Author’s Response to Reviewer #2 Comments

We sincerely thank Reviewer 2 for their thoughtful and constructive review of our manuscript. The reviewer noted that the GRAIN dataset is valuable and timely, with a logically structured and robust methodology, while identifying several points that would benefit from clarification and contextualization. We have carefully addressed each comment in an itemized manner. Reviewer comments are italicized, and our responses are provided in blue colored text. Line numbers referenced correspond to the track changes version of the revised manuscript.

Below, we summarize the major revisions and clarifications made to the manuscript in response to Reviewer 2’s comments:

- Explicitly acknowledged potential regional bias arising from the concentration of training data in United States and India.
- Clarified the sampling strategy underlying feature distributions shown in Figure 4.
- Clarified the rationale for using Mean Offset Distance (MOD) as a validation metric.
- Explained the handling of transboundary canals in country-level statistics.
- Expanded the Future Work section to highlight the value of incorporating canal lining and other canal attributes.

We believe that these revisions and clarifications address Reviewer 2’s comments and further strengthen the transparency, rigor, and usability of the GRAIN dataset.

Detailed Responses to Reviewer #2

We thank the Reviewer for their thoughtful and constructive assessment of our manuscript. The reviewer notes that “the manuscript is logically structured, clearly presented, and robust in its methodology,” and that the resulting GRAIN dataset is “highly valuable and timely, effectively filling a significant gap in global-scale irrigation infrastructure data.” The reviewer also raises several specific points aimed at clarifying methodological choices, validation metrics, and the interpretation of selected results.

Reviewer’s general assessment: *‘This paper presents a machine learning-based classification workflow to reclassify OSM hydrographic data, successfully extracting a global network of agricultural irrigation canals. This study and the resulting Global Irrigation Canal (GRAIN) dataset are highly valuable and timely, effectively filling a significant gap in global-scale irrigation infrastructure data. Overall, the manuscript is logically structured, clearly presented, and robust in its methodology. However, I have several specific concerns that need to be addressed’*

Our response: We sincerely thank the reviewer for this positive evaluation of the scientific contribution and relevance of the study. We have carefully gone through every point raised by the reviewer and have addressed these through targeted clarifications and revisions to the manuscript, as detailed in the itemized responses below.

Itemized response to reviewer #2 comments:

Note that all line numbers mentioned below refer to the track-changes version of the revised manuscript

Comment 1: *‘The training data comes primarily from canal datasets in the United States and India, which may introduce regional bias and affect the model's generalization ability in other global irrigation systems. This limitation should be explicitly stated in the discussion.’*

Our Response: We thank the reviewer for highlighting this important point. We agree that the training data used to develop the Random Forest classifier is primarily derived from national-scale canal inventories in the United States and India, and that this may introduce regional bias that could influence model generalization across diverse global irrigation systems.

We have now explicitly acknowledged this limitation in the Discussion section of the revised manuscript between **lines 500-510**. We clarify that while the canal systems in the United States and India encompass a wide range of canal characteristics, including different geometries, sizes, and management practices, the training data does not fully capture the global diversity of irrigation infrastructure. As a result, some degree of regional bias may be present, particularly in regions where irrigation canals differ substantially in geometry, construction practices, or mapping completeness in OpenStreetMap.

Comment 2: ‘*Figure 4 illustrates significant distinctions between natural rivers and artificial canals across five geometric and topographic features (e.g., sinuosity/straightness ratio, slope, turning angle). Did the authors ensure that the sampled rivers and canals were located within similar geographical or topographical environments during this comparison?’*

Our Response: We thank the reviewer for this clarification. We would like to note that strict geographical matching was not done between river and canal geometries used for training. Figure 4 presents feature distributions computed using the full set of 20,000 training samples used to train the Random Forest classifier. These samples were drawn through random sampling from all locations where reliable in-situ reference data were available. For canal samples, this corresponds primarily to national-scale inventories in the United States and India, while river samples were drawn from OSM river segments intersecting the SWORD database across all 95 countries included in the GRAIN workflow.

This sampling was intended to expose the classifier to a wide range of geometries, particularly for natural rivers, which can exhibit strong regional variability. While the canal training data spans a more limited set of geographic contexts, irrigation canals are engineered features constrained by design and operational requirements and therefore tend to exhibit more consistent geometric characteristics across regions. As a result, differences in geographic context are less likely to substantially alter the distinguishing geometric signatures between canals and natural rivers. This has been clarified between **lines 225-235**.

Comment 3: ‘*In the validation section, the authors use Mean Offset Distance (MOD) to evaluate the spatial accuracy of the GRAIN dataset. However, the machine learning model designed in this paper acts as a classifier to distinguish whether an OSM segment belongs to a "canal" or a "river", and does not require any geometric coordinate correction of the original OSM vectors. Therefore, the canal positions in the final dataset will fully follow the inherent spatial biases existing in OSM. Why was MOD chosen as the core validation metric to evaluate the "accuracy" of this classification method?’*

Our Response: We thank the reviewer for this important clarification and agree with their assessment. As noted, no geometric corrections are applied to the original OpenStreetMap (OSM) vectors in the GRAIN workflow, and the resulting canal geometries therefore inherit any spatial biases present in the source OSM data.

Mean Offset Distance (MOD) is therefore not intended to evaluate the performance of the Random Forest classifier itself, which functions solely as a semantic classifier distinguishing canals from rivers. Instead, MOD is reported alongside recall as a complementary metric to assess the spatial usability of the final GRAIN product when compared against independent in-situ canal datasets that are assumed to have higher positional accuracy. In this context, MOD provides users with a quantitative indication of the typical spatial offset that may be expected when using OSM-derived, classified canal geometries for regional or global analyses.

This has now been explicitly clarified in the Validation section of the revised manuscript between **lines 370-380.**

Comment 4: *'Some canals are transboundary, such as the famous Karakum Canal. When calculating statistics in Figure 9, how were the lengths or proportions of such canals allocated between nations for these statistics?'*

Our Response: We thank the reviewer for this clarification. Country-level statistics presented in Figure 9 were computed by aggregating canal lengths within national administrative boundaries. Accordingly, in the case of transboundary canals, such as the Karakum Canal, only the portion of the canal geometry that lies within a given country's boundary contributes to that country's total canal length and associated density statistics.

This boundary-based accounting avoids double counting and ensures consistency in national-level comparisons. We have clarified this approach in the revised manuscript between **lines 455-465.**

Comment 5: '*Figure 10 is very interesting, highlighting a non-linear relationship between crop yield and the density of the agricultural canal network.'*

Our Response: We thank the reviewer for this positive observation. We agree that Figure 10 does highlight an interesting relationship between cereal yield and canal density. However, we would like to note that the figure is intended to present a high-level contextualization of the association between national-scale agricultural canal density and cereal yield. In line with other reviewer comments, we have also added clarification in the revised manuscript between **lines 430-440** to indicate that the observed relationship is likely influenced by multiple confounding factors and should be treated as indicative, rather than causal.

Comment 6: *'If the lining status of the canals could be provided based on the canal network while presenting the global canal network, it would have greater significance for calculating conveyance efficiency in agricultural irrigation and water resource management applications. Of course, this issue is not one that this article needs to solve, but rather a possible direction for the future.'*

Our Response: We thank the reviewer for this valuable suggestion and agree that information on canal lining status would substantially enhance the applicability of a global canal network for conveyance efficiency and agricultural water management analyses. At present, canal lining information is not available at a global level in OSM or other open datasets. We have therefore noted this in the future works section between **lines 540-545**.