WHITE PAPER
SETTING OPTIMAL BASELINES FOR REDD+ PROJECTS
A Case Study for the Arc of Deforestation in the Colombian Amazon

Nature Reserve Guayabero – Cocodrilo, Department of Meta, Colombia.
Photo: Juan Carlos Rivas/Permian Global.

Permian Global
July 2022

Table of Contents

Attribution

The development of this white paper was led by the Technical Team at Permian Global. It has been developed as part of the Science and Technical Innovation Program to improve our understanding of tropical deforestation and implications for the effectiveness of nature base solutions (NBS) projects, particularly REDD+ and Forest Restoration efforts.

Please cite the work as follows: Permian Global. 2022. White paper. Setting Optimal Baselines for REDD+ Projects. A Case Study for the Arc of Deforestation in the Colombian Amazon. Bogotá, July, 2022.

The Permian Global task team responsible for this white paper was composed of Dr. Leonardo Sáenz, Ronald Rico, David Stone and Juan Chang.

Acknowledgments

We are grateful to Dr. Jonah Busch and Dr. Rene Zamora for helpful comments, which made this manuscript stronger.

Abstract

This study explores the use of dense annual time series of land cover data to test what ranges of historical reference periods could be more optimal for setting baselines of REDD+ initiatives in the arc of deforestation of Colombia. Our methods represent an improvement over previous similar studies by providing a more robust temporal depth of time series analysis and a more direct comparison with standards or methodologies used to design REDD+ projects.

We focused the analysis on four municipalities in the arc of deforestation of Colombia's Amazon biome (La Macarena, San Jose del Guaviare, Cartagena del Chaira and San Vicente del Caguan). These areas represent northern Amazon landscapes where sudden, nonlinear, unforeseen, or dramatic shifts in land use have occurred, over the last two decades[1]. The analysis builds on a historic assessment of forest cover changes associated with deforestation between the years 2001 and 2020. The combined extend of these municipalities expands more than 2.5 million hectares, which makes results also useful to help implement nesting approaches in these areas of the Amazon biome of Colombia.

Five different scenarios were used to project future deforestation after 2001 and four hypothetical historical reference periods (four, six, eight, and 10 years) were chosen to assess what period or periods resulted in better projections of deforestation after the year 2001. The projections were generated by extrapolating the hypothetical baseline and were then compared with real estimates of deforestation within hypothetical monitoring periods. The combination of municipalities, hypothetical historical reference periods and scenarios resulted in 80 regressions, which were compared using indicators of goodness of fit such as R2 and RMSE for each of the hypothetical historical reference periods.

Our results show that deforestation has been a highly random phenomenon over the last two decades in these municipalities, but with an interannual increasing pattern, and that 10 years historical reference periods were better suited to capture such variability, when compared with four, six, and eight years periods. This randomness shows that deforestation has been a highly variable process due to the occurrence of interannual nonlinear shifts in the historical record, while depicting a clear increasing signal in these municipalities of Colombia. Four years historical reference periods showed the least agreement when contrasting calibration and monitoring periods, resulting in highly volatile deforestation projections. Similarly, six and eight years periods showed intermediate performances depending on the scenario tested and the characteristics of each municipality, but generally showed statistically larger errors than the 10 years period. Therefore, the six years period showed closer performance to the four years period, and eight years period closer to the 10 years period. The latter finding suggests that longer historical reference periods should always be preferred over shorter ones when generating REDD+ projects baselines,

[1] https://www.globalforestwatch.org/blog/data-and-research/2017-was-the-second-worst-year-on-record-for-tropical-tree-cover-loss/

particularly in high-forest cover high-deforestation areas where non-linear shifts are known to occur, as was demonstrated for these areas of Colombia.

Finally, critics of REDD+ projects argue that project baselines are commonly inflated and that one of the solutions to that is to shorten their historical reference periods. Our study has shown that this is not an ideal solution when data and funding limitations, as well as political, social and regulatory barriers may also play a role. Instead, we argue that it is important to focus greater attention on ensuring that project baselines are more adequately projected into the future. Also, we highlight the importance that baseline reassessment efforts consider the history of deforestation in the area, and in particular, the evidence of interannual nonlinear shifts in the historical record. Although, the purpose of the paper is not to extrapolate general conclusions across the tropics from these sites in Colombia, where deforestation has been going up over this time period, we recommend that this type of the analysis should be replicated in all tropical forest basins, by the most robust standards, to better inform the setting of historical reference periods, and thus deforestation baselines. With the increasing requirements for projects to nest within larger jurisdictional initiatives, the need to consider these spatial, regional and temporal differences in deforestation becomes paramount to guide and continue to incentivize optimal investment decisions and to preserve the integrity and additionality of REDD+ projects.

High-level messages

- Longer historical reference periods are better able to capture the random nature of deforestation in the arc of deforestation in Colombia, which has been subject to nonlinear shifts over the past two decades.

- Deforestation has shown to be a highly random and variable phenomenon within the historical time series studied and the different municipalities selected for analysis.

- Hypothetical baselines using 10 years historical reference periods tended to be more optimal, in terms of projecting the rate of deforestation, for all the scenarios and municipalities under study. Model performance was higher and consistently showed lower projection errors. In short, more data observations within a deforestation sample, in general, resulted in statistically better model performance.

- Although, the purpose of the study is not to extrapolate general conclusions across the tropics from these sites in Colombia, where deforestation has been increasing over this time period, we recommend that this type of the analysis should be replicated in all tropical forest basins, by the most robust standards, to better inform the setting of historical reference periods.

- However, rather than choosing shorter historical reference periods, greater attention should be put to more adequately project historical baselines into validity periods.

- When deforestation follows a plausible trend, linear models can be more suitable as shown here, even projecting deforestation well within 10 years hypothetical monitoring periods.

- When deforestation observations are highly random, which means that no patterns, either increasing or decreasing can be identified, then the use of the simply historical average models could be more optimal. However, historical average models showed to underestimate deforestation largely in these Colombian municipalities, likely incorrectly labelling any projects as of lower additionality.

- If significant nonlinear shifts are confirmed to have occurred after project start date, which are not represented in the historical reference period chosen, projects should consider a shorter reassessment period to prevent that they start relying significantly on their pool buffers, while maintaining 10 years historical reference periods.

- However, the fact that hypothetical baselines using 10 years historical reference periods showed to represent better the highly abrupt forest cover change in this region of Colombia in 2016, indicates that longer historical periods should always be considered when generating project baselines, even when a shorter baseline reassessment period could be considered. Fortunately, the availability of high-resolution remote sensing data, at least over the last two decades, facilitates the creation of robust 10 years historical reference periods.

- Therefore, baseline reassessment efforts should always consider the history of deforestation in proposed project regions and if interannual nonlinear shifts are identified in the historical record, maintaining a 10 years historical reassessment period could still be plausible as demonstrated here.

- Studies that have argued REDD+ project baselines are commonly inflated often fail to replicate all determinates of deforestation affecting the project. While some projects in the voluntary carbon market may lack additionality, the inference that all projects inflate their baselines is unsubstantiated.

1. Introduction

One of the important debates in REDD+ is how to establish forest reference emission levels or forest reference levels, which are the benchmarks used to calculate carbon credits. At the REDD+ project level, emissions reductions from avoided planned or unplanned deforestation activities are calculated based on a baseline scenario. Standards, such as the VCS standard (Verra, 2020), and others, require that project developers produce a baseline scenario (a counterfactual scenario) to represent what would happen to the forests if project activities were not implemented.

Currently, project baselines are developed with data derived from a representative historical reference period, which can be between eight to 12 years in the past depending on project activity (Verra, 2020). However, Verra and other standards are proposing to use shorter historical reference periods in order to align with a five years update of national scale Forest Reference Levels (FRELs) as per high level policy recommendations, which resulted from the Paris Agreement.

In summary, a reference level or project baseline is the expected net carbon stock change (expressed in metric tons of carbon dioxide equivalent per year) in a counterfactual scenario that would most likely occur without intervention. Therefore, a credible and accurate reference level or project baseline is critical to ensure additionality of funding for REDD+.

1.1 Context to the Additionality of REDD+ Projects

The concept of additionality refers to the notion that greenhouse gas (GHG) reductions would not have occurred in the absence of a REDD+ project or initiative. If reductions had happened anyway – i.e., without any prospect for project owners to sell carbon credits – then they would not be additional. Additionality is essential therefore for the quality of carbon credits. If their associated GHG reductions are not additional, then purchasing offset credits in lieu of reducing a company's emissions could in turn make climate change worse. In a similar way, national or subnational REDD+ programs should also follow closely the principle of additionality to make sure their activities are effectively contributing to achieving their National Determined Contributions (NDCs).

1.2 The Need for Setting Sound Projects Baselines

A credible and accurate project baseline or jurisdictional reference level is therefore critical to ensure additionality of funding for REDD+. The choice of baseline is of great importance for the amount of carbon credits awarded, and thus potentially also for the benefits that local people could receive as compensation for preventing further deforestation or forest degradation. A baseline that is too generous than is realistic based on longer term data will bring the credibility of REDD+ initiatives into question. Lower estimates will result in projects being incorrectly labelled as of low additionality and risks important forest protection as being labelled unnecessary.

1.3 Context to Historical Reference Periods Currently Used for both, Projects and Jurisdictional Initiatives

Essentially, a reference level is the expected net carbon stock change (expressed in metric tons of carbon dioxide equivalent per year) in a baseline scenario without intervention. Various approaches have been developed to establish reference levels and include simple averages of historical net emissions, or linear extrapolations of historical emission levels that to some extent may account for expected changes in drivers of deforestation or forest degradation (Huettner, Leemans, Kok, & Ebeling, 2009). In addition, spatial modelling is often required across entire reference regions when an additionality case is difficult to make from using simple historical average models only (Shoch, et al. 2013; VERRA, 2020).

However, the additionality of voluntary REDD+ projects has been called into question by critics concerned that deforestation baselines might be in some cases intentionally inflated, resulting in credits that are not additional, the so called "hot air" (Seyller et al., 2016). In this direction, Thales et al., (2020) studying the reductions in forest loss attributed to REDD+ projects in the Brazilian Amazon, conclude that crediting baselines assume consistently higher deforestation than the forest loss observed in counterfactual sites using a synthetic control approach. Nevertheless, their study faced certain limitations, which prevented it from providing a direct comparison between project areas and their synthetic control sites. The limitations included that their method may not have included all relevant structural determinants of deforestation and that the period of analysis may not have been long enough to observe significant REDD+ impacts in some cases. Amid these uncertainties, the authors also conclude that while there are projects that clearly have not performed as expected, their results did not imply that voluntary REDD+ projects cannot achieve their objectives if designed and implemented effectively.

For the case of national or jurisdictional approaches, forest reference levels have been proposed to be updated every five years[2] with some authors recommending to use standardized considerations of national circumstances to adjust for particular likely forest transitions in coming years (Köthke, Schröppel, and Elsasser 2014). In practice the adjustment for national circumstances has been implemented as a percentage change above or below the national average (Mertz, et al., 2017). However, Ramankutty & Coomes, (2016) argue that many of these national scale submissions do not account for changes in known or the emergence of new future drivers of deforestation, which may set off rapid and unpredictable shifts in land use and deforestation rates, as often occurred in the past. Moreover, Ankersen et al. (2015) questions the validity of reference levels for subnational

[2] https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf

or local initiatives as these are made at national level, and therefore are likely to be poor predictors of the actual trends in forest cover at smaller scales. At local scales, the dynamics and drivers of change are better known but can be difficult to predict from national scale assessments.

1.4 **Deforestation complexities to consider when setting project baselines**

It is important to highlight that land use does not always follow stable historical trends, but can be subject to nonlinear change that is difficult to predict and that can be dramatic (Ramankutty, N. and Coomes, O. T. 2016; Mertz, O. et al,. 2017). For example, nonlinear shifts can be the result of the impact of large El Niño years on increased burned area and thus deforestation and CO2 emissions, which tend to occur in cycles of more than six years (Burton, et al., 2020). Also, they can result from the influence of promoting the introduction of more intensive cash cropping systems (Vongvisouk, T. et al., 2016).

Furthermore, nonlinear shifts in land use can be a consequence of large waves of production of illicit crops, such as coca cultivation for cocaine production, with evidence suggesting that resulting deforestation peaks can occur at intervals that have historically exceeded five years. Moreover, Ankersen, (2015) argues that nonlinear forest transitions are likely to cause inaccuracies to deforestation baselines, since they are hard to predict and can be caused by political regulation not previously expected. If these nonlinear phenomena are not properly represented in historical reference periods, the viability of REDD+ projects will become much more unpredictable.

However, there is evidence in the literature that relatively long historical reference periods (between 8 to 12 years), can better provide observations of such non-linear forest cover shifts, which can be beneficial when setting the baselines of projects. This characteristic of longer historical reference periods can be explained by the fact that samples with a greater number of observations describing a given population always result in smaller standard and prediction error of predictive models, when compared with those derived from smaller samples of the same population. Also, larger samples also provide more information to correlate deforestation data with other environmental, demographic, political and economic dynamics driving forest cover changes (Müller, D. et al., 2014; Mertz, O. et al,. 2017; Ankersen, J. et al., 2015line projects; Brown, et al., 2007).

1.5 The spatial scale of national/jurisdictional and project-based analysis is dramatically different

Finally, it is important to highlight that there can be significant spatial and temporal mismatches between deforestation drivers at national or jurisdictional scales vs project scales. There is evidence that deforestation rates of second or third level subnational jurisdictions may not be captured by national scale FRELs, because the effect of deforestation drivers may be very localized in nature (Ankersen et al. 2015). Such drivers may include shifts in historical land uses, socio-economic trends, armed conflict and changes in laws and regulations, which can show greater variability locally. Therefore, these spatial and temporal differences can be very pronounced at the project scale.

With the increasing requirements for projects to nest within larger jurisdictional initiatives the need to consider these spatial and temporal differences, when setting project baselines, becomes paramount to guide and continue to incentivize optimal investment decisions and to preserve the integrity and additionality of REDD+ projects.

2. The focus of the White Paper:

In this paper, we have two main objectives:

1) to demonstrate how dense annual time series of land cover data can provide spatially and temporally detailed analysis of deforestation in the arc of deforestation of Colombia;

2) to take a retrospective approach to testing what ranges of historical reference periods could be more optimal to setting baselines for REDD+ Initiatives and thus calculate REDD+ credits. This is similar to what was done by Mertz, et al., (2017), but here we can provide a more comprehensive and robust analysis for REDD+ initiatives given the improved temporal depth of the time series used and the more direct comparison with standards or methodological requirements regarding the end of historical reference periods relative to project start dates.

Specifically, we tested what would be the result if REDD+ projects had been planned in the past based on different historical reference periods. We discuss what extent of additionality would have been achieved by REDD+ initiatives using different hypothetical historical reference periods. Four municipalities in Colombia's arc of deforestation in the Amazon biome were used as case studies for this historical reference periods assessment. These areas were chosen as they effectively represent northern amazon landscapes where sudden, nonlinear, unforeseen, or dramatic shifts in land use practice are occurring[3]. The combined extend of these municipalities expands more than 2.5 million hectares, which makes results from this study useful to help guiding future jurisdictional scale approaches in these areas of the Colombian Amazon.

[3] https://www.globalforestwatch.org/blog/data-and-research/2017-was-the-second-worst-year-on-record-for-tropical-tree-cover-loss/

3. Methods

The analysis was conducted at the municipal level comprising four municipalities in the arc of deforestation in Colombia namely Cartagena del Chaira, San Jose del Guaviare, La Macarena and San Vicente del Caguan with a specific focus to test the change in deforestation predictions on four municipalities with a contrasting history of forest cover change. Cartagena del Chaira and San Jose del Guaviare were notably affected by the change in national circumstances after the signature of the peace deal with FARC in 2016, and La Macarena and San Vicente del Caguan, which aside from seeing the impact of the latter peace accord in 2016, also experienced land use change impacts after the failed peace process with FARC, which ended in 2002 (Table 1; Figure 1).

The analysis builds on a historic assessment of forest cover changes associated with deforestation between 2001 and 2020. Deforestation was assumed as a permanent land use change over the hypothetical reference period using data from Global Forest Watch. The definition of forest used was canopy cover above 30% in 30-meter pixels as per Hansen et al., (2013). In this way forest loss, as provided from Global Forest Watch, was converted to deforestation by filtering pixels, where forest cover dropped below 30% over the historical reference period.

Five different scenarios were used to project future deforestation after 2001 and four hypothetical historical reference periods (4, 6, 8 and 10 years) were chosen to assess what period or periods resulted in better predictions of deforestation after 2001. The projections were generated by extrapolating the hypothetical baseline and were then compared with the real estimates of deforestation for the hypothetical monitoring period. No spatial modelling is used here and only the spatial rates of deforestation are considered, both in the hypothetical historical reference period and the hypothetical monitoring period. Finally, the combination of municipalities, hypothetical historical reference periods and scenarios resulted in 80 regressions, which were compared using indicators of goodness of fit such as R2 and RMSE for each of the hypothetical historical reference periods. Figure 1 presents the location of the study area.

Table 1. Characterization of forest cover changes in the municipalities under study.

Municipality	polygon Area	Forest 2001	Forest 2020	Forest loss 2001-2020	Annual Forest Loss	Annual Forest Loss Rate
	ha				ha/year	%
San José del	1,599,033	1,457,087	1,297,750	159,338	8,386	0.6

Guaviare						
Cartagena de Chairá	970,418	874,503	662,396	212,107	11,164	1.7
La Macarena	503,271	495,353	338,146	157,207	8,274	2.4
San Vicente	753,537	553,790	342,182	211,608	11,137	3.3

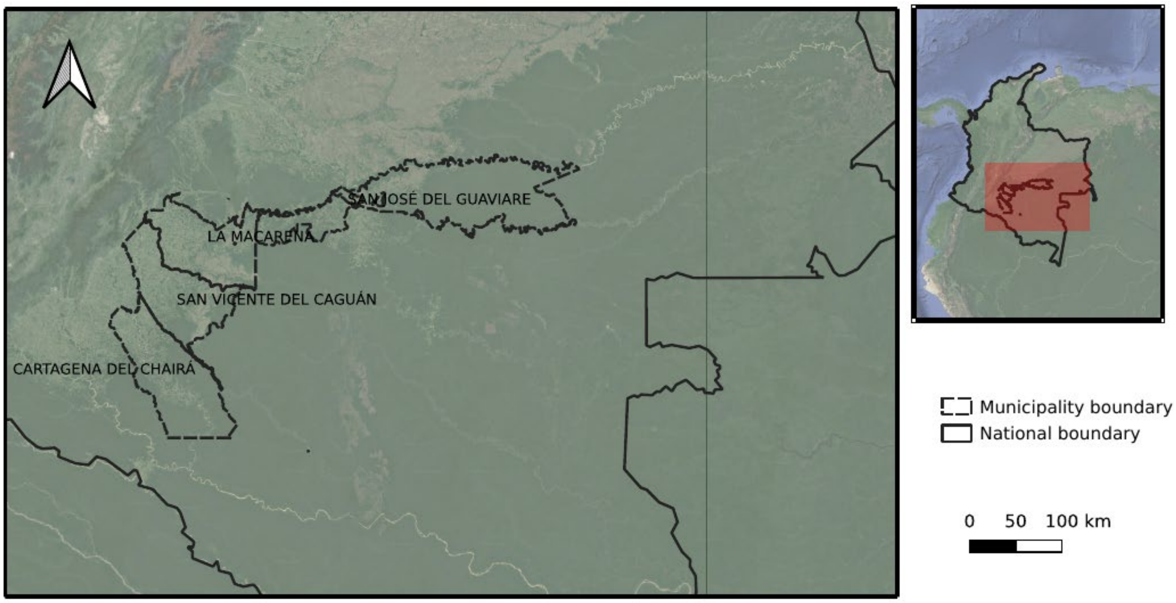


Figure 1. Study area considering four municipalities in the arc of deforestation in Colombia namely: San Jose del Guaviare, Cartagena del Chaira, La Macarena and San Vicente del Caguan.

3.1 Description of Scenarios

A detailed description of scenarios is presented below.

A. Replicate of the Mertz et al., 2017 study for Colombia, with all hypothetical historical reference periods starting in 2009 and forecasting until 2020. For this scenario the four years hypothetical baseline will be compared with eight years monitoring period between 2013 and 2020, whereas the 10 years hypothetical baseline will be compared with a hypothetical monitoring period of only two years between 2019 and 2020. Six and eight years hypothetical baselines will have six and four years hypothetical monitoring periods respectively. This scenario is particularly statistically hard with the 10 years hypothetical baseline as the goodness of fit of the projection will be calculated against two hypothetical monitoring years only, within a highly random distribution of annual deforestation observations, which also coincides with a significant change in forest national circumstances in Colombia since 2016. Also, it may not be as relevant for the start date of REDD+ projects. However, we consider it useful because it can show the performance of different historical reference periods to project deforestation into the future using the same starting year.

B. Replicate of the VCS standard for REDD+ projects, which requires that historical reference periods end at most two years before project start date. Hypothetical project start date is assumed in 2020 with an end to the hypothetical historical reference period in 2018. For this scenario all four hypothetical baselines will be compared with a hypothetical monitoring period of only two years between 2019 and 2020. This scenario also results in a hard but fairer comparison of model performance between the different hypothetical baselines when compared with scenario A. For this scenario, a four years hypothetical baseline starts in 2015, whereas a 10 years hypothetical baseline starts in 2009. Six and eight years hypothetical baselines start in 2013 and 2011 respectively. This scenario is more relevant for the start date of REDD+ projects.

C. Test of a hypothetical monitoring period covering five years before the signature of Colombia's peace deal with FARC. This scenario results in a fair comparison between hypothetical baselines because all scenarios are forced to project five years during the 2011 and 2015 period, when deforestation in Colombia was relatively constant or experiencing declines (in particular municipalities) and had not been affected by the significant change in forest national circumstances resulting from the peace accord with FARC in 2016. The scenario also serves to explore the suitability of a five years reassessment period when compared with a 10 years hypothetical monitoring period scenario. For this scenario a four years hypothetical baseline starts in 2007, whereas a 10 years hypothetical baseline starts in 2001. Six and eight years hypothetical baselines start in 2005 and 2003 respectively. This scenario is also relevant for the start date of REDD+ projects with the advantage that it provides a longer and identical historical monitoring period.

D. Test of a hypothetical monitoring period covering five years after the signature of Colombia's peace deal with FARC starting in 2016. This scenario results in a hard but fair comparison between hypothetical baselines, because all scenarios were forced to project five years during the 2016 and 2020 period, when forest cover in Colombia was affected by a drastic change in forest national circumstances in 2016. The scenario can also help shed light into whether longer or shorter historical reference periods can provide better information for baselines generation when random and abrupt nonlinear shifts occur. The scenario also serves to explore the suitability of a five years reassessment period when compared with a 10 years hypothetical monitoring period scenario. For this scenario a four years hypothetical baseline starts in 2012, whereas a 10 years hypothetical baseline starts in 2006. Six and eight years hypothetical baselines start in 2010 and 2008 respectively. This scenario is also relevant for the start date of REDD+ projects with the advantage that it provides a longer and identical historical monitoring period.

E. Test of a historical baselines against a hypothetical monitoring period covering 10 years, coinciding with the signature of the peace deal with FARC. This scenario results in a hard but fear comparison between hypothetical baselines, because all scenarios are forced to project 10 years during the 2011 - 2020 period when forest cover in Colombia was affected by a significant change in forest national circumstances in 2016. The scenario can also help shed light into whether longer or shorter historical reference periods can provide better information for baselines generation when random nonlinear shifts occur. The scenario also serves to explore the suitability of a 10 years reassessment period when compared with scenarios C and D. For this scenario a four years hypothetical baseline starts in 2007, whereas a 10 years hypothetical baseline starts in 2001. Six and eight years hypothetical baselines start in 2005 and 2003 respectively. This scenario is also relevant for the start date of REDD+ projects with the advantage that provides a longer and identical historical monitoring period.

3.2 Spatial filtering of polygons and data download

In order to estimate deforestation a spatial filter was applied to the polygons of the municipalities, complying with the restriction of not including protected areas or areas of a special nature (i.e. zonas de reserva campesina), which could bias the distribution of deforestation. Adjusted municipality polygons were loaded on the Global Forest Watch (GFW) platform to extract deforestation data from 2001 and 2020. Data were downloaded in text format for each municipality in hectares per year.

Figure 2 shows the distribution of deforestation in the study since 2001.

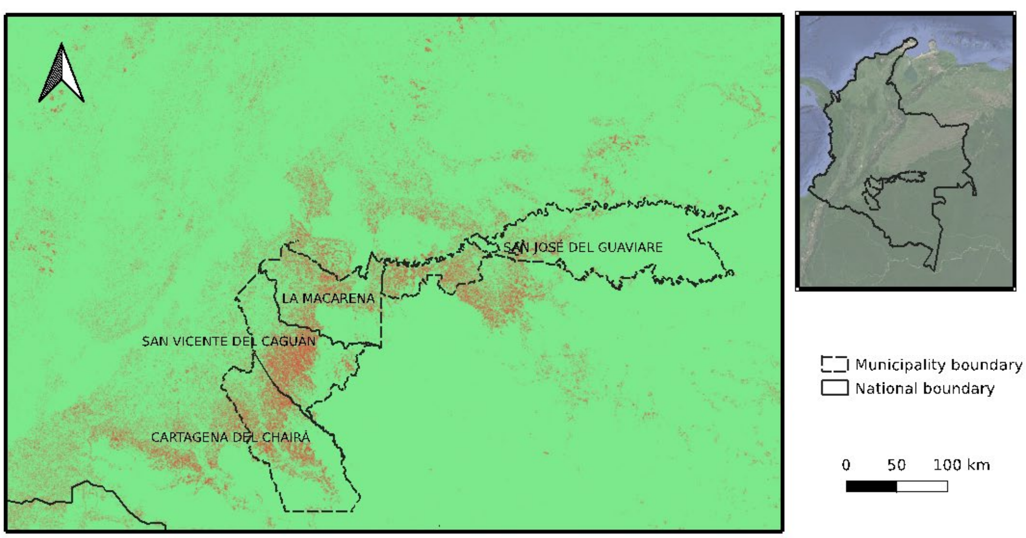


Figure 2: Distribution of deforestation in the study area since 2001.

3.3 Data processing

Two levels of analysis were created: one concerning the length of the historical reference period (4, 6, 8 and 10 years), and another related to the beginning (t0) and the end (t1) of the particular historical reference periods in each of the five scenarios described above.

Baseline projections were made based on simple historical average models and linear regression models, as recommended by VCS standard (Verra, 2020), where the independent variable **x** was the year, and the dependent variable **y** was the corresponding amount of deforested area estimated for that year, obtained from GFW. The range of deforestation values used to build the prediction models were obtained for the period of time in years between t0 and t1 for each scenario respectively. Using this methodology, the different hypothetical baselines for each scenario and municipality were constructed.

No non-linear models were applied first, because the study was focused on the change of forest cover only as predictor of deforestation (y - dependent variable) over the years of analysis (x - independent variable). Thus, not multivariate models were utilized; and second, because even though a non-linear model (e.g polynomial) may fit better the pattern of observations within the historical reference period, it may likely over-extrapolate results and introduce volatility beyond that period, considering the highly random nature of deforestation observations described above.

Finally, simple historical averages and linear regressions were applied to validate the historical reference periods within already existing historical and hypothetical monitoring periods. Therefore, subsequent analyses do not suggest that these models are the approach to follow in real projects because to assess how many years any trend could be realistic should be specific of each project. However, evaluating empirically the best method to project baselines into the future was outside of the scope of this paper. We nevertheless provide recommendations in the discussion section about how this should be done in future studies.

3.4 Accuracy evaluation

In order to assess the suitability of projected baselines, it is necessary to evaluate model performance against a hypothetical monitoring period. For linear regressions, the R2 indicator was used to assess model goodness of fit. The R2 is a widely used metric, which indicates the strength of the linear relationship. The greater the R2 the greater the agreement between projected values and sample observations. For this application we present the predicted R2, which is calculated by extrapolating the linear regression into the hypothetical monitoring period, and provides an idea of model agreement relative to the observed values within such validation period.

In order to test another indicator of model performance not based on normalized model efficiency but on true absolute error we used the RMSE. The use of this indicator has been reported for assessing the error of prediction models (Wooldridge, 2015), and it has also been used in forest cover change and deforestation research (Jaffé et al., 2021). The RMSE is essentially the sample standard deviation of the predicted errors, so a smaller RMSE is preferable to a larger one. In this work we use both, the R2 of the prediction and RMSE of the prediction, as described in Table 2. The differences in percentages between RMSE for the different historical reference periods are also presented in order to shed light into the what historical reference period results in lower errors when compared with the rest.

Table 2. Indicators used to test model performance.

Metric	**Equation**	**Parameter description**
R2	$r^2 = 1 - \frac{\sum_i^n (y - \hat{y})}{\sum_i^n (y - \bar{y})}$	y = *observed value* $\hat{y}$ = *estimated value* $\bar{y}$ = *mean of predicted values*
RMSE	$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y})^2}$	y = *observed value* $\hat{y}$ = *estimated y value*

4. Results

In order to assess the suitability of historical reference periods (4, 6, 8 and 10 years), hypothetical historical baselines were created in hectares of forest lost per year and were projected as linear regression models into a hypothetical monitoring period to assess projection performance as described for each of the scenarios.

4.1 Scenario A. Replicating Mertz et al., 2017 study for Colombia

Results indicate that a 10 years historical reference period shows a more consistent representation of the increasing deforestation during the last decade in all these municipalities and particularly for San Jose del Guaviare, San Vicente del Caguan and Cartagena del Chaira, which experienced greater deforestation with the significant change in forest national circumstances after the signature of the peace deal with FARC (Figure 3).

In contrast, a four years historical reference period performed very poorly for San Jose del Guaviare and consistently underestimated deforestation for all municipalities, except for La Macarena, which was the municipality with the lowest deforestation amongst the four municipalities selected (Figure 3).

A six years historical reference period showed erratic suitability, performing poorly for La Macarena and San Jose del Guaviare, and more consistently for Cartagena del Chaira and San Vicente del Caguan, but still somewhat underestimating deforestation after the signature of the peace deal with FARC (Figure 3).

Similarly, the eight years historical reference period, which ended in 2016, did not show suitable results for municipalities with a sharp increase in deforestation after the signature of the peace deal with FARC in 2016, such as San Jose del Guaviare and san Vicente del Caguan (Figure 3). So, the baseline projection from this latter historical reference period underestimated deforestation into the hypothetical monitoring period as it was not able to capture the sharp increase of deforestation in the year 2016.

4.2 Scenario B. Replicate of the VCS standard for AUDD projects

Results indicate that a 10 years historical reference period again shows a more consistent representation of the increasing deforestation in all the municipalities studied (Figure 4). In contrast, a four years historical reference period resulted in large overestimation of deforestation for all municipalities (Figure 4). This result can be explained by the fact that the four years historical reference period only captured the abrupt changes in forest cover after the year 2016 within the 2015 – 2018 period, but was not able to capture the lower deforestation seen in 2019, which can be attributed to a reduced number of fires during a wetter year. Therefore, this result shows that a four years historical reference period represents a sample of observations too small to characterize the highly random behavior of deforestation in these municipalities.

The mean historical average though for a four years historical reference period showed a better agreement when compared with the four years linear regression model. However, under an evident increasing trend the four years historical mean consistently underestimated the deforestation observed during the hypothetical monitoring period and therefore the linear regression model for a 10 years historical reference period showed a better agreement. Therefore, because a four years historical reference period represents a sample of observations too small to describe the highly random behavior of deforestation then it resulted in an underestimation of deforestation within an increasing deforestation trend of larger amplitude (Figure 4).

Similarly, a six years historical reference period showed a large overestimation of deforestation when using a linear regression and an underestimation when using a simple historical average, and when compared with the linear regression model of the 10 years historical reference period. Finally, the eight years historical reference period performed closer to the 10 years period but still showing a slightly higher (less conservative) projection.

In general, the simple historical average of the 10 years historical reference period did not vary significantly when compared with the shorter historical reference periods and therefore in the absence of a clear trend it should be used because a sample of 10 observations holds significantly lower standard and prediction errors when compared with smaller samples such as four or six years.

4.3 Scenario C. Five years hypothetical monitoring period (2011 - 2015) prior to the signature of Colombia's peace deal

For this scenario all the municipalities experienced a peak in deforestation around the year 2007 except for San Jose del Guaviare. This increasing trend in the mid 2000s can be explained by the fact that the failed peace process with FARC in 2002 resulted in greater deforestation in the municipalities covered by the demilitarized zone immediately after the process was cancelled (i.e. San Vicente del Caguan and La Macarena). During this transition period land speculators took advantage of the lack of military presence. When the army returned to these areas and the confrontation with FARC intensified, deforestation declined again and showed no obvious consolidated trend in the region between the years 2008 to 2015, increasing slightly in some municipalities and decreasing or remaining relatively constant in others.

Under these circumstances linear models fitted to four years historical reference periods (covering the years 2007 to 2010), performed poorly because they reported a declining trend that underestimated deforestation after the year 2008. In contrast linear regressions fitted to the 10 years historical reference periods performed much better. These models accurately projected a declining trend in La Macarena, and an increasing trend in Cartagena del Chaira (a municipality not affected much by the demilitarized zone). Although they overestimated deforestation slightly in San Vicente del Caguan and San Jose del Guaviare, where deforestation showed a more random behavior during the hypothetical monitoring period. For the latter municipalities the simple historical average model was a good predictor for the 10 years historical reference periods. However, simple historical averages of four years historical reference periods tended to overestimate deforestation because these only captured the spike in deforestation round 2007 after the failed peace process, but did not capture a scenario where armed forces where already well established in the territory. For San Jose del Guaviare, a four years historical reference period showed a relatively good performance because this municipality was not affected by the demilitarized zone and thus historical deforestation observations were less random. However, a 10 years historical reference period was able to capture better the dynamics of forest loss resulting from the interactions of armed actors. The 10 years baseline projections tended to be closer to the observed deforestation in the hypothetical monitoring period, either using a linear regression or a simple historical average.

Six and eight years historical reference periods showed performances within the range observed for four years and 10 years, again being closer to observed values when deforestation randomness was lesser and showing lower performance for municipalities most affected by the influence of the demilitarized zone.

4.4 Scenario D. Five years hypothetical monitoring period (2016 - 2020) after to the signature of Colombia's peace deal

For this scenario all the municipalities experienced a peak in deforestation in 2017 and 2018, which was not captured in any of the historical reference periods, as these were in contrast showing slight declines or relatively stable values. Therefore, all linear models performed badly (except for Cartagena del Chaira), because they were not able to capture such a sharp increase in deforestation after the year 2016.

However, the result was different for Cartagena del Chaira, where a linear regression fitted to a 10 years historical reference period showed to be the best model. This result can be explained by the fact that the latter municipality was showing a steady increase in deforestation, in contrast to the other three, which was captured within the historical reference period and was carried forward within the hypothetical monitoring period.

For the other municipalities apart from Cartagena del Chaira, the simple historical average model showed to be more appropriate, but still largely underestimating disforestation. Nonetheless, the underestimation was lower for simple historical average models fitted to a 10 years historical reference period. This finding can be explained by the fact that a 10 years period, for this scenario, represented the spike in deforestation in 2007 after the failed peace process with FARC, while shorter reference periods did not. This outcome highlights the importance of using longer historical reference periods to characterize the highly random nature of deforestation in the arc of deforestation of Colombia.

Nevertheless, it is important to note that when a sudden nonlinear shift in forest national circumstances occurs, projects that may have established their baselines just before the sudden change should consider revising their baseline earlier since projects could clearly start heavily relying on their non-permanence buffer, when deforestation increases significantly. If deforestation significantly declines in proxy areas, projects should also consider revising their baseline earlier providing nonlinear shifts have not been observed in the historical reference period. If projects' baselines are adjusted and then nonlinear shifts do occur, then projects could start relaying heavily on their pool buffer.

However, the fact that the 10 years historical reference period showed a more consistent agreement for simple historical average models, under the highly abrupt forest cover changes observed after 2016, indicates that longer historical reference periods would be the best alternative to generate project baselines. Nevertheless, projects should be able to choose a shorter baseline reassessment period, depending on whether historical deforestation follows a clear trend or shows to be highly random, while ensuring conservativeness.

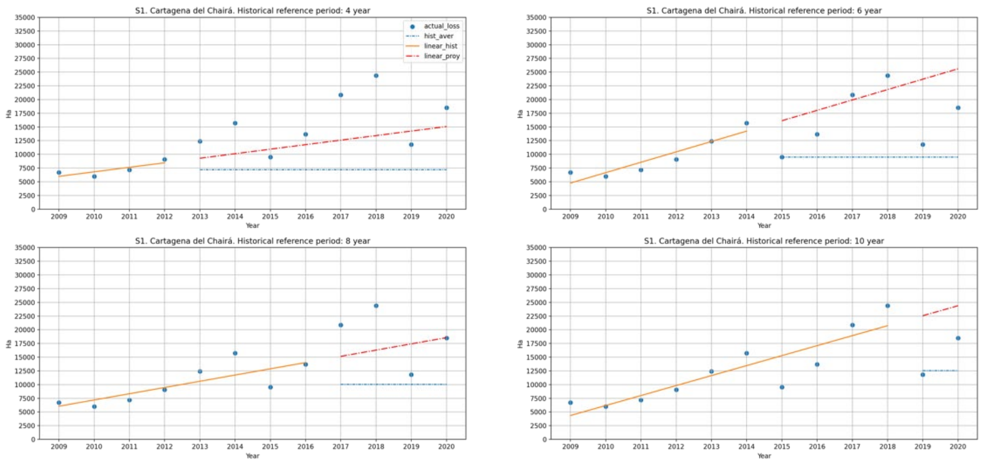
4.5 Scenario E. 10 years hypothetical monitoring period (2011 - 2020) coinciding with the signature of Colombia's peace deal

For this scenario, models fitted using a 10 years historical reference period show the most consistent and, in most cases, the best performance, even capturing the abrupt change in forest national circumstances after the year 2016. Particularly, linear models fitted to a 10 years historical reference period more optimally captured the upward trend in deforestation since the year 2001 in the municipalities of Cartagena del Chaira, San Jose del Guaviare and San Vicente del Caguan, and the downward trend in La Macarena.

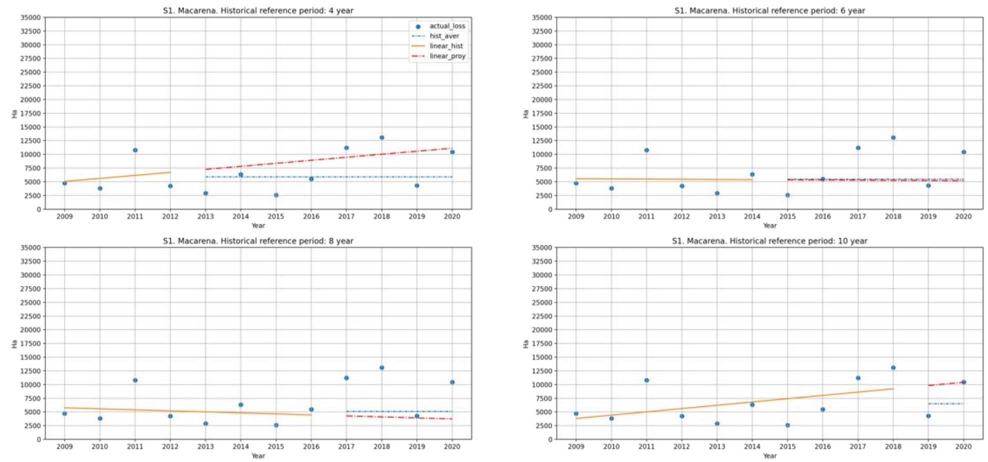
In contrast, linear regressions applied to four years historical reference periods performed poorly for all municipalities except for San Jose del Guaviare, where there was a slight upward trend in deforestation between 2007 and 2010, which on the one hand connected with the trend of deforestation between 2019 and 2020, but on the other hand over estimated deforestation in the immediate four years, during which the four years baseline should have been valid.

Similarly, mean historical averages show a large underestimation of deforestation for all municipalities, and all historical reference periods, indicating that the simple historical average model was unsuitable to capture the significant increases in deforestation after the year 2016 even when all models captured a peak in deforestation around 2007. Therefore, projects should be able to choose between linear models or simple historical average models, depending on whether historical deforestation follows a clear trend or shows to be highly random, while ensuring conservativeness.

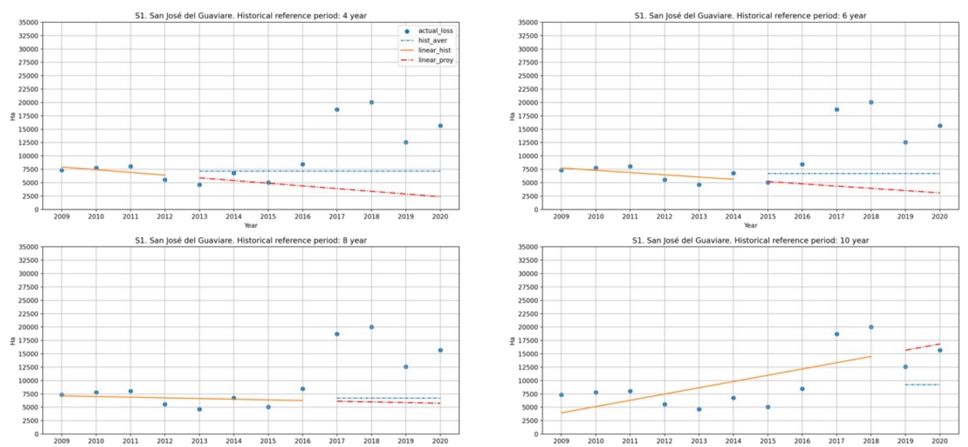
Cartagena del Chaira



La Macarena



San Jose del Guaviare



San Vicente del Caguan

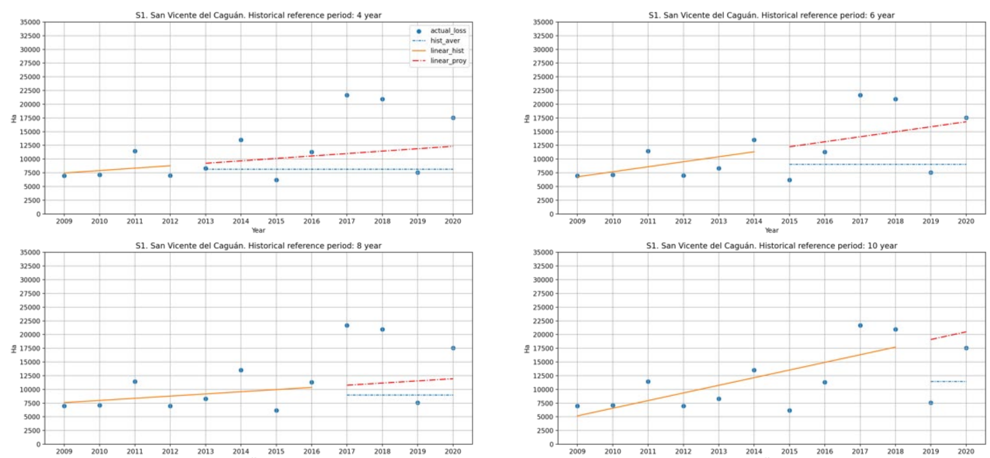
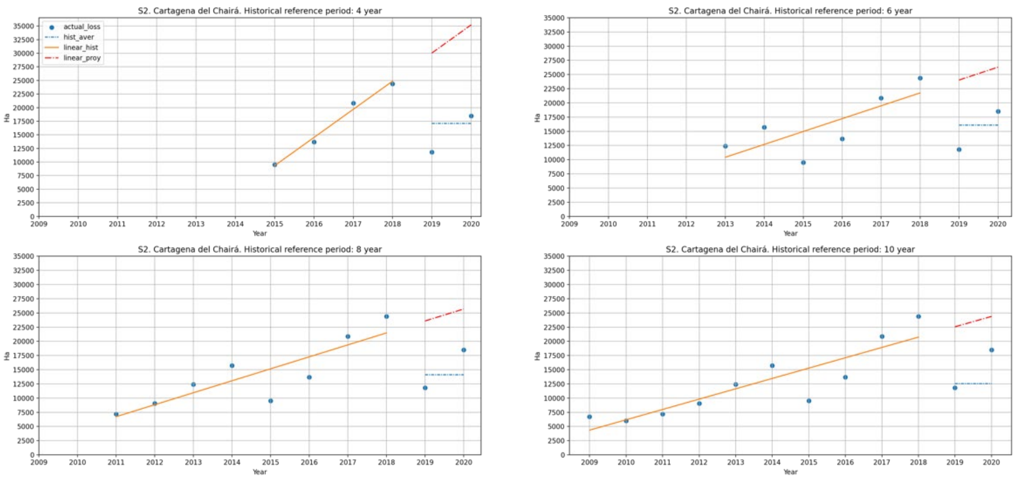
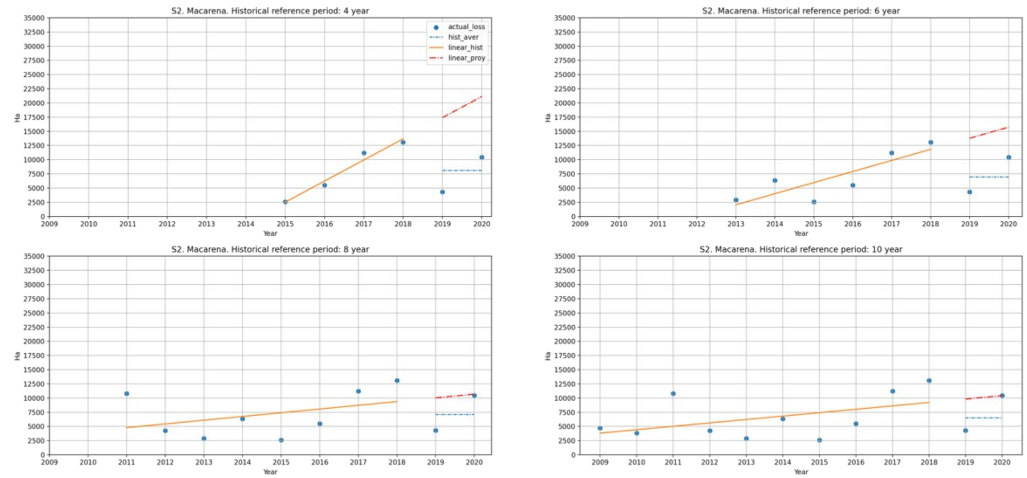


Figure 3. **Scenario A. Replicating Mertz et al., 2017 study for Colombia**

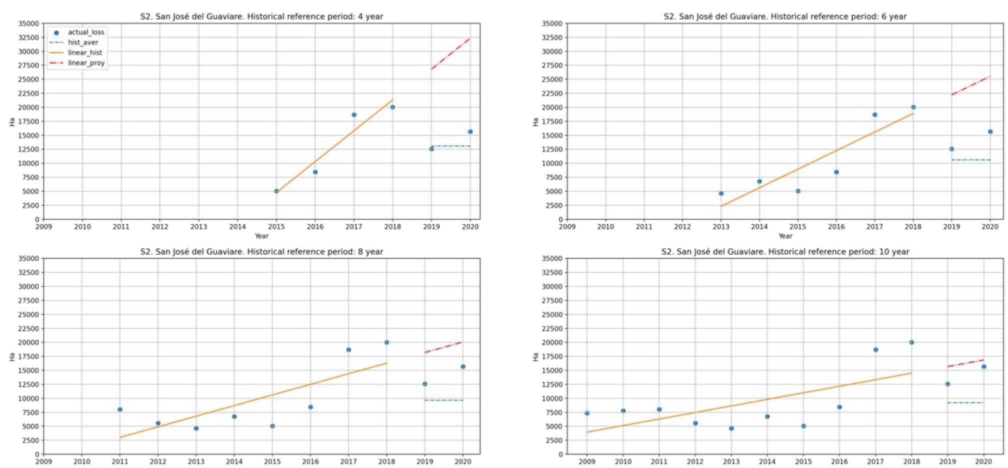
Cartagena del Chaira



La Macarena



San Jose del Guaviare



San Vicente del Caguan

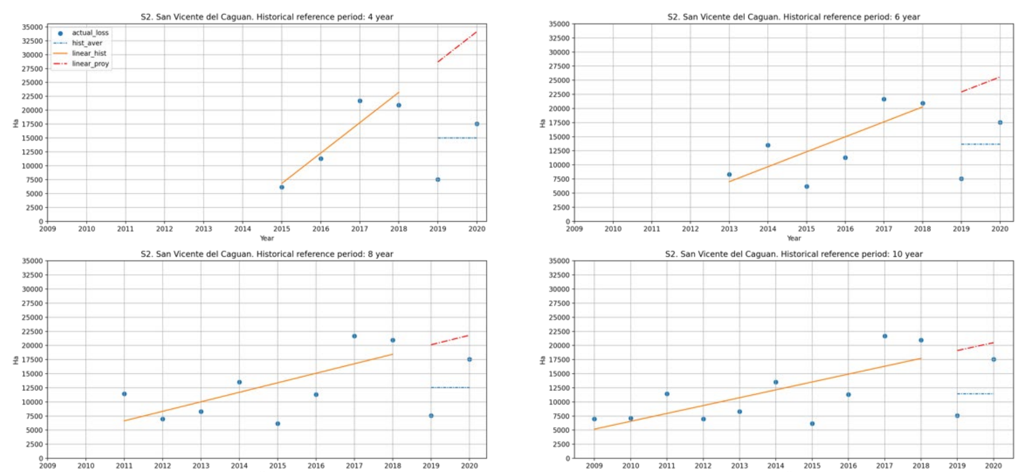
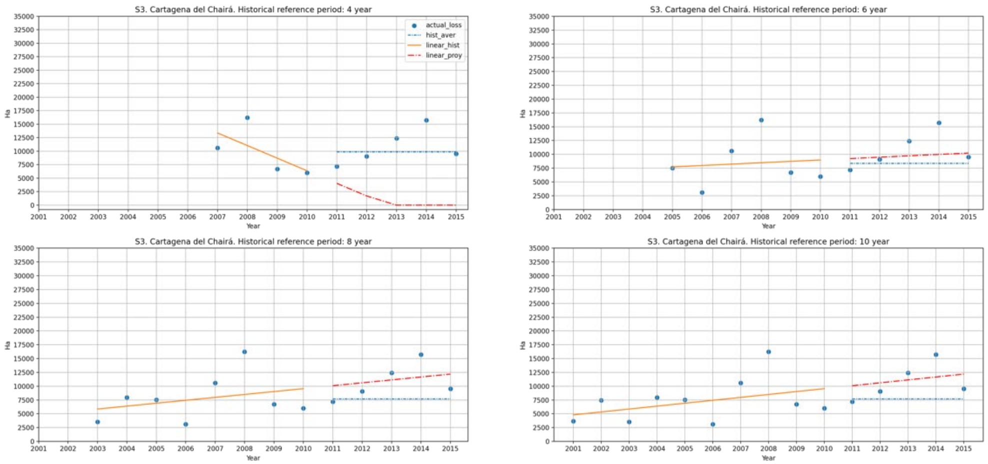
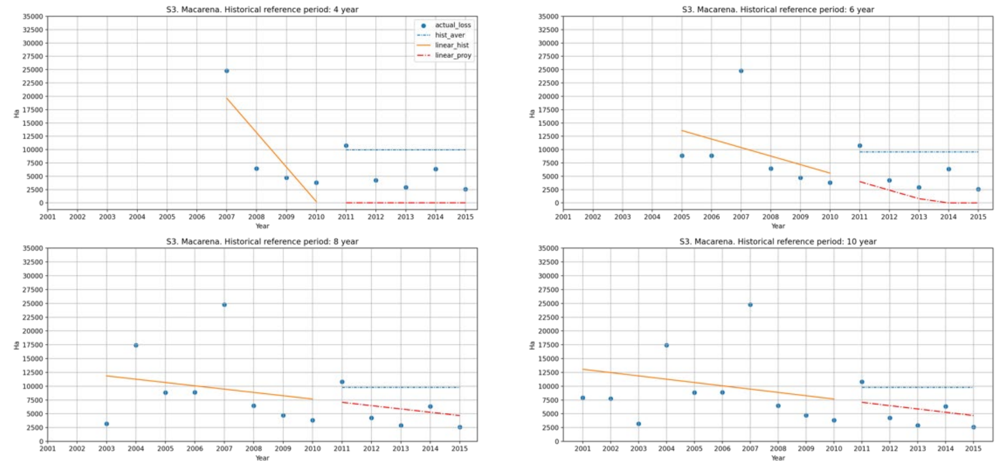


Figure 4. Scenario B. Replicate of the VCS standard for AUDD projects

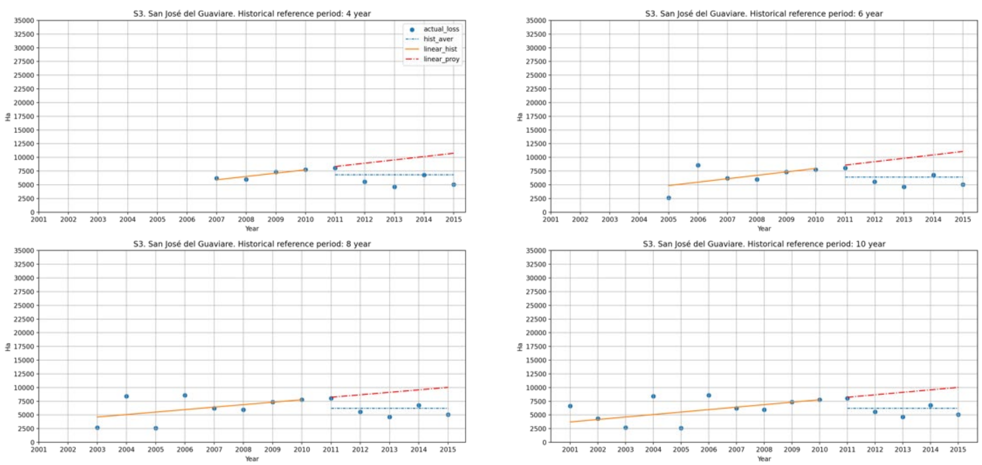
Cartagena del Chaira



La Macarena



San Jose del Guaviare



San Vicente del Caguan

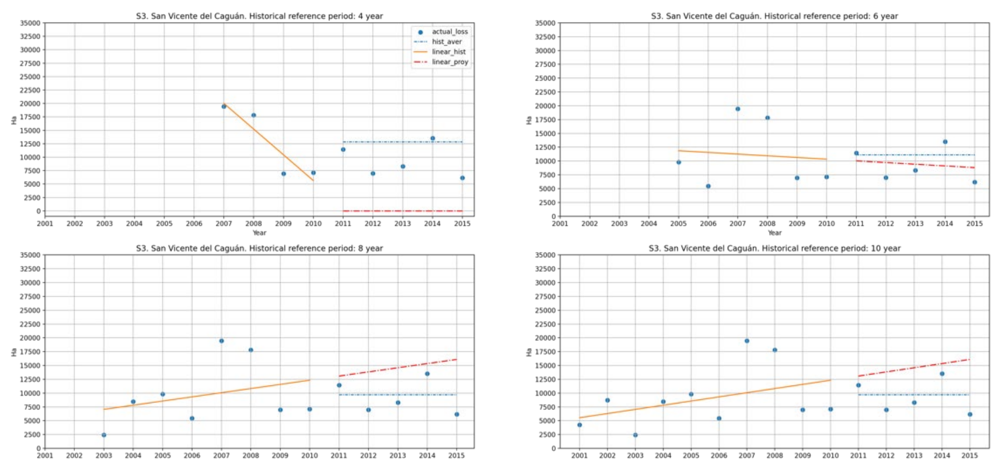
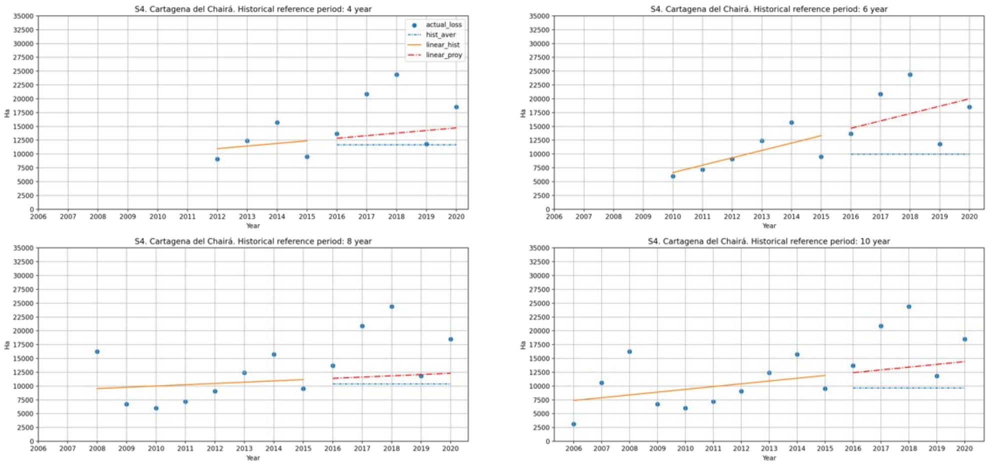
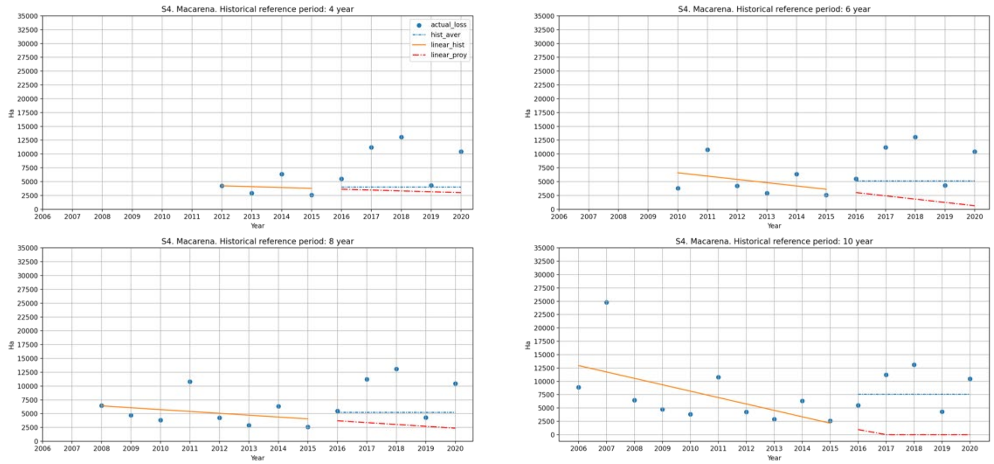


Figure 5. Scenario C. Five years hypothetical monitoring period (2011 - 2015) prior to the signature of Colombia's peace deal

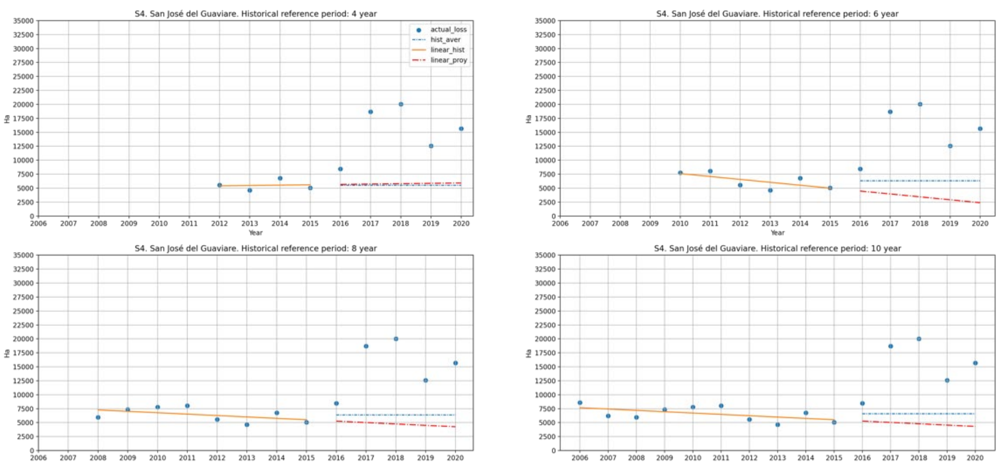
Cartagena del Chaira



La Macarena



San Jose del Guaviare



San Vicente del Caguan

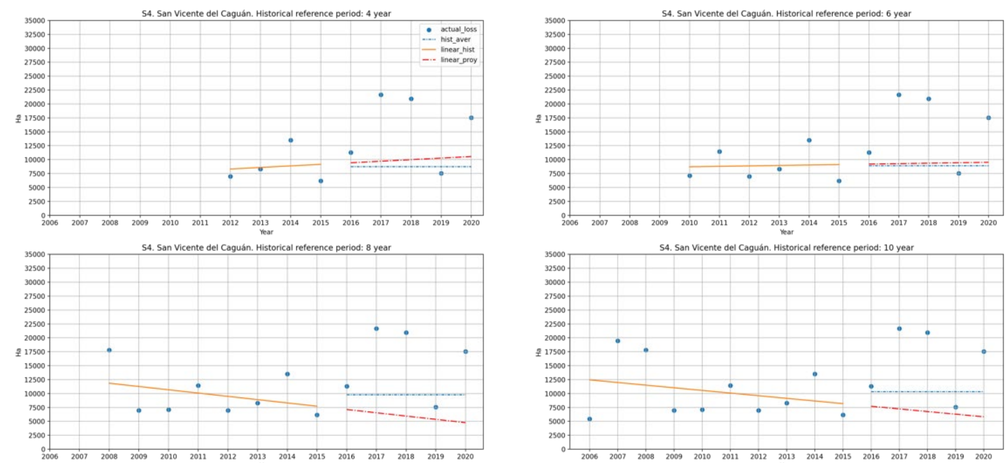
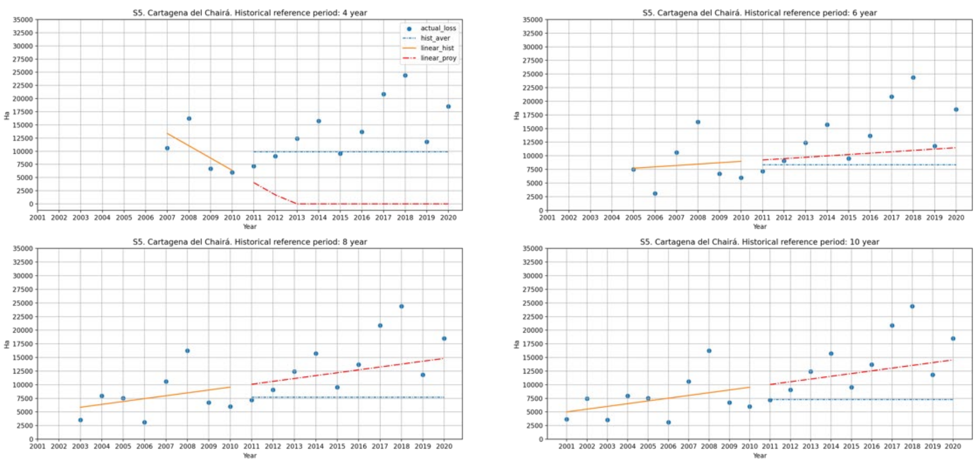
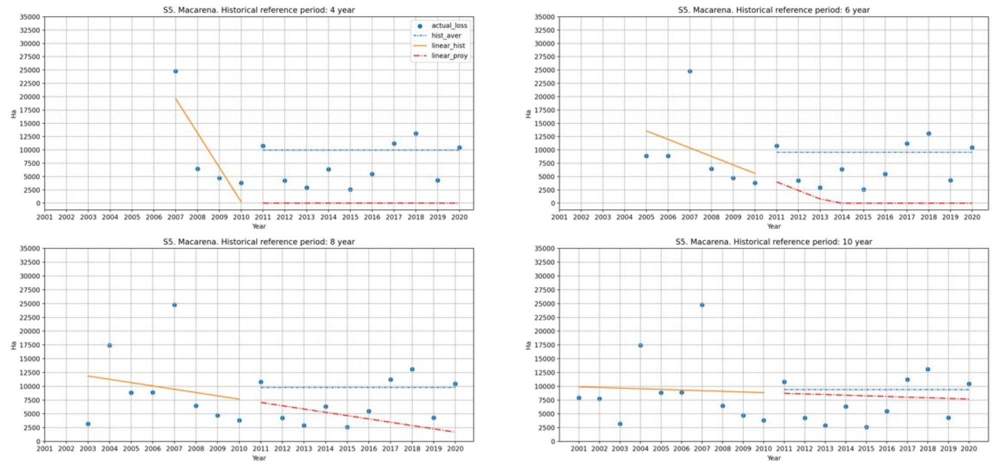


Figure 6. Scenario D. Five years hypothetical monitoring period (2016 - 2020) after to the signature of Colombia's peace deal

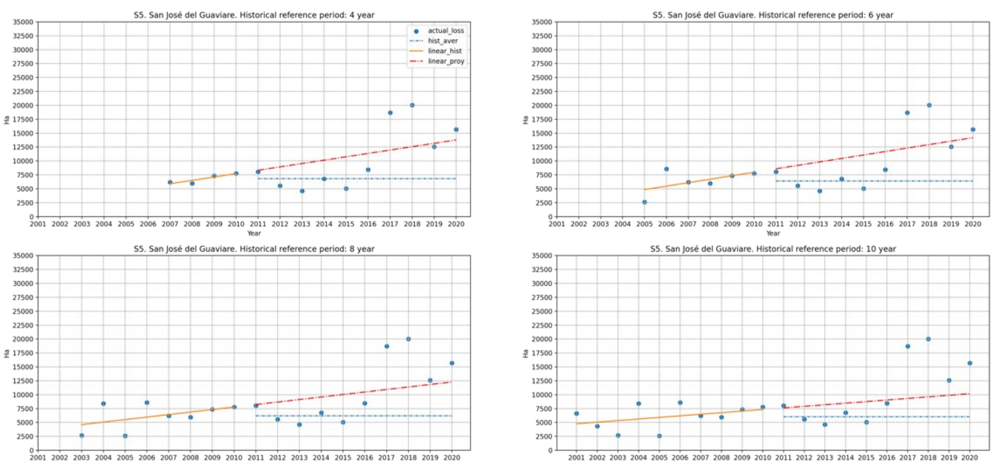
Cartagena del Chaira



La Macarena



San Jose del Guaviare



San Vicente del Caguan

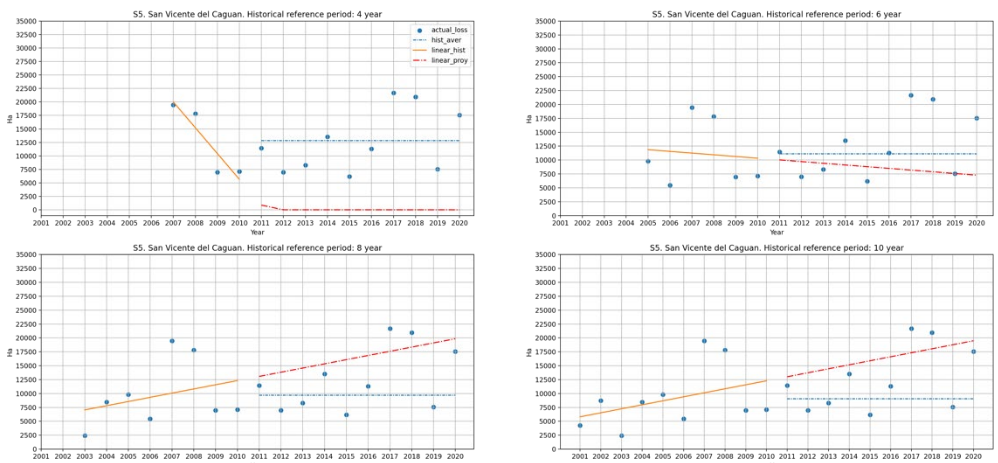


Figure 7. Scenario E. 10 years hypothetical monitoring period (2011 - 2020) coinciding with the signature of Colombia's peace deal

4.6 Model Performance

As indicated earlier, we used two indicators of model performance to illustrate what historical reference periods resulted in more optimal linear regressions, namely R2 and RMSE. We chose to test linear regressions only at this stage of analysis because the simple historic average model showed to significantly underestimate deforestation particularly for scenarios coinciding with the abrupt change in forest national circumstances in Colombia in 2016.

4.6.1 R2

As indicated in the methods section, we used data from all the 80 linear regression to produce a box plot to compare the performance of linear models fitted to each of the historical reference periods. We chose to collapse all municipalities and scenarios into one plot to reduce biases in the comparison, which might result from particularities of each municipality or from the assumptions made when creating each of the scenarios as described in the previous section (i.e. start or end of historical reference periods relative to the change in forest national circumstances in Colombia). This approach to model comparison was assumed to have an adequate statistical representation since each historical reference period accounted for a population of 20 regressions to derive the sample of R2s.

The analysis shows a consistent significantly higher performance of linear models fitted to 10 years historical reference periods when compared with four years linear models (Figure 8). Normalized R2s for 10 years historical reference periods have an average performance of 0.6 with the top 25% of normalized R2s reaching 0.7 on average and the bottom 25% down to around 0.45 on average. In contrast, Normalized R2s for 4 years historical reference periods have an average performance of 0.4 with the top 25% of normalized R2s reaching around 0.52 on average only and the bottom 25% down to around 0.27 only.

In addition, the spread of observations was much greater for the four years regressions highlighting the greater volatility of linear models created using only four points in time. It is important to highlight that the analysis presented here uses normalized R2s to be able to fit all values in a scale from 0 to 1, but predictive R2s can produce negative values due to the highly random nature of deforestation observations and particularly for municipalities showing no clear deforestation trends over the 2001 – 2020 period (Appendix 1, Figure 1).

Finally, model performance varies significantly between municipalities depending again on the randomness of deforestation distributions. For municipalities showing a steadier increasing deforestation trend (San Jose del Guaviare and Cartagena del Chaira) linear models built using a 10 years historical reference period showed significantly better performance than using a four years historical reference period (Appendix 1, Figure 2), particularly for San Jose del Guaviare, a municipality with the highest deforestation after the signature of the peace deal with FARC in 2016 (and very little affected by the failure of the first peace process with FARC). For municipalities with a more random distribution differences in performance were less pronounced but a 10 years historical reference period consistently shows higher performance than the other historical reference periods tested (Appendix 1, Figure 2).

4.6.2 RMSE

In order to test another indicator of model performance not based on normalized model efficiency but on true absolute error we used the RMSE. In a similar way to the R2 analysis we used data from all 80 linear regression to produce a box plot to compare model performance for each of the historical reference periods.

Similarly, to the R2 analysis, the RMSE shows that linear models fitted to 10 years historical reference periods resulted in consistently lower prediction errors and that models fitted to a four years historical reference period not only resulted in greater absolute errors but also in a much more volatile distribution of errors when compared with the other three historical reference periods (Figure 9).

Prediction errors in the four years sample were close to 13,000 hectares on average with the top 25% of errors averaging close to 20,000 hectares and the bottom 25% averaging around 6,000 hectares. In contrast, 10 years models showed prediction errors of just above 5,000 hectares on average with the top 25% of errors averaging around 8,000 hectares and the bottom 25% averaging around 3,750 hectares. Models fitted to six and eight years historical reference periods resulted in lower errors than those fitted to 4 years period, but still significantly higher than 10 years historical reference period, when comparing means (Figure 10; Table 4).

Finally, and in a similar way to the R2 analysis, model performance according to the RMSE varies significantly between municipalities depending again on the randomness of deforestation distributions and on the size and scale of deforestation in each of the municipalities. For San Jose del Guaviare, where a steadier increasing trend was observed during the 2001 – 2020 period, all models showed lower errors when compared with the other municipalities. Nevertheless, models fitted to the 10 years historical reference period consistently showed a much better performance (Appendix 1, Figure 3).

For the other three municipalities, models using a four years historical reference period resulted in much larger errors than the other three periods, highlighting the high volatility that using such a short historical reference period could introduce to predictions. Although, models showed comparable errors for six, eight and 10 years historical reference periods, it is worth noting that for Cartagena del Chaira, a six years historical reference period showed a smaller error spread when compared with the 10 years historical reference period, but almost identical median, which could be attributed to the fact that six years historical reference period for this municipality in general coincided with somewhat stable deforestation. In contrast, in San Vicente del Caguan a six years historical reference period showed a greater error spread than both the eight- and 10 years periods, while having a similar median. Whereas in La Macarena, a 10 years historical reference period resulted in the best performance of linear models amongst all the historical reference periods, with a four years period performing the worst amongst all the municipalities tested (Appendix 1, Figure 3).

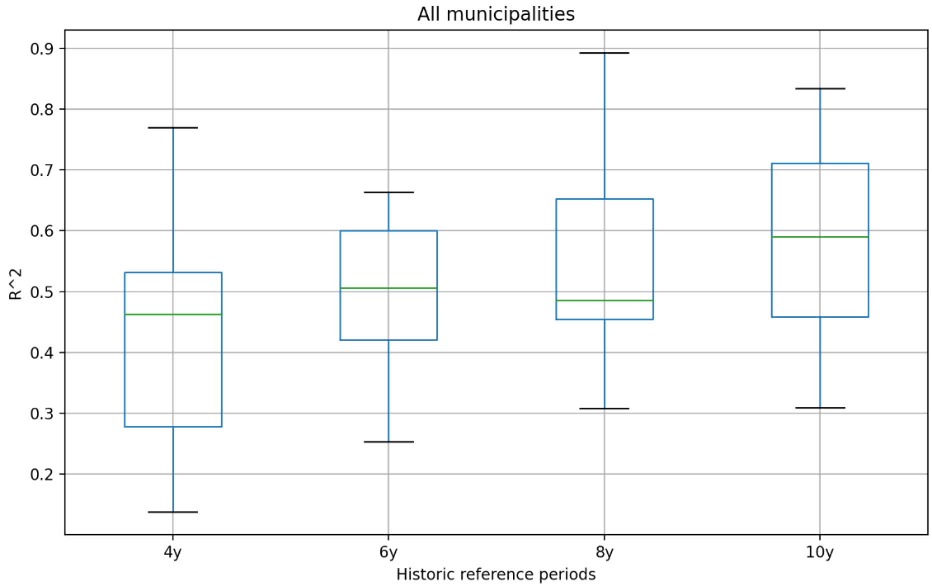


Figure 8. Normalized R2 of 80 linear regressions plotted on a single graph to compare model performance between different historical reference periods and scenarios.

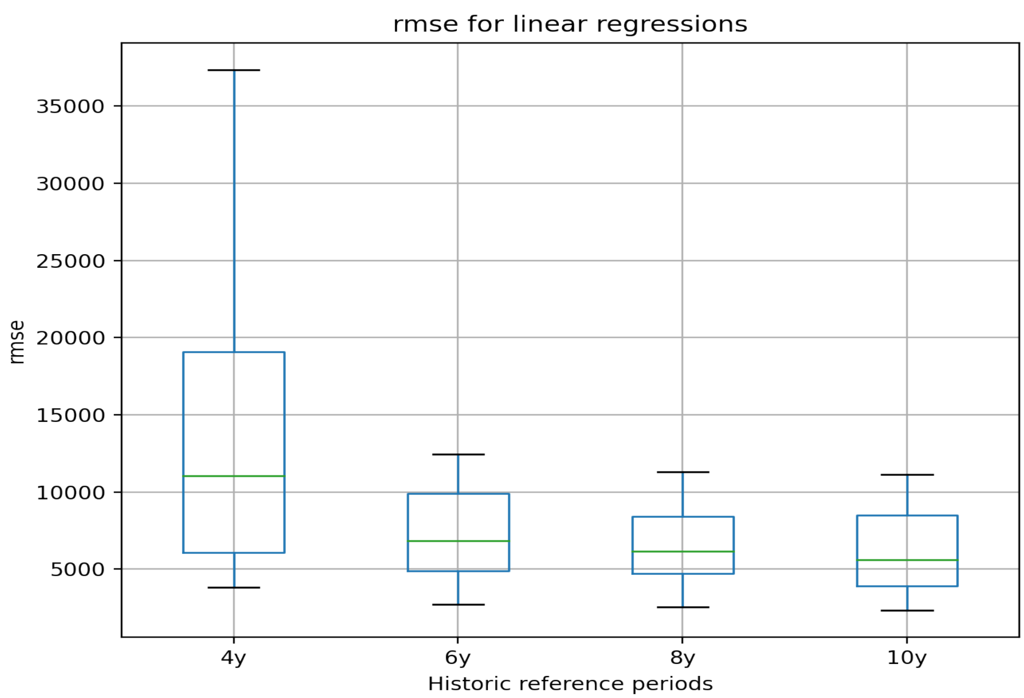


Figure 9. RMSE of 80 linear regressions plotted on a single graph to compare model performance between different historical reference periods and scenarios.

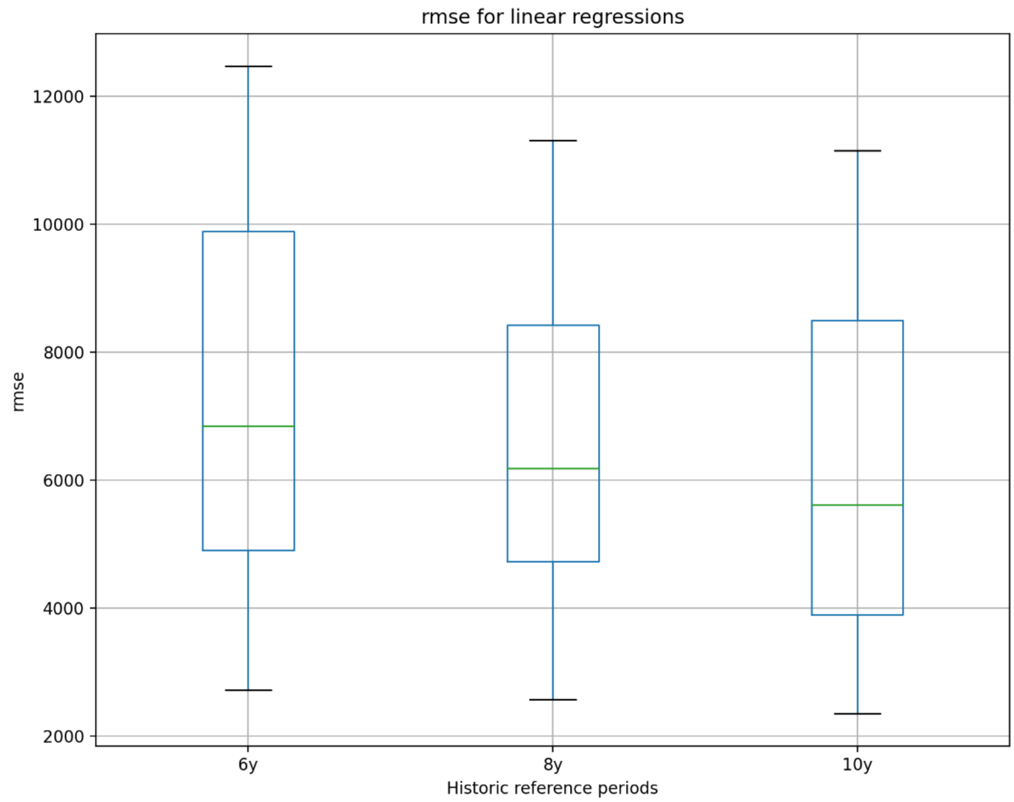


Figure 10. RMSE of linear regressions plotted on a single graph to compare model performance between six, eight and 10 historical reference periods and scenarios. The four years historical reference period is not considered in this graph due to the volatility of predictions using this historical period.

What's more, when comparing the RMSEs of linear models between historical reference periods and for all municipalities, it can be observed that regression errors in the 10 years historical reference period were around 60% lower on average when compared with errors obtained from regressions using four years historical reference periods (Table 3). In addition, errors were 50% and 44% lower for eight and six years respectively when compared with the four years historical reference period, which highlights the volatility of predictions resulting from using such a small number of observations.

Furthermore, RMSEs are compared against the 10 years historical period in order to get a better idea of error differences with regards to six and eight years historical reference periods, which show to be less volatile than the four years period (Table 4). It was found that errors using a 10 years historical reference period were 19.2% on average lower than for regressions using a six years period and 10.5% lower than for regressions using an eight years historical reference period. Therefore, the use of a 10 years historical reference period results in statistically significant lower projection errors than using six and eight years historical reference periods for these municipalities in Colombia. Although, it is important to note that for San Vicente del Caguan the average error (%) was lower for six years than for 10 years. This result could be explained by the fact that a six years period showed better agreement for scenario C, capturing a decreasing deforestation trend better for this municipality. Nonetheless, a 10 years period showed better performance for other scenarios applied to this municipality. These results highlight the random nature of deforestation in these areas of the Colombian Amazon.

Table 3. RMSE error comparison (%) between 4 years historical reference periods and other periods. The average value shows how small the error (%) is when compared with four years projections. It is calculated as (xy-4y)/average(xy,4y). When results are negative, it means that the RMSE of four years is greater on average, and thus worse than the period it is compared against. In the equation, “x” refers to the error of a particular year “y”.

Municipality	4y vs 6y (%)	4y vs 8y (%)	4y vs 10y (%)
Cartagena	-59,2	-59,6	-56,2
San Vicente	-65	-49,8	-52,4
San José	-1,6	-17,6	-57
La Macarena	-51,2	-73	-72,4
AVERAGE	**-44%**	**-50%**	**-60%**

Table 4. RMSE error comparison between 10 years historical reference periods and six and eight years historical reference periods. The average value shows how smaller the error (%) is for 10 years projections, when compared with six and eight projections. It is calculated as (xy-10y)/average(xy,10y); when results are positive it means that the RMSE of 10 years is lower on average, and thus more precise than the period it is compared against. In the equation, “x” refers to the error of a particular year “y”.

Municipality	6y VS 10y (%)	8y vs 10y (%)
Cartagena	0,02	-3,06
San Vicente	-16,36	3,47
San José	58,70	44,20
La Macarena	34,30	-2,54
AVERAGE	**19,2**	**10,5**

5. Discussion

5.1 The role of dense annual time series of land cover data

The analysis above shows that more data observations within a deforestation sample, in general consistently result in a statistically better model performance of linear models used to project deforestation baselines. We arrived at this conclusion by comparing the R2 and RMSE of 80 regressions resulting from five scenarios applied to four municipalities in the arc of deforestation of Colombia and assuming four distinct historical reference periods (4, 6, 8 and 10 years). However, it is important to note that deforestation has shown to be a highly random phenomenon within the historical time series analyzed, as well as a highly variable activity when comparing the different municipalities under study.

5.2 What ranges of historical reference periods could be more optimal

When replicating the Mertz et al., 2017 study, starting in 2009 (to allow for a 10 years period to have at least two years monitoring period), models using a 10 years historical reference periods performed better. This outcome could be explained by the fact that 10 years of observations managed to capture the abrupt change in forest national circumstances in Colombia in 2016, which shorter historical reference periods were unable to represent.

When testing the scenario designed to replicate the requirements of the VCS standard in terms of the end of the historical reference period relative to project start date, again the 10 years historical reference period showed a more consistent representation of the increasing deforestation in all the municipalities studied, because 10 years of observations not only captured the change in forest national circumstances in 2016, but also because it effectively represented lower deforestation rates in previous years, resulting in a more conservative projection of the baseline. In contrast, a four years historical reference period resulted in large overestimation of deforestation for all municipalities because it only captured the change in national circumstances in 2016 but failed to capture lower deforestation seen in 2019, which can be attributed to reduced fire frequencies during a wetter year.

Nevertheless, the mean historical average though for a four years historical reference period showed a better agreement than the four years linear regression model. However, it was still unfit to replicate the increasing deforestation trend of larger amplitude observed between 2016 and 2020, resulting therefore in an important underestimation of deforestation. Therefore, even when simple historical averages obtained from shorter historical reference periods - such as four years - can show some agreement, the analysis shows that using a longer historical reference period is recommended because the longer the prediction sample the lower the standard and prediction errors of the models (Table 3 and Table 4), and therefore projections are likely to be less volatile.

When we tested the performance of models designed to project five-year baselines just before the change in forest national circumstances in Colombia, models based on four years historical reference period performed very poorly because they only captured a decline in deforestation observed after the mid 2000s, underestimating deforestation within the monitoring period. Models fitted using a 10 years historical reference period performed more consistently. A simple historical average model performed better for four years projections. However, simple historical average projections still significantly overestimated deforestation because a four years historical reference period only captured the spike in deforestation after the failed peace process with FARC after 2002 but did not capture a scenario where armed forces were again well established in the territory. Nonetheless, a 10 years historical reference period was able to capture the dynamics of armed actors in the territory better, and therefore predictions using either a linear regression or a simple historical average tended to be closer to the observed deforestation in the hypothetical monitoring period.

When testing the performance of models designed to project five-year baselines after the signature of the peace deal with FARC, all linear models performed badly (except for Cartagena del Chaira), since most municipalities were depicting a slight deforestation decline in the historical reference period and thus models were not able to capture the sharp increase in deforestation in 2016. For this particular case, a simple historical average model for all historical reference periods showed to be more appropriate, but still largely underestimating deforestation in all historical periods. Nevertheless, the 10 years historical reference periods showed less underestimations of deforestation because 10 years of observations captured better the spike in deforestation after the failed peace process with FARC in the 2000s, as described earlier. The latter outcome highlights the importance of having longer historical reference periods as they are able to better capture the random nature of deforestation in Colombia, which has been subject to important nonlinear shifts over the last two decades.

When testing the performance of models designed to project 10 years baselines coinciding with the signature of the peace deal with FARC, the 10 years historical reference period showed to be the most consistent and, in most cases, exhibited the best performance amongst the different historical reference periods tested. Such 10 years baselines even captured the abrupt change in forest national circumstances in 2016, and replicated plausibly the particular trends observed in each of the municipalities. Nevertheless, a four years historical reference period performed relatively well for San Jose del Guaviare, but the linear agreement came only five years after the end of the historical baseline period. This result meant that the four years baseline effectively overestimated deforestation largely over the following four years after the end of the historical reference period when the baseline should have been valid, so this apparently more satisfactory performance should be evaluated with caution.

In Summary, the analysis above indicates that hypothetical baseline projections using 10 years historical reference periods tended to be more optimal for all scenarios and municipalities. This result can be explained by the fact that model performance was higher for 10 years periods and consistently showed lower projection errors. Similarly, the simple historical mean model showed to be more robust when derived from 10 years of observations as longer historical reference periods captured nonlinear shifts better at least for these municipalities in the Colombia’s arc of deforestation.

5.3 Implications for the extent of additionality

This study uses simple historical averages and linear regressions to validate the historical reference periods within already existing historical and hypothetical monitoring periods. Therefore, the analysis presented here does not suggest that these models are the approach to follow in real projects because to assess how many years any trend could be realistic should be specific of each project. Thus, projects should continue to use the best methodologies available to determine what models to use to project deforestation baselines.

However, it is important to clarify that a greater analytical effort should be given to define how the baseline is projected to ensure an adequate additionality assessment. When the population of deforestation observations follow a plausible trend, linear models can be more suitable as shown here, even projecting deforestation well within 10 years hypothetical monitoring periods. However, when deforestation observations are highly random, then the use of simply historic average models could be more optimal. The latter models showed to underestimate deforestation largely in Colombia though, after the year 2016, which indicates that it would be plausible to consider a shorter reassessment period when choosing this model approach.

It could also be plausible to recommend that baselines created using a linear projection also use a shorter reassessment period to ensure adjustment for any plausible change in trend. Therefore, projects should be able to choose between linear models or simple historical average models, or rolling averages, depending on whether historical deforestation follows a clear trend or shows to be highly random, and to consider a shorter reassessment period, while ensuring conservativeness.

5.4 Implications for reassessment periods

It is important to note that deforestation has shown to be a highly random phenomenon within the historical time series analyzed, as well as a highly variable activity when comparing the different municipalities under study, especially considering the abrupt change in forest national circumstances occurred in 2016 after the signature of the peace deal with FARC. Therefore, our analysis recommends that when a sudden and significant nonlinear shift in deforestation is identified, projects that established their baselines before such a change should consider revising their baseline earlier since projects will clearly have to start relying on their risk buffers significantly. A five years reassessment period should be considered plausible.

However, the fact that hypothetical baselines using 10 years historical reference periods showed to represent better the highly abrupt forest cover change in this region of Colombia in 2016, indicates that longer historical periods are always better to generate project baselines, while perhaps choosing a shorter baseline reassessment period when significant nonlinear shifts can be verified.

It is important to highlight though that a reassessment period of five years for projects starting in 2011 (this is a reassessment in 2016) would have resulted in important underestimation of deforestation, whereas a 10 years reassessment period (this is a reassessment in 2021) would have not, at least for these Colombian municipalities. The latter finding highlights the fact that baseline reassessments should also consider the history of deforestation in the area and if interannual nonlinear shifts are identified in the historical record, maintaining a 10 years historical reassessment period could still be plausible, especially in areas with a confirmed increasing trend of large amplitude, as it is the case in the high forest cover high deforestation areas of the arc of deforestation in the Colombia's Amazon.

5.5 Limitations

The limitations of the study are associated to the assumptions made in the design of the scenarios tested. This is that the start and end of historical reference periods were selected either to make sure to allow a minimum of two years as a hypothetical monitoring period for all historical periods; or to test how the different historical periods predicted monitoring years before or after the abrupt changes in forest national circumstances in Colombia in 2016. For this case scenarios B and E were the most optimal to evaluate transitions because they did not end or start just before or after these changes. Particularly Scenario B was very useful because the historical reference period captured two years after the signature of the peace deal with FARC in 2016. However, a more extensive array of scenarios could be explored in future studies, which was beyond the scope of the present study.

In addition, the length of the global forest watch data of 20 years only, prevented the inclusion of previous years, which could have been useful to confirm an increasing deforestation trend before the year 2000. Also, for the municipalities evaluated an increasing trend of deforestation was identified over the last 20 years, and thus linear models showed to perform better than simple historical averages. It may be that this deforestation characteristic is not the case in other large tropical forests basins, but testing this question is outside of the scope of this paper. Nevertheless, the approach undertaken to compare 80 multiple regressions generated for these different historical reference periods, municipalities and scenarios, ensured an unbiassed assessment of historical reference periods and their suitability for setting project baselines in this region of Colombia. Although, we recognize that it is difficult to extrapolate general conclusions across the tropics from these sites in Colombia, where deforestation has been going up over this time period, we recommend that this type of the analysis should be replicated in future studies covering all tropical forest basins to better inform the setting of historical reference periods for REDD+ projects.

6. Conclusions

Our study shows that deforestation has been a highly random phenomenon over the last two decades at least for the municipalities studied in the arc of deforestation of Colombia. To capture this variability when setting project baselines, our results indicate that more data observations within a historical deforestation sample, in general, consistently result in statistically better projections. For this analysis, hypothetical baselines based on 10 years historical reference periods tended to be more optimal for all scenarios and municipalities under analysis. The latter finding suggests that longer historical reference periods should always be preferred over shorter ones when generating project baselines in these areas of Colombia, which have shown an increasing deforestation signal, which was not adequately captured within short historical reference periods.

Nevertheless, our results highlight the fact that greater analytical effort should be given to define how the baseline is projected into the validity period to ensure an adequate additionality assessment. In this direction, our study shows that when the population of deforestation observations follows a plausible trend, linear models can be more suitable, even projecting deforestation well within 10 years hypothetical monitoring periods. In contrast, when deforestation observations are highly random, then the use of simple historical average models could be more optimal. However, simple historical average models showed to underestimate deforestation largely in these Colombian municipalities after the year 2016, because they failed to capture an increasing deforestation signal, which could result in projects being incorrectly labelled as of lower additionality, while making them rely largely on their pool buffers.

In terms of reassessment periods, our study highlights the fact that baseline reassessment efforts should always consider the history of deforestation in the area and if interannual nonlinear shifts are identified in the historical record, maintaining a 10 years historical reassessment period could be plausible. This scenario is particularly probable especially in areas with a confirmed increasing deforestation signal, with observed deforestation waves of larger amplitude within the historical record. This case is precisely the one identified in the high forest cover high deforestation areas of the arc of deforestation of Colombia, a region that continues to face increasing and often significant migration and legal and illegal development pressures. However, it should be plausible for projects to consider a shorter reassessment period (i.e. five years) if significant nonlinear shifts are confirmed to have occurred after the project start date and if such shifts are not properly represented in the historical reference period.

In summary, our study concludes that short term changes in deforestation trends should not be used to disqualify projects when no context is provided in terms of the length of the historical reference period used and the degree of randomness of the deforestation observations for the area. Therefore, a representative number of observations within a historical reference period have to be chosen to avoid biases, because non-linear shifts can always occur. Although, the purpose of this study is no to extrapolate general conclusions across the tropics from these sites in Colombia, where deforestation has been steadily increasing over this time period, we recommend that this type of the analysis should be replicated in all tropical forest basins, by the most robust standards, to better inform the setting of historical reference periods for REDD+ projects.

Finally, some REDD+ studies published in the literature have argued that project baselines are commonly inflated. However, several of those studies have often failed to replicate all relevant structural determinants of deforestation affecting project areas; or their periods of analysis have been insufficient to better inform their additionality conclusions. In addition, other studies have stressed the fact that shorter historical reference periods may not be the only alternative to consider, when setting project baselines, because data and funding limitations, as well as political, social and regulatory barriers may also play a role. This is not to say that all projects are additional across the voluntary market, but to say that the majority of projects inflate their baselines is safely exaggerated. Our study has shown that one of the solutions to this is not to shorten the historical reference periods to create project baselines, but to ensure that project baselines are better projected into the future, and to make sure baseline reassessment efforts consider the history of deforestation in the area, particularly the evidence of interannual nonlinear shifts in the historical record. With the increasing requirements for projects to nest within larger jurisdictional initiatives the need to consider these spatial and temporal differences in deforestation becomes paramount to guide and continue to incentivize optimal investment decisions and to preserve the integrity and additionality of projects.

7.References

Ankersen, J., Grogan, K., Mertz, O., Fensholt, R., Castella, J.-C., Lestrelin, G., Nguyen, D. T., Danielsen, F., Brofeldt, S., & Rasmussen, K. (2015). Vietnam's forest transition in retrospect: Demonstrating weaknesses in business-as-usual scenarios for REDD+. *Environmental Management*, *55*(5), 1080–1092.

Burton, C., Betts, R. A., Jones, C. D., Feldpausch, T. R., Cardoso, M., & Anderson, L. O. (2020). El Niño driven changes in global fire 2015/16. *Frontiers in Earth Science*, *8*, 199.

FAO, 2014. emerging approaches to Forest Reference Emission Levels and/or Forest Reference Levels for REDD+. www.fao.org/publications.

Gibbs, H. K., Brown, S., Niles, J. O., & Foley, J. A. (2007). Monitoring and estimating tropical forest carbon stocks: Making REDD a reality. *Environmental Research Letters*, *2*(4), 045023.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850–53. Data available on-line from:http://earthenginepartners.appspot.com/science-2013-global-forest.

Huettner, M., Leemans, R., Kok, K., & Ebeling, J. (2009). A comparison of baseline methodologies for "Reducing Emissions from Deforestation and Degradation." *Carbon Balance and Management*, *4*(1), 4. https://doi.org/10.1186/1750-0680-4-4

Jaffé, R., Nunes, S., Dos Santos, J. F., Gastauer, M., Giannini, T. C., Nascimento Jr, W., Sales, M., Souza, C. M., Souza-Filho, P. W., & Fletcher, R. J. (2021). Forecasting deforestation in the Brazilian Amazon to prioritize conservation efforts. *Environmental Research Letters*, *16*(8), 084034.

Köthke, M., Schröppel, B., & Elsasser, P. (2014). National REDD+ reference levels deduced from the global deforestation curve. *Forest Policy and Economics*, *43*, 18–28.

Mertz, O., Grogan, K., Pflugmacher, D., Lestrelin, G., Castella, J.-C., Vongvisouk, T., Hett, C., Fensholt, R., Sun, Z., Berry, N., & others. (2018). Uncertainty in establishing forest reference levels and predicting future forest-based carbon stocks for REDD+. *Journal of Land Use Science*, *13*(1–2), 1–15.

Müller, D., Sun, Z., Vongvisouk, T., Pflugmacher, D., Xu, J., & Mertz, O. (2014). Regime shifts limit the predictability of land-system change. *Global Environmental Change*, *28*, 75–83.

Ramankutty, N., & Coomes, O. T. (2016). Land-use regime shifts: An analytical framework and agenda for future land-use research. *Ecology and Society*, *21*(2).

West, A. P., T., Börner, J., Sills, E., O. and Kontoleon, A. (2020) Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *Proceedings of the National Academy of Sciences. 24188-24194. V. 117. 10.1073/pnas.2004334117*

Shoch, D., Eaton, J., and Settelmyer, S. (2013) Project Developer's Guidebook to VCS REDD Methodologies, Version 2.0. Conservation International.

Seyller, C.; Desbureaux, S.; Ongolo, S.; Karsenty, A.; Simonet, G.; Faure, J.; Brimont, L. (2016) The "virtual economy" of REDD+ projects: Does private certification of REDD+ projects ensure their environmental integrity? *Int. Rev.* **18**, 231–246.

Verra. (2020). *VM0007 REDD+ Methodology Framework (REDD+MF)*. https://verra.org/methodology/vm0007-redd-methodology-framework-redd-mf-v1-6/

Vongvisouk, T., Lestrelin, G., Castella, J.-C., Mertz, O., Broegaard, R. B., & Thongmanivong, S. (2016). REDD+ on hold: Lessons from an emerging institutional setup in Laos. *Asia Pacific Viewpoint*, *57*(3), 393–405.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Appendix 1. Predictive R2 and RMSE plots per municipality.

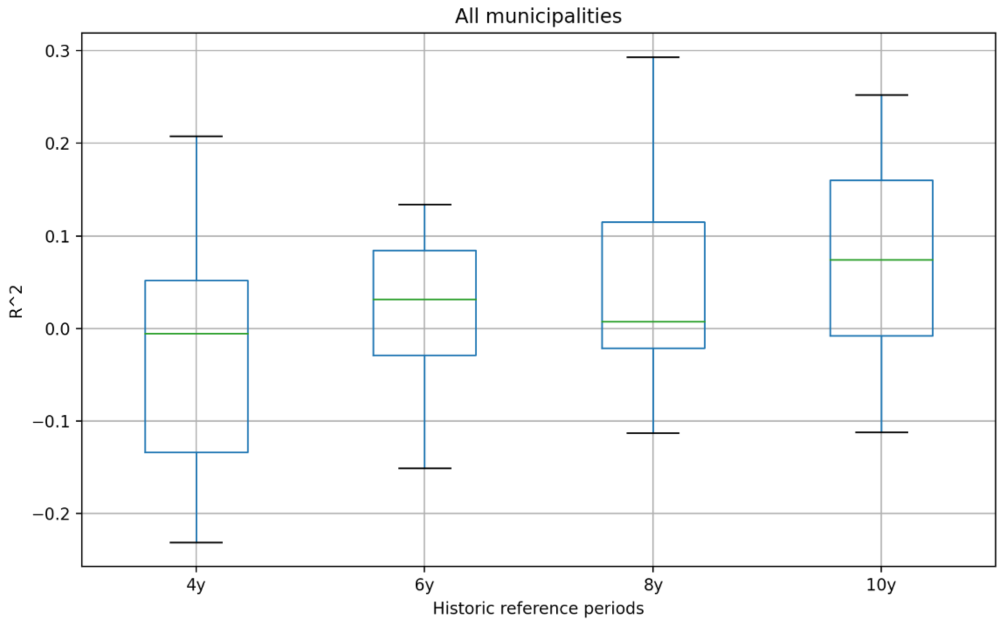


Figure 1. Native R2 of 80 linear regressions plotted on a single graph to compare model performance between different historical reference periods and scenarios.

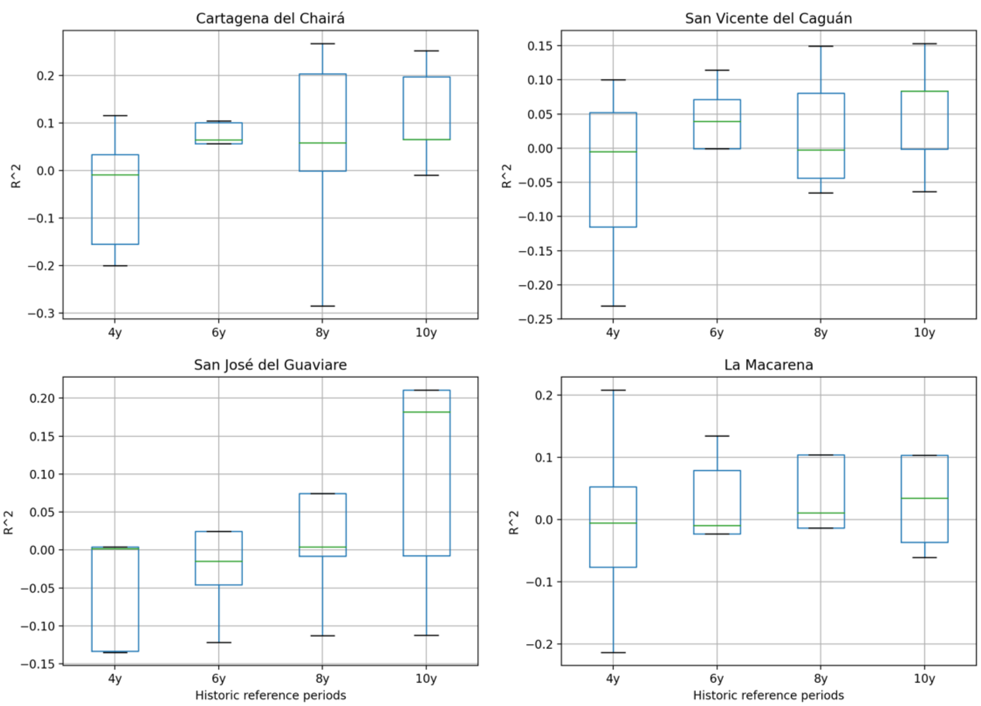


Figure 2. Native R2 per municipality of 20 linear regressions plotted on a single graph to compare model performance between different historical reference periods and scenarios.

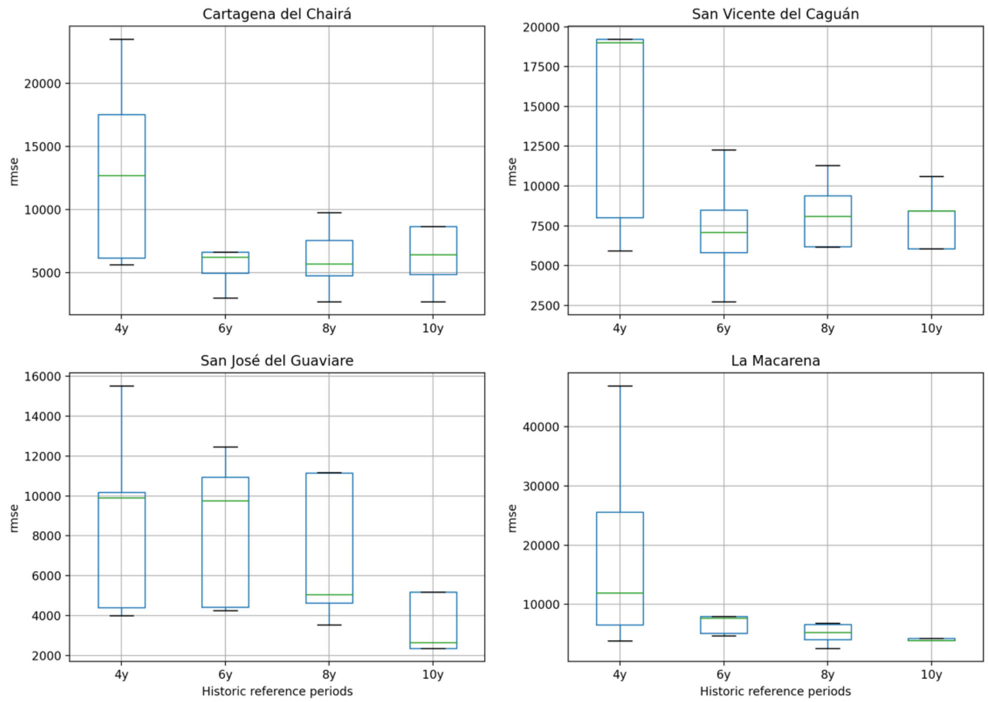


Figure 3. Native RMSE per municipality of 20 linear regressions plotted on a single graph to compare model performance between different historical reference periods and scenarios.