34	0.24	0.09

⁹⁰ In this case, the non-linear terms of the model need to be checked, which
⁹¹ shows that once again, changes in forest cover impact changes in flow as indi-
⁹² cated by the p-value of 0. The total deviance explained for the model is 0.5 and

the model has an AIC of 3166.9317897. These performance statistics are similar
to the first (baseline) model.

One thing that is of interest is the fact that while a non-linear relationship
between the change in forest cover and the change in streamflow was conceptu-
alised, the actual relationship turns out to be linear. This can be shown from
the plot and by running `gam.check`.

```
D <- visreg(Forest_model_nonlin, "DeltaF_perc", gg = T, plot = T) + theme_bw()
A <- visreg(Forest_model_nonlin, "Area_km2", gg = T, plot = T) + theme_bw() +
  scale_x_log10()

(A + D)
```
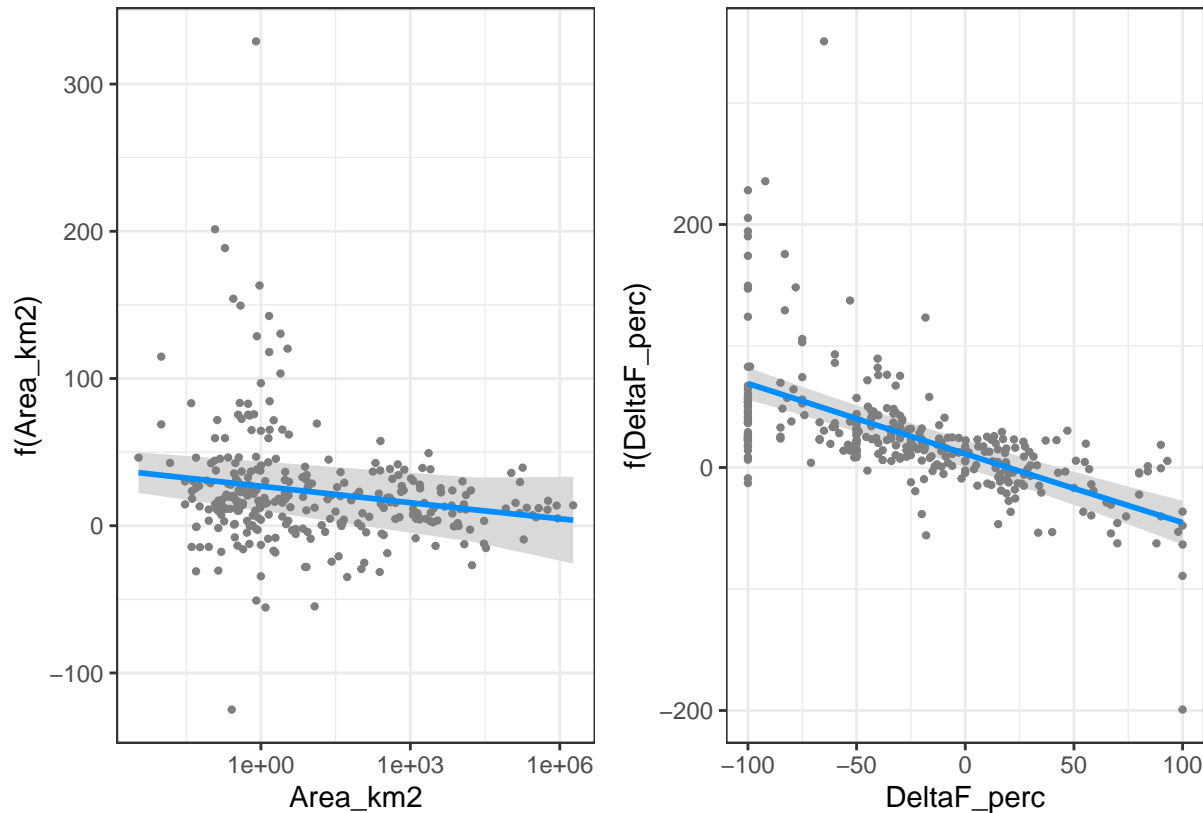

The procedure `gam.check` highlights that the p-value of the Basis dimension
check of the smooth on `DeltaF_perc` is very small and the actual edf is close to
1 suggesting a linear relationship, similar to the smooth on log10(`Area_km2`).

The residual analysis shows that including this non-linearity makes no differ-
ence in terms of he distribution of the residuals. While there is some remaining
weak trend in the data, the overall distribution of the residuals is approximately

normal.

```
par(mfrow= c(2,2))
gam.check(Forest_model_nonlin)
```


Resids vs. linear pred.

Histogram of residuals

Response vs. Fitted Values

107
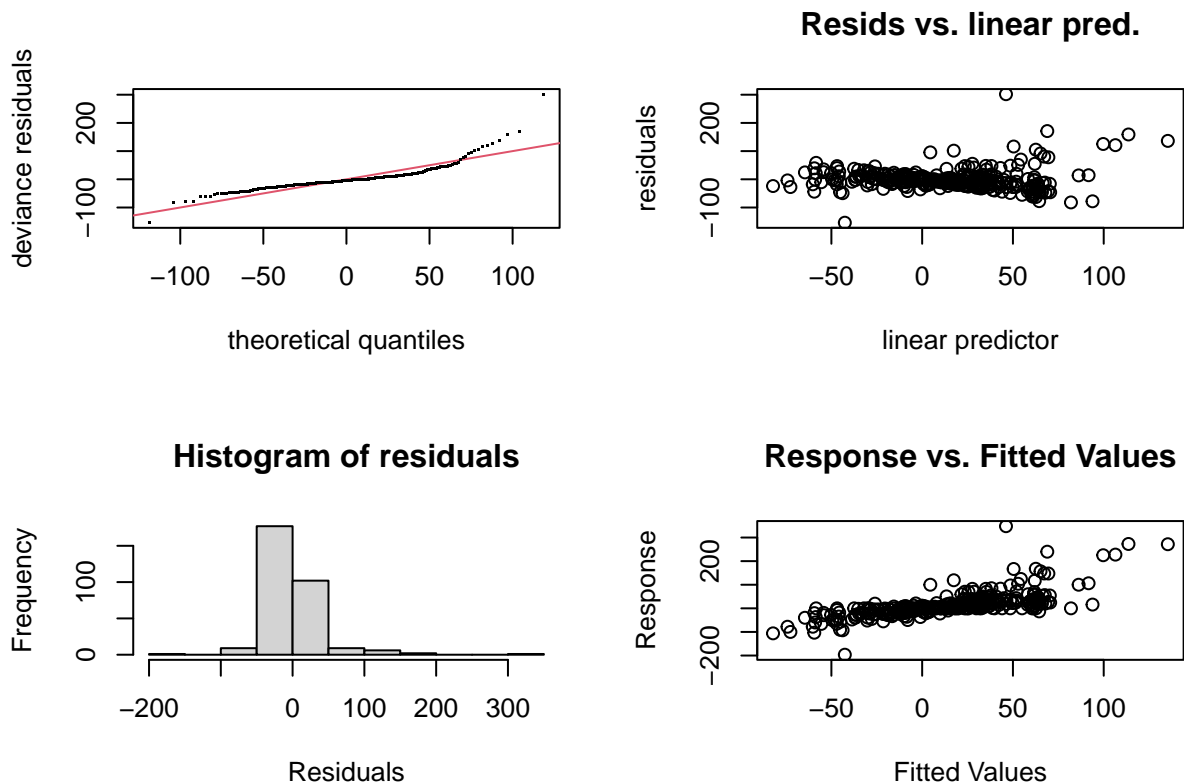
```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 15 iterations.
## The RMS GCV score gradient at convergence was 0.004151441 .
## The Hessian was positive definite.
## Model rank =  69 / 69
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                        k'    edf k-index p-value
## s(DeltaF_perc)      9.000  1.113    0.89    0.03 *
## s(log10(Area_km2))  4.000  0.756    0.96    0.20
```

8

```
## s(Dryness)             9.000  4.703     0.94     0.13
## s(length)             34.000  4.448     0.92     0.10 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References

[1] S. Wood. *Generalized additive models: an introduction with R.* CRC Press, Boca Raton, FL, 2006. ISBN 978-1584884743.

[2] Mingfang Zhang, Ning Liu, Richard Harper, Qiang Li, Kuan Liu, Xiaohua Wei, Dingyuan Ning, Yiping Hou, and Shirong Liu. A global review on hydrological responses to forest change across multiple spatial scales: Importance of scale, climate, forest type and hydrological regime. *Journal of Hydrology*, 546:44–59, 2017. ISSN 0022-1694. doi: https://doi. org/10.1016/j.jhydrol.2016.12.040. URL `http://www.sciencedirect.com/science/article/pii/S0022169416308307`.