

Response to Reviewers: Statewide Forest Monitoring Data for California

Journal: Earth System Science Data

Authors: Christopher B. Anderson and David C. Marvin

Overview

The authors would like to thank the three referees for their fair and thoughtful feedback on our manuscript. These reviews shared important concerns about the temporal stability of the fuel predictions, the reliability of the estimates (regarding both geographic generalization and measurement sensitivity), the lack of independent validation, and raised questions about how to interpret the data. These are all key points we directly address in our revised manuscript. In this response, we first provide an overview of how we sought to incorporate the common feedback across referees into the manuscript, then we provide a line-based response to the specific points from each review.

We significantly expanded the manuscript by adding two new methods sections and results sections: a) *Geographic generalization and time series stability*, and b) *Intercomparisons with independent reference data*. These included multiple intercomparison analyses, four new figures, and a new table. We refined the discussion section based on referee feedback, and added a new discussion section titled *Lessons learned and open questions for future fuel monitoring systems*. We believe these new additions substantially improve the quality of the manuscript and address the key issues raised in the reviews.

One caveat that frames our response and updated manuscript is that the California Forest Observatory (CFO) was published six years ago. This manuscript documents data that have been accessed by thousands of users and cited in over 70 studies, but were never formally described. We do not plan to update the CFO models, and we considered further sensitivity analyses that would require re-fitting models out of scope. We sought instead to add analyses to better describe the published data and characterize their behavior in more depth.

Updated Framing and Positioning

We revised the language to describe the work, moving from an “operational” to a “prototype” monitoring system. This is now reflected in the article title, which better characterizes the quality and scope of the original deployment.

We learned many lessons from this prototype that directly informed current operational monitoring systems, such as the [Forest Carbon Diligence](#) and [Forest Carbon Monitoring](#) products. We believe it is more precise to position the CFO as a large-scale prototype, even if the outputs were adopted in several operational contexts (as described in the Impact section), as it will better ground expectations regarding the quality and interpretation of the data, particularly the quality of the understory vegetation metrics.

Field Observations

There are two complementary throughlines in the referee's comments related to field data. First regards the ecological interpretability of 3 meter scale canopy height and canopy cover estimates. Second is the request for further independent corroboration of the results.

On the first point, our perspective is that high resolution (meter and even sub-meter scale) mean canopy height and canopy cover retrievals from airborne LiDAR are well-established, canonical, and non-controversial. They represent a fundamentally distinct measurement from what one would collect in the field.

These metrics are derived from the height and density of discrete, small footprint LiDAR pulses. Gridded canopy height represents the average height of vegetation at each pixel location, typically estimated via [Delaunay triangulation](#). These metrics do not directly map onto field measurements where the individual tree is the unit of measurement. But, when averaged over plots, mean canopy height best correlates with the stand-level aggregate Lorey's Height. The Supplemental Material of [Anderson et al. 2026](#) has a detailed analysis comparing different LiDAR and stand-based field measurements, if you are interested in deeper review. Furthermore, the aggregation of mean canopy height is used by wildfire fuels products like LANDFIRE, which specifies the set of definitions our metrics adhere to.

This is all to say that, while it can be challenging to reconcile tree-oriented measurements with gridded height data, this is a general issue that is not specific to our paper, and is not strictly germane to this domain of canopy fuel modeling under current definitions. We tried to clarify these points briefly in the LiDAR processing section (Sec 2.2), which expands on previous text on the topic.

Independent Validation and Intercomparison

The referees highlighted the need for additional independent reference data to evaluate our results. We concur with their assessment. We have added several new analyses with multiple observation modalities to independently assess our data products, including intercomparisons with LANDFIRE, the de facto standard in the US fuel modeling community.

Open access vegetation measurements over the domain of interest — the State of California in our case — are available through the US Forest Inventory and Analysis (FIA) program and

from other remote sensing products. FIA publishes their database openly but intentionally randomizes plot geolocation information. This means FIA data cannot be used for plot-level validation. Fortunately, the NASA [Carbon Monitoring System](#) published a 64,000 ha scale hex grid with aggregated plot data. We have now added an analysis where we aggregate CFO and data to this scale and compare with the FIA aggregation (Sections 2.8 and 3.3).

In this same section we developed an intercomparison analysis with [GEDI RH98 data](#), which provides spaceborne LiDAR measurements of top of canopy height. These data represent a dense observation record of canopy height across the state that can be used for independent evaluation of both the CFO model predictions and the LANDFIRE benchmark. GEDI measurements provide coverage of all of California's ecoregions, which is particularly valuable for characterizing out of sample model performance, and can be used to fairly compare the CFO and LANDFIRE products with independent reference data. Overall, we compared nearly 900,000 GEDI footprints to evaluate correlations between GEDI, CFO and LANDFIRE.

Each of these analyses is best framed as an intercomparison, not a validation. Validation assumes one source of data is an error-free source of truth, and that the two metrics being compared are equivalent. In this case the FIA hex analysis represents a modeled quantity (forest cover) with associated uncertainty, and the GEDI measurements of RH98 are similar but not equal to ALS-derived mean canopy height. We tried to fairly characterize these factors in the manuscript, providing two large-scale, independent analyses of performance using high quality reference data.

Time Series Stability

The referees identified year-to-year prediction variation as something we discussed as a limitation but did not adequately describe from a technical perspective. To address this feedback, we included a new time series analysis (Sections 2.7 and 3.3) that characterizes time series stability and trends over time. This analysis does not provide independent temporal validation, nor does it quantify any temporally-correlated error structure in the predictions; it primarily provides insights into the magnitude of time series variance. We further elaborated on the challenges posed by time series instability in the Discussion (Section 4.2), and we now provide additional reflections on how to better characterize time series stability and temporal error structure in a new "lessons learned" Discussion section (Section 4.3).

Prediction Generalization

The referees identified that there is little information provided to evaluate out-of-sample test performance to evaluate geographic generalization. We added two analyses to address this.

First is an analysis quantifying the geographic bias of the training data, spatially mapping areas that are proportionally under or overrepresented in the training data extent. These

maps indicate how well the biogeographic conditions across the state are represented within the training data extent, providing a metric of expected generalization skill.

Second is the GEDI RH98 analysis described above. GEDI data are available statewide beyond the extent of our airborne LiDAR collection, providing an effective test of generalization performance. Since this is a fully independent measurement modality — reporting a separate metric of canopy height — we frame this as an intercomparison analysis and report adjusted R^2 scores, correlation coefficients, bias scores, and mean residual error rates as a way to holistically assess the quality of the model predictions against a dense independent dataset.

We also report GEDI intercomparison numbers for the CFO models inside our training data footprint (in sample) and across the rest of the state (out of sample). We find no difference in model performance inside and outside, even weighting by distance to the training data, indicating reasonable generalization performance.

Interpreting the Data

The referees also expressed concern about the reliability of understory vegetation predictions from overhead Earth observations, and about differences in the ranges of values for metrics like Canopy Bulk Density. We have included an extended discussion of each of these topics in the Discussion section (Section 4.3), as there are several practical considerations that these questions raise about the use of this data.

Overall, we thank the reviewers for their excellent feedback on the manuscript, and we now recognize their contributions in the manuscript (Acknowledgements). Next, we respond to the line item feedback from each referee.

Detailed Responses

Referee comments are formatted as left-justified text, and our replies as bullet points. Related comments were grouped together.

Anonymous Referee #1

L112 How is the lower canopy layer defined? How is ladder fuel density defined?

- These are defined in text just after the figure caption in question.

L115 why is Z max 120m?

- Clarified (Sec 2.2). It's the height of Hyperion, the tallest tree in the world (~120 m)

L122 Why is the maximum number of vertical canopy intervals set to four?

- It is not, and we cannot identify what this comment refers to (such a limitation is not stated in preprint L122). The scatter plot of layer count predictions appears to saturate ~4 vertical layers, but the lidar-derived layer count metrics extend further (Fig. 5).

L127-128 What is the reference for this method? (canopy layer count derivation)

- There is no independent reference - we developed this methodology in this paper.

L91 How do the authors know that the airborne lidar data have sufficient canopy penetration to reliably characterize understory structure?

- We understand the cautionary interpretation. For reference, the authors worked for over 5 years collecting airborne LiDAR data in tropical and temperate forests around the world and have published several articles characterizing vertical structural complexity with airborne LiDAR data.
- It is indeed the case that collection parameters like spot spacing and laser power play important roles in determining the density of understory observations. This is difficult to standardize across opportunistic USGS data collections, which were sourced from a series of instrument configurations.
- The methodologies we developed here were designed to normalize values as much as possible across collections. This includes the 10x10m binning for vertical structure layers, meaning that, at an average point spacing of 1 shot per m^2 , and an average return density of two LiDAR returns per shot (conservative estimates, with most California collections collecting >1 shot per m^2), we would have an average density of 200 point values per pixel from which to voxelize our vertical structure estimates. The canopy height, canopy cover, and ladder fuel metrics are all designed to minimize sensitivity to point density and report consistent metrics across configurations.

L191-192 The linkage between fuel-type classification and fuel-load quantification is somewhat confusing... The basis for these relationships needs to be more clearly explained and justified.

L184-192 The referenced classification scheme for surface fuels does not appear to provide a sufficient basis for classifying non-surface fuel types. The authors claim to have developed a hierarchical fuel classification scheme that represents the full vertical fuel structure, but some of the criteria appear loosely or inconsistently defined. For example, why is green vegetation cover considered a good proxy for grass density, and why is ladder fuel density considered a reliable proxy for shrub and understory density? In particular, ladder fuels generally refer to combustible material that provides vertical continuity between surface fuels and the canopy and should not simply be treated as equivalent to shrubs or understory vegetation.

- Great suggestion. We have added some citations to back up these claims. First are some citations from the literature on spectral mixture analysis demonstrating linear

relationships between vegetation cover fractions and grass biomass density. Second is the paper by Kramer et al. (2014) who indeed define ladder fuels in a structural fashion (the density of vegetation from 1-4 m tall), which provides the basis for our method.

L212-218 Although the reference ALS data may have some ability to characterize subcanopy fuel conditions, none of the satellite-based predictors used here have been demonstrated to be sensitive to subcanopy vegetation or fuel conditions, such as surface fuels or shrubs. What evidence supports the validity of using these satellite observations to infer these subcanopy fuel metrics?

- Thanks for making this point. The ladder fuel metric can be consistently mapped for understory or shrubland vegetation from airborne LiDAR, but the retrieval quality from satellite data is likely worse in closed forests than in open shrublands. We have clarified and expanded this in the Discussion (Sec. 4.3).

L274- 279 what is the ecological base for a 3m resolution canopy height and canopy cover product when medium tree crown size exceed 9m and large mature tree crowns easily exceed 15m.

- Added some text clarification and a reference to an intercomparison analysis demonstrating the relationships between LiDAR and field-derived metrics (Sec 2.2).

L284 with regards to the time domain, it was not very clear how the models handle the training data from various years, and whether they developed annual models or single model.

- While the exact breakdown of train/test performance by year is not available to us, the new section on time series stability should address this point (Sections 2.7 and 3.3)

L309-310 this is informative, however, despite the known limitation with pixel-level accuracy, the model is still applied to produce pixel level maps?

- Added a discussion point on this topic (Sec. 4.3)

Figure 5 & 6, given California vegetation and fuel structure is highly diverse, the authors didn't touch upon how they evaluate the models for out-of-training domain performance? The figure 1 shows large portion of the southern and central region lack training data.

- Good call. This is now addressed with the new sections on geographic bias and an evaluation of out of sample performance using statewide GEDI data.

Section 4&5, both the discussion and conclusion describe great goals and needs of such data, but without reliable independent references and assessment of ability to apply model cross domain and the uncertainties, the level of saturation and mis-estimation could undermine and mislead lots of the goals than the product could contribute.

- Indeed. The new Discussion section (Sec. 4.3) addresses these issues directly. The lack of reliable reference data is of far larger importance than signal saturation, in our estimation, but the point is well taken.

Thank you for your time and consideration.

Anonymous Referee #2

Spatial independence of model evaluation. The 64×64 pixel tiles were randomly divided into 80/10/10 training, validation, and testing sets. Could neighboring tiles or tiles from the same LiDAR acquisition occur in both the training and testing sets? If so, spatial autocorrelation could lead to optimistic performance estimates.

- Thanks for pointing this out. Clarified in sections 3.2 and 3.3.

Canopy-height saturation. Figure 7 clearly shows canopy-height saturation, particularly above about 30 m. Were any correction or post-processing approaches applied, or are the saturated predictions retained in the final products? Please discuss whether this bias may affect derived variables that depend on canopy height.

- Added a sentence on this topic in the discussion section where this effect is compared among data products (Sec. 4.2)

Zero-value patterns in Figs. 5 and 6. The canopy-cover 2-D histograms show many observations with substantial canopy cover but predicted values near zero. This pattern is especially clear for the 3 m model. Could the authors explain the source of this pattern?

- Good catch. Added a sentence clarifying this effect (Sec. 3.1).

Generation of 3 m and 10 m response variables. Section 2.2 states that canopy height and canopy cover were first rasterized at 1 m resolution, while the models operate at 3 m and 10 m resolutions. Please clarify how the 1 m LiDAR response variables were aggregated or resampled to these resolutions.

- Clarified in the LiDAR processing section (Sec. 2.2)

Figure 8 and canopy bulk density. Figure 8A shows substantially lower canopy bulk density estimates from the Forest Observatory than from LANDFIRE. The manuscript explains why the distributions have different shapes, but it does not fully explain the difference in magnitude. Please discuss the causes and effects.

- Clarified these patterns in the results section (Sec. 3.4) and elaborated on the effects in the discussion (Sec. 4.3)

Uncertainty in derived fuel products. Canopy bulk density combines several modeled or externally derived quantities, while the surface fuel product is generated using a rules-based classification. Please provide more discussion of uncertainty in these derived products and how errors in the inputs may propagate.

- This is a great point. While not directly addressed, we discuss the general issue of uncertainty and framed the core question — how to quantify and propagate uncertainty — as a frontier in Section 4.3.

Thank you for your time and consideration.

Anonymous Referee #3

Surface fuels max height is not defined anywhere

- No it is not, and the authors cannot determine what the referee is referring to here.

135: why did you go with > 3% change?

- This is described in the next line (Sec 2.3).

147: canopy bulk density units not defined until line 163

- The units are defined in the following sentence (Sec. 2.3)

153: LC

- LC is a variable in Eq. 4, and the in-text reference is describing the equations' variables.

241: U-nets not until 263

- The citation of Ronneberger et al. (2015) is used to introduce U-nets.

278: how can you jump from a lower to higher res. model by concatenating predictions?

- Added more contextual information to clarify the resampling. This introduces an implicit propagation of features, akin to using a 1-band, lower resolution foundation model as a predictive feature. This is now clarified in Section 2.5.

Why did you choose to compare with LANDFIRE? (figure 8)

- Justification for the LANDFIRE intercomparison is articulated in Sec. 2.6

Thank you for your time and consideration.