

# High-resolution soybean tracing for deforestation-free supply chains

Corresponding Author: Dr Caspar Chater

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Communications Earth & Environment.

**This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.**

**Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.**

Version 0:

Decision Letter:

**\*\* Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors \*\***

Dear Dr Chater,

Your manuscript titled "High-resolution soybean tracing for deforestation-free supply chains" has now been seen by 3 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised substantial concerns that must be addressed. In light of these comments, extensive revisions will be required before we can further consider the manuscript for publication. We would, however, be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the reviewers' comments useful as you decide how to proceed. If additional work allows you to either incorporate or refute these criticisms, we will be happy to look at a substantially revised manuscript. If you choose to take up this option, please either highlight all changes in the manuscript text file, or provide a list of the changes to the manuscript with your responses to the reviewers.

To be publishable in Communications Earth and Environment the following editorial thresholds need to be addressed.

\*Compellingly validate the model's performance focusing on the generalizability.

\*Clearly discuss the sampling strategies.

\*Transparently report the associated operational limitations/challenges.

**When resubmitting, please provide a point-by-point response to the reviewers' comments.** Please submit your responses as a separate file, distinct from your cover letter where you can add responses to the Editors' comments that you do not want to be made available to the reviewers. Word files are preferred. We recommend that any figures, tables or graphs that are included in the response to reviewers are also included in the main article or Supplementary Information.

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

If the revision process takes significantly longer than three months, we will be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the reviewers' comments with a list of your changes to the manuscript text (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and any completed checklist:

Link Redacted

\*\* This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first \*\*

Please do not hesitate to contact us if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Nandita Basu, PhD  
Consulting Editor, Communications Earth & Environment  
Associate Editor, Communications Sustainability  
Nature Portfolio

## EDITORIAL POLICIES AND FORMAT

If you decide to resubmit your paper, please ensure that your manuscript complies with our editorial policies and complete and upload the checklist below as a Related Manuscript file type with the revised article:

For Manuscripts that fall into the following fields:

- Behavioural and social science
- Ecological, evolutionary & environmental sciences
- Life sciences

An updated and completed version of our Reporting Summary must be uploaded with the revised manuscript

You can download the form here:

<https://www.nature.com/documents/nr-reporting-summary.pdf>

For your information, you can find some guidance regarding format requirements summarized on the following checklist: (<https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf>) and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>).

## REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

This manuscript uses data on the locations of sampled soybean farms in South America, as well as attributes of soybeans harvested from those locations, in a predictive model meant to geolocate soy based on its attributes. The problem is timely given the upcoming implementation of the EUDR at the end of 2025 as well as more general concern on supply chain traceability and transparency for commodities. The paper seems to provide an advancement for how to geolocate annual crops. It is generally well written, but could use some clarification and greater detail at points. I also think that the authors have buried an interesting finding in the supplement, and I recommend they move it to the main text (in shortened form, of course). I have some general and specific thoughts that I hope the authors can incorporate during revision, as outlined below.

## GENERAL COMMENTS

**Sample Selection.** The authors do not describe how the locations for samples were selected. This is a critically important part of the methodology that needs to be clearly explained. Was sampling random? Stratified? Convenience? Why did they collect the number of samples that they did? Are there any potential biases in the sample, such as under sampling of recently deforested land, or oversampling of farms held by young people, and are those biases potentially correlated to the variables included in the model (including space and soybean characteristics)? In what way might the sample bias the model?

**Supplement/Climate and Soil Data.** Some of the best parts of the manuscript – i.e., the attempt to use climate and soil data to improve the model – are relegated to the supplement. This (i.e., Section 3 and Section 4 in the supplement) is critical to include in the main text, because as the authors say it has the potential to improve predictive capabilities beyond areas sampled. The finding that soil and climate data do not improve model performance is a fascinating puzzle. Why do the authors think this might be (are soybean characteristics more variable across time than space? Do soy properties vary so predictably across space that x,y coordinates are actually the only variables needed for a good prediction)? I also wonder why the authors didn't incorporate elevation and slope data, which are freely available, sufficiently high resolution for use in these models, and less "modeled" to fill gaps than the soil, precipitation, or temperature data. Or, for that matter, other relevant variables like distance to coast which the authors mention might influence the properties of the soybeans. Why not provide the model with many possible geospatial predictors – just as they did with the soybean sample chemistry data – to

see which set of predictors provide the best performance?

Acronyms. I know the authors are trying to reduce words, but where possible please spell out the concept instead of using acronyms (SIR, TE, etc) which become difficult to track.

Data sensitivity. The authors imply that the data are of a sensitive nature. But soybean maps exist for all of South America, with especially good availability in Brazil. It is well known where soybean farms are. Is there another reason for not releasing the spatial coordinates of the study? Was this part of an agreement with farmers, for instance?

## SPECIFIC COMMENTS

### Main Text

Line 25. Which predictions? This is not clear until reading the whole paper and should be explained in the abstract.

Line 101. There does seem to be some research in this area on oil palm: <https://jopr.mpob.gov.my/a-proof-of-concept-study-determining-the-geographical-origin-of-crude-palm-oil-with-the-combined-use-of-gc-ims-fingerprinting-and-chemometrics/>

Line 177. "Continental scale" seems like an exaggeration, since the sample did not span the whole continent.

Line 191. Did the authors try running the model with just these limited predictors? How did performance change? What was the per-sample cost of analysis for this project?

Line 197. Country of origin seems important for certain types of traceability and transparency, but the authors in the introduction imply that the issue is identifying soybean sourced from recently deforested lands. How is country of origin relevant to this issue? Or are there concerns about, for instance, soybeans from Paraguay being laundered into Brazil supply chains? It would be helpful if the authors could set up this story a bit more effectively.

Line 197-201. Has any work been done on this temporal stability? If not, maybe that should be a recommendation for future work.

Line 202. But could part of the variability be due to what varieties are used in which locations? Might this be additional information (or confusion) for a less industrialized crop?

Line 228. "unprecedented" seems like rather strong language.

Line 229-231. On deforestation frontiers, it requires accuracy on the <1 km level to definitively confirm deforestation rather than just risk (or alignment with statements about origin). Interestingly, there is also the problem of the accuracy and scale of the deforestation maps. It might be worth mentioning that combining these probability-of-soy-origin maps with deforestation maps means dealing with the uncertainties in both maps.

Lines 237-240. Could the authors clarify? I am not sure what this sentence (starting with "Verification only") is trying to communicate.

Line 243. Could the authors provide some examples of local sources of provenance info?

Line 273-275. Why?

Line 276-277 ("we aim...problems"). This is not clear. Please rephrase and clarify.

Section 5.2. How were sampling locations selected? Please see general comment above.

Line 285-286. Does this dry collection mimic the way soybeans are processed? Overall, if this approach used in the current manuscript was applied by regulators/companies to actually verify origins, would it be effective? At what point in the supply chain are beans likely to be tested (e.g., upon entrance to the EU at a port) and would the methodology the authors applied here still be valid in that case?

Line 287 ("at the ... 10 km diagonal"). Why? is this the scale at which prediction is needed for policy purposes, or the scale at which variation in bean properties is likely to be minimal?

Line 343. <30 km – why? Was there testing at greater and smaller cutoffs?

Line 355. What was the exact dataset used? Is this potential croplands or current croplands?

Line 358-361. This is confusing, please clarify: What suggests a greater range? Which previous estimates (there is only one source here). What makes the prior less informative?

Figure 1. As well as deforestation, I recommend that the authors provide a map of soybean areas. This is available for Brazil annually from MapBiomass. I believe there is at least one dataset for all of South America available:

<https://data.globalforestwatch.org/datasets/soy-planted-area-/about>

Figure 4. What is the difference between A and D if both are based on the same data? Is this a comparison of ML and traditional classification? Please edit the caption for clarity.

### Supplementary File 1

Section 3/ Paragraph 1/ "that are hypothesized" – based on what information? I see this is covered in the discussion below but laying out the logic in the methods would be more effective.

### Reviewer #2 (Remarks to the Author):

The manuscript focuses on a relevant topic, which is the traceability of agricultural commodities. The claim of the paper is to establish a method to determine the provenance of the soybean, examining stable isotope ratios and trade elements of soybean samples. Gaussian Process models were applied to, as the final outcome, estimate for "every pixel the probability that overlaps with the harvest location of the sample". The provenance is, therefore, estimated. According to the results presented, the efficacy of the model is emphasized as very satisfactory based on the calculated prediction error and 95% confidence region (CR).

The results obtained and the methodology per se is of the interest because traceability is a major issue for regulators and businesses communities that deal with agricultural commodities trading. Especially for soybean, which is a fungible good, such methodology has potential to be helpful for buyers and exporters that must prove the provenance for regulatory and

consumer needs purposes.

The manuscript anchors the sample collection in Latin America countries alleging that soy expansion associated to deforestation is found on those regions. However, the need to prove provenance is not a requirement only for regions in which deforestation is present. The EUDR, legislation mentioned in the manuscript several times to justify the paper, requires proof of origin (plot of land) from all countries supplying the EU with commodities covered by the legislation. The manuscript, therefore, should have collected samples from all major soybean producers, and not only South American countries. I suggest justify why samples from US were not collected, being the country a major soy producer and subjected to the same requirements as any other country in the scope of the EUDR.

It is necessary to discuss the applicability of the model for real operations. Given that initiatives such as the due diligence of the EUDR and the Soy Moratorium are mentioned by the manuscript, I am assuming that the authors expect the methodology be recognized to be used in large-scale operations such as soybean exports. However, the manuscript is silent about this topic. The authors should have pointed out the limitations of the method to be adopted in large-scale operations. I would like to present the limitations I identified.

The first limitation is that soybeans batches loaded on vessels for exports come from several origins. So collecting samples in vessels, or in batches, will not represent all origins from where the soy where harvested. Soybean is a fungible good passing through several loadings from the farm to the exporting port. The origin control must be done at the points of aggregation, such as warehouses and not at the final point, which is the port.

The second limitation is that collecting samples at the point of aggregation is complex and would need to be integrated with the samples collection already in place for other purposes such as soy quality and moisture classification. However, any warehouse receives soy from hundreds of trucks every day. The very complex process for controlling the origin of each truck at the point of aggregation is the main reason for the Soy Moratorium to identify the provenance using satellite images, which is more applicable for large scale operations. The manuscript quoted the Soy Moratorium, which was appropriated, but it should have made a benchmark of its method with the Soy Moratorium method from the point of view of the applicability for large scale operations. Soy Moratorium methods are well described in the annual reports of the initiative and available at: <https://abiove.org.br/esg/iniciativas/soy-moratorium/>. The second limitation is, therefore, the absence of a benchmark between the method proposed by the manuscript and the Soy Moratorium method.

In my opinion, recognizing the limitations of the proposed methodology for large scale operations and justifying why only Latin America was part of the samples, the manuscript can be published.

Reviewer #3 (Remarks to the Author):

Please see my attached review report

\*\* Visit Nature Portfolio's author and referees' website at <a

href="http://www.nature.com/authors">www.nature.com/authors</a> for information about policies, services and author benefits\*\*

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the <a href="http://platformsupport.nature.com/">Platform Support Helpdesk</a>.

Version 1:

Decision Letter:

<\*\*\* REMEMBER TO ATTACH REVISIONS CHECKLIST (WORD)\*\*\*>

\*\* Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors \*\*

Dear Dr Chater,

Your manuscript titled "High-resolution soybean tracing for deforestation-free supply chains" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment provided you address all the remaining concerns of reviewer 1, particularly regarding the detailed sampling strategies.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

#### EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

\*\*\*\*Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. \*\*\*\*

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

#### SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at <https://www.nature.com/documents/commsj-file-checklist.pdf>.

#### OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Portfolio, please visit <https://www.nature.com/commsenv/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:

Link Redacted

\*\* This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first \*\*

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Nandita Basu, PhD  
Consulting Editor, Communications Earth & Environment  
Associate Editor, Communications Sustainability  
Nature Portfolio

#### REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

see attached

Reviewer #3 (Remarks to the Author):

The authors have done a competent job of addressing my comments. I recommend acceptance of the revised version of the paper.

\*\* Visit Nature Portfolio's author and referees' website at <a href="http://www.nature.com/authors">www.nature.com/authors</a> for information about policies, services and author benefits\*\*

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

## **High-resolution soybean tracing for deforestation-free supply chains**

Responses to reviewers' comments

Editor's comments:

- \*Compellingly validate the model's performance focusing on the generalizability.
- \*Clearly discuss the sampling strategies.
- \*Transparently report the associated operational limitations/challenges.

### **Reviewer #1 (Remarks to the Author):**

This manuscript uses data on the locations of sampled soybean farms in South America, as well as attributes of soybeans harvested from those locations, in a predictive model meant to geolocate soy based on its attributes. The problem is timely given the upcoming implementation of the EUDR at the end of 2025 as well as more general concern on supply chain traceability and transparency for commodities. The paper seems to provide an advancement for how to geolocate annual crops. It is generally well written, but could use some clarification and greater detail at points. I also think that the authors have buried an interesting finding in the supplement, and I recommend they move it to the main text (in shortened form, of course). I have some general and specific thoughts that I hope the authors can incorporate during revision, as outlined below.

### **GENERAL COMMENTS**

**R1C1 (and R1C17, R1C19):** Sample Selection. The authors do not describe how the locations for samples were selected. This is a critically important part of the methodology that needs to be clearly explained. Was sampling random? Stratified? Convenience? Why did they collect the number of samples that they did? Are there any potential biases in the sample, such as under sampling of recently deforested land, or oversampling of farms held by young people, and are those biases potentially correlated to the variables included in the model (including space and soybean characteristics)? In what way might the sample bias the model?**Response:** Sampling locations were selected as a result of optimising multiple factors, including but not limited to maximising spatial coverage, keeping within budget for both field expeditions and downstream sample analyses, obtaining farmers' agreement to participate, avoiding areas that suffered damage to crops (e.g. through disease and droughts), and designing a sampling route that coincides with the harvest period in the various regions so that seeds were collected at the correct developmental stage. Prior location selection sometimes had to be adapted by collectors on the ground due to changes in seasonal crop planting choice as well as local security risks. In each location we aimed to take 5 samples to account for local and regional variation in the data. Sampling was arranged in a square with a 10 km diagonal around a central sampling point in each location.

This sampling strategy was optimised for several objectives: capturing natural variation, accessibility to collectors, reducing the number of farms sampled, keeping to sample quotas.

Our checks did not locate any consistent bias in the data. Demographic data was not collected so we could not test for correlations with farmer characteristics. [lines 377-382, 388-399]

**R1C2:** Supplement/Climate and Soil Data. Some of the best parts of the manuscript – i.e., the attempt to use climate and soil data to improve the model – are relegated to the supplement. This (i.e., Section 3 and Section 4 in the supplement) is critical to include in the main text, because as the authors say it has the potential to improve predictive capabilities beyond areas sampled. The finding that soil and climate data do not improve model performance is a fascinating puzzle. Why do the authors think this might be (are soybean characteristics more variable across time than space? Do soy properties vary so predictably across space that x,y coordinates are actually the only variables needed for a good prediction)? I also wonder why the authors didn't incorporate elevation and slope data, which are freely available, sufficiently high resolution for use in these models, and less “modeled” to fill gaps than the soil, precipitation, or temperature data. Or, for that matter, other relevant variables like distance to coast which the authors mention might influence the properties of the soybeans. Why not provide the model with many possible geospatial predictors – just as they did with the soybean sample chemistry data – to see which set of predictors provide the best performance?

**Response:** We thank Reviewer #1 for their observations and comments. We had to decide carefully what to include in the main text due to article length and figure restrictions. We also tried to ensure that the key steps described in the main text were relevant to how the model works and therefore relegated less immediately applicable approaches to supplemental. Our thinking is that the soil and climate data layers currently offer less for model improvement as a good understanding of how they modulate seed composition is lacking. Closing this gap will likely take time, so developing and improving on the framework we describe in the main text is, in our opinion, the most promising route forward. Similarly, including slope and elevation data (distance to coast requires first identifying the relevant coast) added much more data and complexity to our models, whereas the soy suitability data that we used integrates over the effects of variables directly relevant to soybean growth. We add that in a previous study on oak trees from the US, the inclusion of climate variables also did not improve the accuracy or precision of origin predictions. [lines 263-264]

All in all, we think that the models relying on environmental data, although interesting, do not advance the main theme of this study— a working model for origin determination—because they offer poor predictions in their current state. To avoid exceeding article length limits we have added a paragraph summarising this topic in the discussion [lines 252-270], but would like to keep the full section (~430 words, 2 display items +captions) as supplementary material. Should the editor deem it crucial to include in full in the discussion, we think it would best fit after the discussion of the advantages of spatially explicit models in subsection “Outperforming current approaches”.

**R1C3:** Acronyms. I know the authors are trying to reduce words, but where possible please



spell out the concept instead of using acronyms (SIR, TE, etc) which become difficult to track.

**Response:** Thank you for this observation. We have now reduced acronyms to a minimum and now spell out the concepts at several places throughout the manuscript. In sentences and paragraphs where the acronyms are repeated multiple times, we kept them in place (particularly in the Results section and when referring to the predictive models by name) because avoiding the acronyms would make the text cumbersome, ambiguous, or otherwise more difficult to follow.

**R1C4:** Data sensitivity. The authors imply that the data are of a sensitive nature. But soybean maps exist for all of South America, with especially good availability in Brazil. It is well known where soybean farms are. Is there another reason for not releasing the spatial coordinates of the study? Was this part of an agreement with farmers, for instance?

**Response:** the data is sensitive because the use of exact farm location reveals the farmer's and therefore donor's identity. Protecting the farmer's identity aligns with data protection legislation in several relevant countries. Moreover, some farmers have explicitly conditioned their participation on remaining anonymous. EUDR and soybean traceability are highly contentious subjects in Brazil and in other areas of South America, and collector and donor security must not be compromised.

## SPECIFIC COMMENTS

### Main Text

**R1C5:** Line 25. Which predictions? This is not clear until reading the whole paper and should be explained in the abstract. **Response:** We have now spelled it out in the text. [line 25]

**R1C6:** Line 101. There does seem to be some research in this area on oil palm: <https://jopr.mpob.gov.my/a-proof-of-concept-study-determining-the-geographical-origin-of-crude-palm-oil-with-the-combined-use-of-gc-ims-fingerprinting-and-chemometrics/>

**Response:** Unlike our study, the oil palm study in the link only applies classification models and is therefore non-spatial, like the soybean studies already cited in this sentence. There are numerous classification-based commodity traceability studies out there, we opted to include the most relevant to our study.

**R1C7:** Line 177. "Continental scale" seems like an exaggeration, since the sample did not span the whole continent.

**Response:** We respectfully disagree. This study merits the designation "continental scale" as it covers all of South America north of 40°S (>90% of the continent), it includes data from countries that make up ~70% of the continental land area, and it examines broad patterns and processes that operate at this scale.

**R1C8 (see also R3C1):** Line 191. Did the authors try running the model with just these limited predictors? How did performance change? What was the per-sample cost of analysis

for this project?

**Response:** The set of 22 predictors reported here is the limited set. We have run multiple models with various permutations of the set of predictors. The set reported here was the best performing in terms of prediction accuracy and precision. However, since this study aims to describe a working traceability tool, backed by intensive sampling, chemical analysis and machine learning, it does not focus on a full exploratory analysis and optimisation of the multi-dimensional predictor landscape. This will be the topic of a separate study. [lines 201-205]

**R1C9 (also R3C2):** Line 197. Country of origin seems important for certain types of traceability and transparency, but the authors in the introduction imply that the issue is identifying soybean sourced from recently deforested lands. How is country of origin relevant to this issue? Or are there concerns about, for instance, soybeans from Paraguay being laundered into Brazil supply chains? It would be helpful if the authors could set up this story a bit more effectively.

**Response:** Traceability has so far focused on identifying country of origin in the absence of (1) more precise measures and (2) legislation necessitating greater granularity (because to date, regulations were operating at the country level as origin declaration requirements). Here we aim to move past artificial country borders, and aim to identify the harvest location on a spatial scale independent of country boundaries, in line with EUDR. However, to compare our study with previous studies on agricultural commodities, which only used classification approaches, we included a quasi-classification experiment. We mainly include this to demonstrate the inherent weakness of classification models when applied to a contiguous geographical space. [we have added this in the discussion, lines 115-119, 209-218]

**R1C10:** Line 197-201. Has any work been done on this temporal stability? If not, maybe that should be a recommendation for future work.

**Response:** We are not aware of any such work in soybeans. We now refer to a case study on Beech trees and have added this recommendation in lines 223-227. Temporal variability in annual vs. perennial plants is also discussed in the Supp. Materials (Section 4).

**R1C11:** Line 202. But could part of the variability be due to what varieties are used in which locations? Might this be additional information (or confusion) for a less industrialized crop?

**Response:** This could certainly be part of it as different varieties may differentially take up soil elements due to divergent root traits, for example. We are currently preparing another manuscript which examines soybean genomes across South American varieties and the use of their genetic information to predict harvest origin and support traceability efforts. Based on this, a less industrialised crop would likely have a greater pool of agrobiodiversity which may make them easier rather than more difficult to trace. [lines 229-231]

**R1C12:** Line 228. “unprecedented” seems like rather strong language.

**Response:** We have now changed the wording of this sentence. [lines 277-278]

**R1C13:** Line 229-231. On deforestation frontiers, it requires accuracy on the <1 km level to definitively confirm deforestation rather than just risk (or alignment with statements about origin). Interestingly, there is also the problem of the accuracy and scale of the deforestation maps. It might be worth mentioning that combining these probability-of-soy-origin maps with deforestation maps means dealing with the uncertainties in both maps.

**Response:** Indeed, and because of that uncertainty, the legislation does not rely solely on exact origin identification. However, if a sample is traced to a region where deforestation is known to be ongoing, it is expected that it would trigger closer scrutiny and possibly further action. [we now refer to this point in lines 279-290]

**R1C14:** Lines 237-240. Could the authors clarify? I am not sure what this sentence (starting with “Verification only”) is trying to communicate.

**Response:** We have removed this sentence because it was not directly contributing to the point made in the paragraph and is detailed in other studies.

**R1C15:** Line 243. Could the authors provide some examples of local sources of provenance info?

**Response:** The Soy Working Group holds a register of soy-growing farms and combines it with satellite imagery, field visits and production data to assess violations of the Amazon Soy Moratorium. [lines 296-297]

**R1C16:** Line 273-275. Why?

**Response:** It is unclear what this question refers to. If it means why we chose this number of pixels, that was mainly a technical decision. We attempted to achieve modelling at the highest spatial resolution that our computational resources could support. If it is about retaining the original latitude:longitude proportions of the study area, that was done to maintain the integrity of distance and area measurements. [added in lines 362-365]

**R1C17:** Line 276-277 (“we aim...problems”). This is not clear. Please rephrase and clarify. Section 5.2. How were sampling locations selected? Please see general comment above.

**Response:** Defining a very small study area upfront necessarily directs origin predictions to that area, meaning that the user informs the model about the origin of the sample, which is unlikely to happen in real-world use cases. We have clarified this in the text [lines 367-369].

**\*Please note that this comment accidentally combines two separate comments. The above response addresses the first of the two (the spatial definition of the study), whereas the second comment (about sampling locations) is addressed in our response to R1C1.**

**R1C18:** Line 285-286. Does this dry collection mimic the way soybeans are processed? Overall, if this approach used in the current manuscript was applied by regulators/companies to actually verify origins, would it be effective? At what point in the supply chain are beans likely to be tested (e.g., upon entrance to the EU at a port) and would the methodology the authors applied here still be valid in that case?

**Response:** Dry collection mimics pod harvesting in the field. However soybeans are then

handled in several different ways to produce many downstream products. Soybean imports into the EU and UK vary annually, but just under half of the soybeans imported in the EU and a third of the soybeans imported into the UK arrive as unprocessed whole beans. Once the EUDR and the UK Deforestation Regulation come into effect, we will have a better understanding of the supply chain tests that will take place. Our methodology is effective in the case of whole beans. Downstream processing is likely to result in obfuscated signatures of origin for which further research is required to develop additional traceability tools. As mentioned above, DNA could well be an important alternative for processed commodities because genetic modification licensing varies by country, and cultivars are bred for specific environments, disease resistance, and photoperiods. [lines 383-385 and 310-323]

**R1C19:** Line 287 (“at the ... 10 km diagonal”). Why? is this the scale at which prediction is needed for policy purposes, or the scale at which variation in bean properties is likely to be minimal?

**Response:** To build a spatially effective model, we need to capture both the local and larger spatial variability. The system we use (5 points per location) provides the model with data about local variation, whilst facilitating consistent sampling instructions to collectors, as we work with local collection teams in many different countries. Specifically, 10km was chosen through optimising for several objectives: capturing greater variation (farther is better), accessibility to collectors (nearer is better), reducing the number of farms sampled (nearer is better in case of small farms), keeping to sample quotas (farther is better to cover more ground). [Lines 388-391; Please also note our response to R1C1].

**R1C20:** Line 343. <30 km – why? Was there testing at greater and smaller cutoffs?

**Response:** We explored greater and smaller cut-offs as well, but the value of 30 km resulted in most clusters having exactly 5 samples, adequately reflecting the survey design. Retaining the survey design is important. Too small a cutoff would split spatial clusters and enable the samples in them to be in both the training and testing datasets, thereby causing data leakage that wrongly inflates the model’s predictive ability. Too large a cutoff reduces the model’s predictive power as multiple spatial clusters are grouped into one, which would mask their respective characteristic values and the spatial signal that the model aims to identify. [lines 454-460]

**R1C21:** Line 355. What was the exact dataset used? Is this potential croplands or current croplands?

**Response:** We used a dataset of (potential) attainable soybean yields throughout South America, which does not depend on current land use. [cited in line 472, Reference #65]

**R1C22:** Line 358-361. This is confusing, please clarify: What suggests a greater range? Which previous estimates (there is only one source here). What makes the prior less informative?

**Response:** We have now clarified the sentence [lines 477-480]

**R1C23:** Figure 1. As well as deforestation, I recommend that the authors provide a map of

soybean areas. This is available for Brazil annually from MapBiomass. I believe there is at least one dataset for all of South America

available: <https://data.globalforestwatch.org/datasets/soy-planted-area-/about>

**Response:** We accept this request for context, but we think that overlaying soybean farms and deforestation maps as well as our sampling locations would make the figure difficult to interpret. Instead, to facilitate a clearer and more accurate understanding of soybean-related deforestation, Figure 1 caption now refers to a map—also by Global Forest Watch—of forest lost to soybean farms: <https://gfr.wri.org/forest-extent-indicators/deforestation-agriculture#how-much-forest-has-been-replaced-by-soy> .

**R1C24:** Figure 4. What is the difference between A and D if both are based on the same data? Is this a comparison of ML and traditional classification? Please edit the caption for clarity.

**Response:** In panels a-c the pixels indicate *true* harvest origin (panels a, b, c) whereas in d-f pixels show *predicted* harvest origin. We have clarified this in the figure caption.

**R1C25:** Supplementary File 1

Section 3/ Paragraph 1/ “that are hypothesized” – based on what information? I see this is covered in the discussion below but laying out the logic in the methods would be more effective.

**Response:** We have expanded on the sources of the hypotheses in the same paragraph.

## **Reviewer #2 (Remarks to the Author):**

The manuscript focuses on a relevant topic, which is the traceability of agricultural commodities. The claim of the paper is to establish a method to determine the provenance of the soybean, examining stable isotope ratios and trace elements of soybean samples. Gaussian Process models were applied to, as the final outcome, estimate for “every pixel the probability that overlaps with the harvest location of the sample”. The provenance is, therefore, estimated. According to the results presented, the efficacy of the model is emphasized as very satisfactory based on the calculated prediction error and 95% confidence region (CR).

The results obtained and the methodology per se is of the interest because traceability is a major issue for regulators and businesses communities that deal with agricultural commodities trading. Especially for soybean, which is a fungible good, such methodology has potential to be helpful for buyers and exporters that must prove the provenance for regulatory and consumer needs purposes.

**R2C1:** The manuscript anchors the sample collection in Latin America countries alleging that soy expansion associated to deforestation is found on those regions. However, the need to prove provenance is not a requirement only for regions in which deforestation is present. The EUDR, legislation mentioned in the manuscript several times to justify the paper, requires proof of origin (plot of land) from all countries supplying the EU with commodities covered by the legislation. The manuscript, therefore, should have collected samples from all

major soybean producers, and not only South American countries. I suggest justify why samples from US were not collected, being the country a major soy producer and subjected to the same requirements as any other country in the scope of the EUDR.

**Response:** This manuscript demonstrates the capabilities of our model to correctly predict soybean origin with high spatial accuracy and precision, regardless of country boundaries. We chose South America as the arena of interest because it encompasses most of the global soybean production in multiple, neighbouring countries. Because plants that grew close to each other are expected to exhibit similar composition, this study area would be the most challenging for provenance prediction, and therefore the most appropriate for testing a model. Other studies have already shown that simple classification models can successfully distinguish between soybeans from different continents, e.g. from China, USA and Brazil (Zhou et al. 2023, Nguyen-Quang et al. 2021; full references in the main text), but they have not demonstrated how to predict the harvest location within a country or non-administrative area. Importantly, the EUDR also operates on a country-risk level, and although technically each exporting country has the same requirement, the level of scrutiny shipments will receive will be dependent on the risk category the exporting country gets. In addition, by far the greatest negative impacts of deforestation come from tropical deforestation, which is virtually non-existent in soybean farming in the US or China. [added in line 110-112]

**R2C2 (see also R2C3, R3C1, R3C4):** It is necessary to discuss the applicability of the model for real operations. Given that initiatives such as the due diligence of the EUDR and the Soy Moratorium are mentioned by the manuscript, I am assuming that the authors expect the methodology be recognized to be used in large-scale operations such as soybean exports. However, the manuscript is silent about this topic. The authors should have pointed out the limitations of the method to be adopted in large-scale operations. I would like to present the limitations I identified.

The first limitation is that soybeans batches loaded on vessels for exports come from several origins. So collecting samples in vessels, or in batches, will not represent all origins from where the soy where harvested. Soybean is a fungible good passing through several loadings from the farm to the exporting port. The origin control must be done at the points of aggregation, such as warehouses and not at the final point, which is the port.

**Response:** We agree that high precision traceability is likely to be more effective when carried out closer to the harvest origin. However, random sampling of shipments and imports, provided that rigorous sampling is undertaken, may be sufficient to identify soybeans from areas flagged for illegal deforestation. We also need to note that under EUDR, there must be harvest location claims for each bean within a container – the regulation does not work on a container but on an individual “product” level, so each harvest location should be listed. The analysis of whole beans in our study ensures a scalable approach from the start to the end of the supply chain. It is not yet known how these regulations will be enforced, and of course, bad actors may try to blend legally- and illegally-sourced commodities, but random sampling may force them to keep the blending to a very low ratio to remain undetected.

To keep the manuscript focused we refrained from discussing these considerations at length. However, we do acknowledge that there are shortcomings and that adaptations will need to take place along the supply chain (added a paragraph in the Discussion, lines 301-323)



**R2C3 (see also R3C1, R3C4):** The second limitation is that collecting samples at the point of aggregation is complex and would need to be integrated with the samples collection already in place for other purposes such as soy quality and moisture classification. However, any warehouse receives soy from hundreds of trucks every day. The very complex process for controlling the origin of each truck at the point of aggregation is the main reason for the Soy Moratorium to identify the provenance using satellite images, which is more applicable for large scale operations. The manuscript quoted the Soy Moratorium, which was appropriated, but it should have made a benchmark of its method with the Soy Moratorium method from the point of view of the applicability for large scale operations. Soy Moratorium methods are well described in the annual reports of the initiative and available at: <https://abiove.org.br/esq/iniciativas/soy-moratorium/>. The second limitation is, therefore, the absence of a benchmark between the method proposed by the manuscript and the Soy Moratorium method.

In my opinion, recognizing the limitations of the proposed methodology for large scale operations and justifying why only Latin America was part of the samples, the manuscript can be published.

**Response:** We have further stressed the limitations of our technology in lines 277-290, 301-323 and clarified why only Latin America was part of the sampling in lines 110-112. Sampling at point of aggregation would be complex indeed, however if we look at the point of EU import, the beans still need to have their declared harvest location (GPS, GEOJSON polygon or shapefile) attached to them, these claims can still be scrutinized with our model. Companies are looking at significant adaptations to their logistics (hence also some of the massive pushback against EUDR), but for our system – the bean and the origin claim are all that's needed. Although satellite imagery might be better for large scale operations, our system is a powerful screening system. Chemical testing is specifically mentioned in the EUDR. Similar legislation for the timber trade (the European Union Timber Regulation - EUTR) has led to a massive uptake by the timber industry of chemical testing for timber origin traceability. We agree that monitoring is crucial for linking specific commodity shipments to deforestation. However, monitoring alone cannot establish provenance of a given sample unless every shipment (and every bean) is tracked continuously throughout its route from the farm to the importing country, which is not feasible. In the timber industry, blockchain technology is used to enable close tracking of individual logs, but the investment in technology and the need for all stakeholders to take it on board slow down the adoption of such methods. Ultimately, achieving traceability in supply chains that currently offer low-level traceability if any, will require adjustment and modifications throughout the supply chain no matter which method is used. The technology we describe here is not meant as a "silver bullet" but a powerful technique to be added to the suite of tools available to law enforcement authorities and other stakeholders.

### **Reviewer 3 comments**

#### **Review of "High-resolution soybean tracing for deforestation-free supply chains"**

This manuscript presents a methodological framework for tracing the geographic origin of soybeans using Gaussian modeling and spatially-referenced chemical fingerprint data. The core contribution is the development of a model that outputs a posterior probability surface,

with performance evaluated through a "Prediction Error," which is the distance from the highest-probability pixel to the true origin pixel, and the area of the "95% Confidence Region," which is the size (area) of pixels that together contain 95% of the posterior probability.

The technical approach is sound and represents a valuable contribution to the field of commodity traceability. However, the practical applicability and scalability of the method are not sufficiently addressed, and several claims regarding its real-world impact appear overstated. The following major and minor points should be addressed to improve the manuscript.

## Major Comments

**R3C1 (see also R2C2, R2C3).** Cost and Scalability Analysis: The authors correctly note that extracting both SIR and TE data is costly (Line 84) but provide no quantitative estimates. For practitioners to adopt this method, a preliminary cost-benefit analysis is essential. The authors should provide rough estimates (or at least a more thorough discussion) for:

a. The financial cost and time required per sample for data collection and analysis.

**Response:** The cost of analysis is highly context-dependent, and varies with service provider (different companies and localities may offer cheaper rates), number of samples (large batches reduce per-sample cost), urgency (premium paid for shorter turnaround time) and so on. Trace Elements analysis costs 50-120 USD per sample, Stable Isotope Ratios comes up at 200-500 USD per sample, and Strontium isotopes ( $^{87}\text{Sr}/^{86}\text{Sr}$ ; potentially useful but not done in this study) costs around 250 USD per sample as well. All prices are dependent on how many samples are tested (which is expected to reduce overall cost as chemical testing is increasingly adopted), how many variables measured, and which variables (some metals are harder than others to detect in TE analysis; some elements cannot be measured simultaneously so multiple runs would may be required; Hydrogen isotope ratio analysis requires more expensive equipment than for Nitrogen or Carbon). We have added a general estimate of price. [lines 303-307]

There will be a need for additional or ongoing sampling irrespective of the traceability method chosen, to meet EUDR, enforcement and stakeholder needs, and companies can play a valuable role in helping to build these databases. However, the pipeline we present is currently the only method that can predict across space—thereby replacing physical samples with statistical predictions—which allows for less dense sampling and lowers the costs of setting up a reference database. [added in lines 316-323]

b. The relationship between sample size (both in terms of geographic coverage and number of samples per location) and model accuracy. This is crucial for designing a cost-effective sampling strategy. Sampling Strategy for Practical Implementation: The claim that the framework can support companies and regulators (Line 111) assumes comprehensive data collection, which is likely prohibitively expensive. The authors should provide a more nuanced discussion of what kind of sampling strategy (i.e., sampling density across origin regions, sampling frequency across traded batches, and the resulting accuracy and confidence of origin predictions) would result in an optimal balance between cost and accuracy, and whether that strategy would be acceptable from a practical point of view.

**Response:** Establishing this relationship will require iterative removal of samples from the



data set and registering how model performance changes by that. We expect that the minimal number of samples required per location is variable and context-dependent (e.g. on species, terrain, farming practices and history etc.), meaning that different samples may carry different importance to the model's predictions. In this paper we describe the technique and demonstrate that it works for an annual agricultural commodity that is not distributed in space in the same way that long-standing perennial populations are, such as in natural forests. The supplemental includes two elimination experiments we have carried out to assess the robustness of origin predictions to systematic data omissions (based on arbitrary latitude and longitude values we chose), but accurately estimating the optimal size of a reference dataset and the ideal locations for sampling would require a separate study. [section 5.2 - Sample collection, lines 391-399]

**R3C2.** Policy-Relevant Performance Metrics: The current metrics (Prediction Error and 95% CR area) are useful but insufficient for policy and enforcement decisions. A more critical performance measure is the decomposition of error into False Positives (incorrectly assigning origin to a deforestation zone) and False Negatives (failing to detect that a shipment originated from a deforestation zone). Under a strict zero-deforestation policy, minimizing false negatives is paramount. The authors should analyze their model's performance from this perspective, as it directly impacts its practical utility for regulators.

**Response:** We carried out the quasi-classification experiments so as to obtain True/False prediction results, meant only to facilitate comparison of model performance between our study and previous commodity traceability studies. This *ad-hoc* addition is meant only for contextualising our study. Table 1 shows the quasi-classification metrics as indication, but our spatially-explicit framework is designed exactly to escape the need to dichotomise prediction results, because the truth is rarely known with certainty in real-world cases. We employ a Bayesian framework to facilitate that, and enable propagation of statistical uncertainty so that the results are a richly detailed distribution of probability values, which the user can then choose to process in any way they wish. Reducing the results to a True/False dichotomy loses much of the information. For example, how far from the harvest origin should a prediction fall to be considered "correct"? 1 km? 10 km? 50 km? We argue that choosing a cut-off is arbitrary, context dependent, and subjective. The results from our model are expressed in absolute terms—spatial extents and probability values—that are readily applicable to real-world uses, and which the user can still dichotomise according to their chosen criteria. Moreover, to establish a link with deforestation zones, we would first need to run our model, and then overlay with satellite imagery in the predicted area to see if deforestation happened before the EUDR cut-off date, and if there is any illegal forest clearing activity still ongoing in the region. The latter part is outside the scope of this study and would require a separate modelling framework (such as the deforestation maps of Global Forest Watch), as we aimed to build a harvest location prediction system that directly links a soybean to a location. [we have clarified the purpose and interpretation of the classification experiments in lines 115 and 209-218]

**R3C3.** Operational Trade-offs: Determination vs. Verification: The authors mention two use cases: "origin determination" and "origin verification" (Line 235). It is implied that verification (discounting unlikely locations) may be less costly. The authors should elaborate on this.

What are the specific savings in data requirements, computational cost, or time when using the model for verification? A discussion, or even a conceptual cost-accuracy curve (characterizing the relationship between accuracy and cost), would greatly help readers assess the practical usefulness of the tool developed, as well as help users understand how to deploy this tool efficiently.

**Response:** Origin verification generally requires a smaller reference dataset, and relies on simpler statistical models that are less computationally costly to run. For example, if a claim is a certain area in a country, then under verification we would only need data from that specific area to either establish plausible doubt or not. It can only establish whether a tested sample fits within the reference set (see Deklerck 2023, Mortier et al. 2024; full references below). However, if the sample does not fit, it is not possible to obtain further information on where it might actually be from. Origin determination requires a larger reference dataset, but it does not need any prior information on the tested sample to establish the likely origin.

We have altered language relating to these concepts throughout the manuscript to clarify the differences between origin verification and origin determination models.

#### References:

- Deklerck, V. (2023). Timber origin verification using mass spectrometry: Challenges, opportunities, and way forward. *Forensic Science International: Animals and Environments*, 3, 100057. <https://doi.org/10.1016/j.fsiae.2022.100057>
- Mortier, T., Truszkowski, J., Norman, M., Boner, M., Buliga, B., Chater, C., Jennings, H., Saunders, J., Sibley, R., Antonelli, A., Waegeman, W., & Deklerck, V. (2024). A framework for tracing timber following the Ukraine invasion. *Nature Plants*, 10(3), 390–401. <https://doi.org/10.1038/s41477-024-01648-5>

**R3C4 (see also R2C2, R2C3).** Limitation: Model Adaptation to Changing Landscapes: The point on legislation needing to keep up with "rapid landscape modification" (Line 258) is well-taken, but this is also a significant limitation of the proposed methodology. As production shifts to new frontiers, the model's predictive power will degrade without continuous, costly re-sampling and re-training. The authors should explicitly discuss this limitation and the associated logistical and financial challenges of maintaining an up-to-date model.

**Response:** We have added more text on the limitations and costs in lines 303-310. In addition, the reviewer identified a valid and important point, but there are a number of factors mitigating the problem. First, in the context of deforestation-risk, commodities from "new frontiers" carry higher deforestation risk by definition, and will therefore be scrutinised as the legislation is designed precisely to disincentivise such products. Second, unlike the classification models used so far, the method we describe extrapolates beyond the locations of reference data points, and the prediction uncertainty is measured in geographically meaningful units (area, distance etc). Model training time is not a limiting factor as it only needs to be done once per reference dataset, and can be done before the relevant commodity arrives at the EU port. Occasional additions and updates to the reference data will be necessary to account for emerging variability in the commodity itself, potentially arising from temporal variation or the introduction of new cultivars. Such updates will be required irrespective of which traceability method, model, or conceptual framework is used (origin verification vs. determination), so this expense is not unique to our method. A good

example to this is the EU JRC Wine database, which undergoes regular updates to the reference dataset. [lines 310-323, 338-346]

Temporal variability and species-specific considerations are also discussed in the Supp. Materials (Section 4).

### **Minor Comments**

**R3C5.** Typos and Language: The manuscript requires thorough proofread. For example, in Line 328, "assessed used" should be "assessed using." Please check the entire document for similar errors.

**Response:** We have corrected this and all other typos we could find.

**R3C6.** Tone Down Exaggerated Claims: The language in the manuscript, particularly in the abstract and conclusion, should be moderated to more accurately reflect the scope of the contribution. For instance, describing the model as a "paradigm shift" is an overstatement. The claims should be framed within the context of a significant methodological advance whose practical implementation still faces considerable hurdles (as outlined in the major comments). **Response:** We have taken these comments on board and moderated the language while better explaining why the claims are warranted in the abstract and conclusion [lines 26 and 350-352]

I appreciate the authors' revisions to the manuscript in response to comments. I continue to believe the study is robust, novel, timely, and has real-world applicability. I have a few follow up points that I hope the authors can address before publication, provided in green in response to the full correspondence to my first set of remarks.

We thank the reviewer for their thorough and careful thought on this study. Our current responses below are given in pink to distinguish from previous correspondence.

## **Reviewer #1 (Remarks to the Author):**

### **GENERAL COMMENTS**

**R1C1 (and R1C17, R1C19):** Sample Selection. The authors do not describe how the locations for samples were selected. This is a critically important part of the methodology that needs to be clearly explained. Was sampling random? Stratified? Convenience? Why did they collect the number of samples that they did? Are there any potential biases in the sample, such as under sampling of recently deforested land, or oversampling of farms held by young people, and are those biases potentially correlated to the variables included in the model (including space and soybean characteristics)? In what way might the sample bias the model?

**Response:** Sampling locations were selected as a result of optimising multiple factors, including but not limited to maximising spatial coverage, keeping within budget for both field expeditions and downstream sample analyses, obtaining farmers' agreement to participate, avoiding areas that suffered damage to crops (e.g. through disease and droughts), and designing a sampling route that coincides with the harvest period in the various regions so that seeds were collected at the correct developmental stage. Prior location selection sometimes had to be adapted by collectors on the ground due to changes in seasonal crop planting choice as well as local security risks. In each location we aimed to take 5 samples to account for local and regional variation in the data. Sampling was arranged in a square with a 10 km diagonal around a central sampling point in each location. This sampling strategy was optimised for several objectives: capturing natural variation, accessibility to collectors, reducing the number of farms sampled, keeping to sample quotas.

Our checks did not locate any consistent bias in the data. Demographic data was not collected so we could not test for correlations with farmer characteristics. [lines 377-382, 388-399]

I appreciate the additions. However, the authors still do not fully describe their sampling strategy (e.g.: to maximize spatial coverage, we did XX; we selected these three countries because YY; because of budget limitations, we aimed to sample XX locations in each county; we randomly identified potential sample sites along a 1 km buffer of major roadways in soy producing regions and approached the farmers in these locations to ask if they would participate. If they declined, we did YY.). To

ensure that the work is reported at the level of reproducibility, a complete sampling strategy that links a specific approach to study objectives or rationale should be provided. It might even differ between countries. This relates, indirectly, to R3s very important comments about the scalability of this general method and the type of sampling strategy that would be needed at scale.

We thank Reviewer 1 for the care they have taken in their response. As our study is centred on identifying deforestation embedded in exported soybeans, our sampling strategy aimed to maximise both spatial coverage and soybean production levels, as well as proximity to deforestation fronts as safely feasible. The collectors then contacted farmers as close as possible to desired sampling locations; if a farmer chose not to be included in the study, the collectors contacted the next nearest farm and so on. We selected the top soy producers—and exporters—in the continent, the countries we sampled produce >94% of the soybeans in South America, and form a contiguous space to enable rigorous testing of our model.

To clarify, the purpose of this study is to develop a model that can identify the origin of soybeans wherever they may be from in South America. As our reference dataset includes hundreds of samples taken from a polygon that spans thousands of kilometres in all directions, the influence of idiosyncratic reference data is strongly reduced. For the same reason, reproducibility of model results would not be reliant on replication of our sampling strategy. As farm locations cannot be disclosed, perfect reproduction of the sampling design is precluded anyhow.

We have reworded the sample collection section to include this information [lines 388-395, 413-416]

**R1C2:** Supplement/Climate and Soil Data. Some of the best parts of the manuscript – i.e., the attempt to use climate and soil data to improve the model – are relegated to the supplement. This (i.e., Section 3 and Section 4 in the supplement) is critical to include in the main text, because as the authors say it has the potential to improve predictive capabilities beyond areas sampled. The finding that soil and climate data do not improve model performance is a fascinating puzzle. Why do the authors think this might be (are soybean characteristics more variable across time than space? Do soy properties vary so predictably across space that x,y coordinates are actually the only variables needed for a good prediction)? I also wonder why the authors didn't incorporate elevation and slope data, which are freely available, sufficiently high resolution for use in these models, and less "modeled" to fill gaps than the soil, precipitation, or temperature data. Or, for that matter, other relevant variables like distance to coast which the authors mention might influence the properties of the soybeans. Why not provide the model with many possible geospatial predictors – just as they did with the soybean sample chemistry data – to see which set of predictors provide the best performance?

**Response:** We thank Reviewer #1 for their observations and comments. We had to decide carefully what to include in the main text due to article length and figure restrictions. We also tried to ensure that the key steps described in the main text were relevant to how the model works and therefore relegated less immediately applicable approaches to supplemental. Our thinking is that the soil and climate data layers currently offer less for model improvement as a good understanding of how they modulate seed composition is lacking. Closing this gap will likely take time, so developing and improving on the framework we describe in the main text is, in our opinion, the most promising route forward. Similarly, including slope and elevation data (distance to coast requires first identifying the relevant coast) added much more data and complexity to our models, whereas the soy suitability data that we used integrates over the effects of variables directly relevant to soybean growth. We add that in a previous study on oak trees from the US, the inclusion of climate variables also did not improve the accuracy or precision of origin predictions. [lines 263-264]

All in all, we think that the models relying on environmental data, although interesting, do not advance the main theme of this study— a working model for origin determination—because they offer poor predictions in their current state. To avoid exceeding article length limits we have added a paragraph summarising this topic in the discussion [lines 252-270], but would like to keep the full section (~430 words, 2 display items +captions) as supplementary material. Should the editor deem it crucial to include in full in the discussion, we think it would best fit after the discussion of the advantages of spatially explicit models in subsection “Outperforming current approaches”.

Okay, I see the authors’ point here. I would consider adding a point to the discussion that the main approach in this paper cannot be extended very well outside regions sampled, but that environmental data could allow for such extension (if indeed those models proved robust). Again, this links with Reviewer 3’s concerns about how to adapt the model to new landscapes.

We thank Reviewer 1 for their comment. The analytic pipeline that we present here relies on the initial creation of a reference dataset which powers the model’s predictions. When generating predictions in regions that are very far from reference data, the model would indeed rely more heavily on auxiliary data such as environmental predictors and as such, uncertainty will increase. Such auxiliary information is itself mostly derived from models rather than measured directly, and its own uncertainty will be propagated into the soybean origin prediction. [lines 258-265]

**R1C4:** Data sensitivity. The authors imply that the data are of a sensitive nature. But soybean maps exist for all of South America, with especially good availability in Brazil. It is well known where soybean farms are. Is there another reason for not releasing the spatial coordinates of the study? Was this part of an agreement with farmers, for instance?

**Response:** the data is sensitive because the use of exact farm location reveals the farmer's and therefore donor's identity. Protecting the farmer's identity aligns with data protection legislation in several relevant countries. Moreover, some farmers have explicitly conditioned their participation on remaining anonymous. EUDR and soybean traceability are highly contentious subjects in Brazil and in other areas of South America, and collector and donor security must not be compromised.

Okay. I suggest adding this as a note to the methods (for clarity) and cite any of the institutional ethics in research protocols at your institutions that cover these sensitive data.

Thanks, we have added a note in the Data Availability section.

## SPECIFIC COMMENTS

### Main Text

**R1C5:** Line 25. Which predictions? This is not clear until reading the whole paper and should be explained in the abstract. **Response:** We have now spelled it out in the text. [line 25]

Thanks for making this change, but what is a target range area? This is not clear and the authors might want to consider further edits for clarity in this general readership journal.

The target area is the range considered for making predictions of origin. This is now clarified in the text [lines 29-30].

**R1C7:** Line 177. "Continental scale" seems like an exaggeration, since the sample did not span the whole continent. **Response:** We respectfully disagree. This study merits the designation "continental scale" as it covers all of South America north of 40°S (>90% of the continent), it includes data from countries that make up ~70% of the continental land area, and it examines broad patterns and processes that operate at this scale.

I say this because like Reviewer 3, I believe it is important to discuss the scope of the work accurately while also recognizing the advance provided by the scholarship. It is true that the authors applied the model to all relatively productive soy suitability area on the South American continent. But the dataset of georeferenced samples definitely doesn't come from all of South America north of 40 degrees south – from figure 1 samples are all below the equator, thus excluding soy grown in Ecuador, Colombia, and Venezuela as well as Peru and Chile. Yet the authors claim in the abstract that they use: "a continental-scale dataset of georeferenced chemical fingerprints from South American soybeans" which is not accurate. I recommend that they accurately describe what was done rather than making such broad claims (e.g.,



instead the language could be: applied a model based on georeferenced chemical footprints from across core soy production and deforestation frontiers in South America, to all South American lands suitable for soy production).

We agree with Reviewer1 that 'continent-scale dataset' could be misconstrued. We now provide alternative wording based on their suggestion: '*....applied a continent-wide model based on georeferenced isotopic and elemental datasets of soybeans from across the main soy growing areas, representing >94% of South American soy production.*' [lines 188-191]

We hope that this clarifies our intention to replace exhaustive sampling with robust statistical inferencing.

**R1C22:** Line 358-361. This is confusing, please clarify: What suggests a greater range? Which previous estimates (there is only one source here). What makes the prior less informative?

**Response:** We have now clarified the sentence [lines 477-480]

I'm not sure this was edited? It was not marked on the track changes doc.

We apologise for missing this and we thank the reviewer for catching the error. Amendments now included in the text [lines 491-495].

**R1C23:** Figure 1. As well as deforestation, I recommend that the authors provide a map of soybean areas. This is available for Brazil annually from MapBiomass. I believe there is at least one dataset for all of South America available:

<https://data.globalforestwatch.org/datasets/soy-planted-area-/about>

**Response:** We accept this request for context, but we think that overlaying soybean farms and deforestation maps as well as our sampling locations would make the figure difficult to interpret. Instead, to facilitate a clearer and more accurate understanding of soybean-related deforestation, Figure 1 caption now refers to a map—also by Global Forest Watch—of forest lost to soybean farms:

<https://gfr.wri.org/forest-extent-indicators/deforestation-agriculture#how-much-forest-has-been-replaced-by-soy> .

I would still recommend that the authors use soy locations instead of or in addition to deforestation locations, since it would help readers understand the sample relative to the overall coverage of soy in South America. While I recognize that the location of soy relative to deforestation is also of interest, it was not a factor in the sampling strategy. My broader point here is transparency to readers about how the sampled locations compare to all the locations of soybean in the study region, and to provide (limited) insight into how the sampling strategy captured geographic variation in soy production.



We interpret “soy locations” in this context as “current soybean farming”, from which we understand that the reviewer thinks that the current range of soy farming is more relevant a context for our study than soy-driven deforestation. The underlying logic would be to show how well our reference dataset represents potential soybean origins, because that is the space into which the model generates predictions. If we follow this logic through, the map of soy suitability (placed in the supplementary material) is very useful as it provides information on where soybeans could come from, whether or not known to be grown there at present.

Nonetheless, we have re-plotted Figure 1 to include soybean production data, and placed the original Figure 1 in the supplementary material alongside a figure showing sampling locations over the soybean suitability map (replacing the original Figure S3).

I appreciate the authors' revisions to the manuscript in response to comments. I continue to believe the study is robust, novel, timely, and has real-world applicability. I have a few follow up points that I hope the authors can address before publication, provided in green in response to the full correspondence to my first set of remarks.

**Reviewer #1 (Remarks to the Author):**

GENERAL COMMENTS

**R1C1 (and R1C17, R1C19):** Sample Selection. The authors do not describe how the locations for samples were selected. This is a critically important part of the methodology that needs to be clearly explained. Was sampling random? Stratified? Convenience? Why did they collect the number of samples that they did? Are there any potential biases in the sample, such as under sampling of recently deforested land, or oversampling of farms held by young people, and are those biases potentially correlated to the variables included in the model (including space and soybean characteristics)? In what way might the sample bias the model?

**Response:** Sampling locations were selected as a result of optimising multiple factors, including but not limited to maximising spatial coverage, keeping within budget for both field expeditions and downstream sample analyses, obtaining farmers' agreement to participate, avoiding areas that suffered damage to crops (e.g. through disease and droughts), and designing a sampling route that coincides with the harvest period in the various regions so that seeds were collected at the correct developmental stage. Prior location selection sometimes had to be adapted by collectors on the ground due to changes in seasonal crop planting choice as well as local security risks. In each location we aimed to take 5 samples to account for local and regional variation in the data. Sampling was arranged in a square with a 10 km diagonal around a central sampling point in each location. This sampling strategy was optimised for several objectives: capturing natural variation, accessibility to collectors, reducing the number of farms sampled, keeping to sample quotas.

Our checks did not locate any consistent bias in the data. Demographic data was not collected so we could not test for correlations with farmer characteristics. [lines 377-382, 388-399]

I appreciate the additions. However, the authors still do not fully describe their sampling strategy (e.g.: to maximize spatial coverage, we did XX; we selected these three countries because YY; because of budget limitations, we aimed to sample XX locations in each county; we randomly identified potential sample sites along a 1 km buffer of major roadways in soy producing regions and approached the farmers in these locations to ask if they would participate. If they declined, we did YY.). To ensure that the work is reported at the level of reproducibility, a complete sampling strategy that links a specific approach to study objectives or rationale should be provided. It might even differ between countries. This relates, indirectly, to R3s very important comments about the scalability of this general method and the type of sampling strategy that would be needed at scale.

**R1C2:** Supplement/Climate and Soil Data. Some of the best parts of the manuscript – i.e., the attempt to use climate and soil data to improve the model – are relegated to the supplement. This (i.e., Section 3 and Section 4 in the supplement) is critical to include in the main text, because as the authors say it has the potential to improve predictive capabilities beyond areas sampled. The finding that soil and climate data do not improve model performance is a fascinating puzzle. Why do the

authors think this might be (are soybean characteristics more variable across time than space? Do soy properties vary so predictably across space that x,y coordinates are actually the only variables needed for a good prediction)? I also wonder why the authors didn't incorporate elevation and slope data, which are freely available, sufficiently high resolution for use in these models, and less "modeled" to fill gaps than the soil, precipitation, or temperature data. Or, for that matter, other relevant variables like distance to coast which the authors mention might influence the properties of the soybeans. Why not provide the model with many possible geospatial predictors – just as they did with the soybean sample chemistry data – to see which set of predictors provide the best performance?

**Response:** We thank Reviewer #1 for their observations and comments. We had to decide carefully what to include in the main text due to article length and figure restrictions. We also tried to ensure that the key steps described in the main text were relevant to how the model works and therefore relegated less immediately applicable approaches to supplemental. Our thinking is that the soil and climate data layers currently offer less for model improvement as a good understanding of how they modulate seed composition is lacking. Closing this gap will likely take time, so developing and improving on the framework we describe in the main text is, in our opinion, the most promising route forward. Similarly, including slope and elevation data (distance to coast requires first identifying the relevant coast) added much more data and complexity to our models, whereas the soy suitability data that we used integrates over the effects of variables directly relevant to soybean growth. We add that in a previous study on oak trees from the US, the inclusion of climate variables also did not improve the accuracy or precision of origin predictions. [lines 263-264]

All in all, we think that the models relying on environmental data, although interesting, do not advance the main theme of this study— a working model for origin determination—because they offer poor predictions in their current state. To avoid exceeding article length limits we have added a paragraph summarising this topic in the discussion [lines 252-270], but would like to keep the full section (~430 words, 2 display items +captions) as supplementary material. Should the editor deem it crucial to include in full in the discussion, we think it would best fit after the discussion of the advantages of spatially explicit models in subsection "Outperforming current approaches".

Okay, I see the authors' point here. I would consider adding a point to the discussion that the main approach in this paper cannot be extended very well outside regions sampled, but that environmental data could allow for such extension (if indeed those models proved robust). Again, this links with Reviewer 3's concerns about how to adapt the model to new landscapes.

**R1C4:** Data sensitivity. The authors imply that the data are of a sensitive nature. But soybean maps exist for all of South America, with especially good availability in Brazil. It is well known where soybean farms are. Is there another reason for not releasing the spatial coordinates of the study? Was this part of an agreement with farmers, for instance?

**Response:** the data is sensitive because the use of exact farm location reveals the farmer's and therefore donor's identity. Protecting the farmer's identity aligns with data protection legislation in several relevant countries. Moreover, some farmers have explicitly conditioned their participation on remaining anonymous. EUDR and soybean traceability are highly contentious subjects in Brazil and in other areas of South America, and collector and donor security must not be compromised.

Okay. I suggest adding this as a note to the methods (for clarity) and cite any of the institutional ethics in research protocols at your institutions that cover these sensitive data.

#### SPECIFIC COMMENTS

##### Main Text

**R1C5:** Line 25. Which predictions? This is not clear until reading the whole paper and should be explained in the abstract. **Response:** We have now spelled it out in the text. [line 25]

Thanks for making this change, but what is a target range area? This is not clear and the authors might want to consider further edits for clarity in this general readership journal.

**R1C7:** Line 177. “Continental scale” seems like an exaggeration, since the sample did not span the whole continent. **Response:** We respectfully disagree. This study merits the designation “continental scale” as it covers all of South America north of 40°S (>90% of the continent), it includes data from countries that make up ~70% of the continental land area, and it examines broad patterns and processes that operate at this scale.

I say this because like Reviewer 3, I believe it is important to discuss the scope of the work accurately while also recognizing the advance provided by the scholarship. It is true that the authors applied the model to all relatively productive soy suitability area on the South American continent. But the dataset of georeferenced samples definitely doesn’t come from all of South America north of 40 degrees south – from figure 1 samples are all below the equator, thus excluding soy grown in Ecuador, Colombia, and Venezuela as well as Peru and Chile. Yet the authors claim in the abstract that they use: “a continental-scale dataset of georeferenced chemical fingerprints from South American soybeans” which is not accurate. I recommend that they accurately describe what was done rather than making such broad claims (e.g., instead the language could be: applied a model based on georeferenced chemical footprints from across core soy production and deforestation frontiers in South America, to all South American lands suitable for soy production).

**R1C22:** Line 358-361. This is confusing, please clarify: What suggests a greater range? Which previous estimates (there is only one source here). What makes the prior less informative?

**Response:** We have now clarified the sentence [lines 477-480]

I’m not sure this was edited? It was not marked on the track changes doc.

**R1C23:** Figure 1. As well as deforestation, I recommend that the authors provide a map of soybean areas. This is available for Brazil annually from MapBiomass. I believe there is at least one dataset for all of South America available: <https://data.globalforestwatch.org/datasets/soy-planted-area/about>

**Response:** We accept this request for context, but we think that overlaying soybean farms and deforestation maps as well as our sampling locations would make the figure difficult to interpret. Instead, to facilitate a clearer and more accurate understanding of soybean-related deforestation, Figure 1 caption now refers to a map—also by Global Forest Watch—of forest lost to soybean farms: <https://gfr.wri.org/forest-extent-indicators/deforestation-agriculture#how-much-forest-has-been-replaced-by-soy>.

I would still recommend that the authors use soy locations instead of or in addition to deforestation locations, since it would help readers understand the sample relative to the overall coverage of soy in

South America. While I recognize that the location of soy relative to deforestation is also of interest, it was not a factor in the sampling strategy. My broader point here is transparency to readers about how the sampled locations compare to all the locations of soybean in the study region, and to provide (limited) insight into how the sampling strategy captured geographic variation in soy production.

## **Review of “High-resolution soybean tracing for deforestation-free supply chains”**

This manuscript presents a methodological framework for tracing the geographic origin of soybeans using Gaussian modeling and spatially-referenced chemical fingerprint data. The core contribution is the development of a model that outputs a posterior probability surface, with performance evaluated through a "Prediction Error," which is the distance from the highest-probability pixel to the true origin pixel, and the area of the "95% Confidence Region," which is the size (area) of pixels that together contain 95% of the posterior probability.

The technical approach is sound and represents a valuable contribution to the field of commodity traceability. However, the practical applicability and scalability of the method are not sufficiently addressed, and several claims regarding its real-world impact appear overstated. The following major and minor points should be addressed to improve the manuscript.

### **Major Comments**

1. **Cost and Scalability Analysis:** The authors correctly note that extracting both SIR and TE data is costly (Line 84) but provide no quantitative estimates. For practitioners to adopt this method, a preliminary cost-benefit analysis is essential. The authors should provide rough estimates (or at least a more thorough discussion) for:
  - a. The financial cost and time required per sample for data collection and analysis.
  - b. The relationship between sample size (both in terms of geographic coverage and number of samples per location) and model accuracy. This is crucial for designing a cost-effective sampling strategy.
2. **Sampling Strategy for Practical Implementation:** The claim that the framework can support companies and regulators (Line 111) assumes comprehensive data collection, which is likely prohibitively expensive. The authors should provide a more nuanced discussion of what kind of sampling strategy (i.e., sampling density across origin regions, sampling frequency across traded batches, and the resulting accuracy and confidence of origin predictions) would result in an optimal balance between cost and accuracy, and whether that strategy would be acceptable from a practical point of view.
3. **Policy-Relevant Performance Metrics:** The current metrics (Prediction Error and 95% CR area) are useful but insufficient for policy and enforcement decisions. A more critical performance measure is the decomposition of error into False Positives (incorrectly assigning origin to a deforestation zone) and False Negatives (failing to

detect that a shipment originated from a deforestation zone). Under a strict zero-deforestation policy, minimizing false negatives is paramount. The authors should analyze their model's performance from this perspective, as it directly impacts its practical utility for regulators.

4. **Operational Trade-offs: Determination vs. Verification:** The authors mention two use cases: "origin determination" and "origin verification" (Line 235). It is implied that verification (discounting unlikely locations) may be less costly. The authors should elaborate on this. What are the specific savings in data requirements, computational cost, or time when using the model for verification? A discussion, or even a conceptual cost-accuracy curve (characterizing the relationship between accuracy and cost), would greatly help readers assess the practical usefulness of the tool developed, as well as help users understand how to deploy this tool efficiently.
5. **Limitation: Model Adaptation to Changing Landscapes:** The point on legislation needing to keep up with "rapid landscape modification" (Line 258) is well-taken, but this is also a significant limitation of the proposed methodology. As production shifts to new frontiers, the model's predictive power will degrade without continuous, costly re-sampling and re-training. The authors should explicitly discuss this limitation and the associated logistical and financial challenges of maintaining an up-to-date model.

## **Minor Comments**

1. **Typos and Language:** The manuscript requires thorough proofread. For example, in Line 328, "assessed used" should be "assessed using." Please check the entire document for similar errors.
2. **Tone Down Exaggerated Claims:** The language in the manuscript, particularly in the abstract and conclusion, should be moderated to more accurately reflect the scope of the contribution. For instance, describing the model as a "paradigm shift" is an overstatement. The claims should be framed within the context of a significant methodological advance whose practical implementation still faces considerable hurdles (as outlined in the major comments).