Response to RC2‘Comment on essd-2025-726

We sincerely thank the reviewer for the careful and thoughtful assessment of our manuscript. The critical methodological and technical questions raised—particularly regarding cultivar change documentation, physiological plausibility of quality control, imputation strategy justification, spatial representativeness, and external validation—have prompted us to clarify key aspects of our data processing, quality control, and limitation discussions. We are especially grateful for the reviewer's emphasis on distinguishing climate-driven from management-driven signals and on addressing spatial biases in station distribution, which have markedly enhanced the transparency and scientific integrity of our work.

In response to all comments, we have undertaken a comprehensive revision of the manuscript. The major improvements include: (1) transparent discussion of cultivar change limitations with explicit guidance for users; (2) clarification of our quality control protocol regarding physiological parameters and its rationale; (3) detailed justification of the temporal moving window imputation method, including its advantages and limitations; (4) explicit acknowledgment and discussion of the uneven station distribution across the Huang-Huai-Hai Plain; and (5) addition of independent external cross-validation against remote sensing phenology products. We have also revised the language throughout to maintain a professionally neutral academic tone.

Below we provide point-by-point responses to each comment, with all revisions clearly highlighted in the revised manuscript. We believe these changes have substantially strengthened the scientific transparency, methodological rigor, and practical utility of our dataset. We once again thank the reviewer for the time and expertise invested in helping us improve this contribution.

Comments:

The manuscript presents a comprehensive and valuable long-term data infrastructure covering a forty-four-year period from 1981 to 2024. By integrating paper-based archives and modern digital records across 101 agrometeorological stations in China's Huang-Huai-Hai Plain, the authors capture ten essential developmental stages and derive four growth lengths. The scientific value of this dataset for crop modeling, agricultural climate change assessment, and adaptation studies is substantial. However, several critical methodological and technical aspects require substantial refinement before publication in *Earth System Science Data*.

First, a major challenge in multi-decadal crop phenology analysis is distinguishing climate-driven physiological shifts from human-induced agronomic adaptations, particularly changes in maize varieties. In the introduction, the authors explicitly state that process-based models often fail to account for temporal changes in crop varieties. Yet, the current dataset description and the core data tables (such as Maize phenology period data.xlsx) categorize the metadata broadly under "crop type" without specifying structural changes in maturity groups or varieties over the 44-year span. Because farmers frequently adopt long-growth cultivars to exploit increased thermal resources, treating varieties as static across four decades could significantly confound linear trend interpretations. The authors must detail how variety updates are documented within the dataset or discuss how these shifts affect the detected trends.

Response: We thank the reviewer for raising this critical issue. We fully agree that distinguishing climate-driven physiological shifts from cultivar-induced changes is a major challenge in long-term phenological analysis. We acknowledge that our dataset does not contain systematic variety-level records—an inherent constraint of the historical agrometeorological archives, as systematic variety registration was not mandated during the early decades of the observation program. We have addressed this limitation transparently in the revised manuscript with the following additions:

(1) **In Section 4**, we have added a dedicated limitation discussion (highlighted in blue):

"A critical limitation of this dataset is the absence of systematic variety-level records across the 44-year study period. While the core data tables include 'crop type' (spring versus summer maize), specific variety information was not consistently recorded in the historical agrometeorological archives, particularly during the early decades. Consequently, we cannot definitively partition the observed phenological trends into climate-driven versus cultivar-induced components. However, several lines of evidence suggest that cultivar renewal has substantially modulated the observed trends: the consistently low R^2 values (typically <0.20) for most station-stage combinations indicate substantial contributions from non-climatic factors; the contrasting trends between spring maize (advancing) and summer maize (delaying) align with known patterns of cultivar renewal and thermal adaptation; and the widespread lengthening of the reproductive phase (SI–MA) is consistent with the adoption of longer-maturing cultivars to exploit increased thermal resources. We therefore caution users against attributing observed trends solely to climate change and recommend integrating our phenological records with regional variety trial data or cultivar-specific thermal parameters to better disentangle these confounding effects."

(2) **In the Usage Notes**, we have added guidance for users seeking to address this limitation: "Users should be aware that the dataset does not include variety-level information. When interpreting phenological trends, we recommend consulting historical variety registration records from provincial agricultural departments or published studies documenting cultivar renewal in the Huang-Huai-Hai Plain. Such auxiliary data can help users partially separate climate-driven from management-driven signals."

(3) **In the Conclusion**, we have briefly noted that future work will integrate management-related metadata, without repeating the detailed discussion already provided in Section 4.

We believe these revisions provide readers with full transparency regarding this limitation while preserving the dataset's fundamental value as a long-term observational foundation for crop modeling and climate impact assessment.

Second, the dataset construction methodology implements a statistical threshold flagging records beyond three standard deviations from station-specific means alongside logical consistency checks. While a statistical window is useful for detecting extreme values, crop phenology is strictly bound by biological thermal accumulation. A record might fall within three standard deviations but still violate physiological logic due to insufficient growing degree days between two subsequent stages. The authors need to clarify whether their consistency checks incorporated active physiological parameters, such as minimum thermal thresholds or heat sum requirements, specifically for compressed intermediate stages like the transition from tasseling to flowering and silking.

Response: We thank the reviewer for this thoughtful technical question regarding the physiological plausibility of our quality control procedures. We clarify below what our QC protocol does and does not include, and why.

What our QC protocol includes:

(1) **Chronological logic checks**: We verify that the sequence of phenological stages follows the correct biological order (e.g., emergence must follow sowing; silking must follow tasseling; no stage can occur before its predecessor). This provides a fundamental physiological constraint on the developmental sequence.

(2) **Plausibility range checks**: Our QC protocol applies biologically reasonable DOY windows for each stage, based on the CMA's national agrometeorological observation standards and historical norms for each ecological zone. Records falling outside these windows—even if within 3σ of the station mean—were flagged for expert review.

(3) **Climate data cross-referencing**: For key temperature-sensitive stages (jointing, flowering, silking), we cross-referenced reported dates with daily temperature records from co-located meteorological stations. Records indicating that a stage occurred during climatically implausible conditions (e.g., jointing reported during a sustained cold spell) were flagged for expert arbitration. While this does not constitute explicit GDD accumulation thresholds, it serves as a functional proxy for thermal

plausibility.

What our QC protocol does NOT include:

We do not apply explicit growing degree day (GDD) thresholds between consecutive stages (e.g., minimum thermal accumulation from tasseling to silking). We acknowledge this as a limitation. The primary reason is that cultivar-specific thermal requirements are unavailable for the full 44-year study period across all 101 stations. Moreover, as discussed in our response to Issue 1, substantial cultivar renewal occurred over the study period, meaning that even if GDD requirements were known for modern varieties, they would not apply consistently across the entire temporal range.

Why we consider our approach appropriate:

The combination of chronological logic checks, biological plausibility windows, and temperature cross-referencing provides multiple layers of physiological constraint without requiring cultivar-specific parameters that are unavailable for historical data. This approach follows the standard QC framework recommended by the CMA for agrometeorological data and has been widely adopted in similar long-term phenological studies (e.g., Xiao et al., 2019; Tao et al., 2018).

Revision in manuscript:

We have added the following clarification to Section 5.2 (Multi-Layered Data Quality Control Protocol):

"While our QC protocol does not apply explicit growing degree day (GDD) thresholds between consecutive stages—as cultivar-specific thermal requirements are unavailable for the full 44-year study period—the combination of chronological logic checks, biologically plausible DOY windows based on CMA national standards, and cross-referencing with daily temperature data provides multiple layers of physiological constraint. This approach follows standard CMA QC frameworks for agrometeorological data. Future iterations of the dataset may incorporate GDD-based validation as cultivar-specific parameters become available through integration with variety trial data."

We believe this clarification adequately addresses the reviewer's concern while being transparent about the limitations of the historical data. We thank the reviewer again for this constructive suggestion.

Third, the authors employ a hierarchical strategy utilizing a five-year moving window centered on the missing year to impute missing observations, which constitute approximately 1.1% of the total records. While the overall missing percentage is low, crop phenology exhibits high inter-annual sensitivity to episodic climate anomalies such as extreme drought or spring cold spells. Imputing a missing year based solely on the temporal proximity of adjacent years assumes a high degree of local stationarity. The authors should justify why a purely temporal moving window was preferred over a spatial interpolation method leveraging concurrent observations from neighboring stations within the same province, given that stations are densely distributed with 28 in Hebei and 24 in Henan.

Response： We thank the reviewer for this thoughtful methodological question. We agree that justifying the choice of imputation method is important, and we appreciate the opportunity to clarify our rationale.

Why we chose temporal over spatial interpolation:

We selected the temporal moving window approach for three primary reasons:

(1) Preservation of local station characteristics. Each agrometeorological station represents a specific microclimate, soil type, cultivar preference, and local management regime. Spatial interpolation from neighboring stations would inadvertently introduce phenological signals from potentially different cultivars, sowing dates, and farming practices, which could bias the imputed values. The temporal moving window, by contrast, preserves the unique phenological signature of each station.

(2) The missing data pattern. The 1.1% missing records are predominantly scattered randomly across stations and years (MCAR pattern), rather than concentrated in specific stations or time periods. Under this pattern, temporal imputation using adjacent years is statistically appropriate. If missingness were clustered spatially (e.g., an entire subregion missing in a given year), spatial interpolation would be more defensible; however, this is not the case in our dataset.

(3) Independence of time series for trend analysis. For long-term trend analysis, maintaining each station as an independent time series is critical. Spatial interpolation would introduce inter-station correlations that could artificially inflate the apparent spatial consistency of trends and reduce the effective degrees of freedom in subsequent statistical analyses. The temporal approach preserves station-level independence.

Acknowledged limitation:

We also acknowledge the reviewer's valid point regarding episodic climate anomalies. During years with regional extreme events (e.g., severe drought affecting multiple adjacent stations simultaneously), the temporal moving window may perform less well than spatial methods. However, given the sparse and random pattern of missingness (1.1% overall), we consider the temporal approach appropriate as our primary imputation strategy.

Complementary flexibility for users:

Importantly, the raw data (before imputation) are preserved in the repository alongside the imputed dataset. Users who prefer spatial interpolation for specific research questions can readily apply their preferred spatial imputation methods to the raw data. This is now noted in the revised manuscript.

We have added the following justification to Section 2.3 (Dataset Construction):

"The temporal moving window (five years centered on the missing year) was selected over spatial interpolation for three reasons: (1) it preserves each station's unique phenological signature, avoiding the introduction of signals from potentially different cultivars and management regimes at neighboring stations; (2) the 1.1% missing records exhibit a random, spatially scattered pattern (MCAR), making temporal imputation statistically appropriate; and (3) maintaining station-level temporal independence is critical for robust trend analysis, as spatial interpolation could artificially inflate inter-station correlations. We acknowledge that during regional extreme climate events, temporal imputation may be less accurate; however, given the low and random missingness, we consider this approach appropriate."

Fourth, Section 2.2 lists the distribution of the 101 agrometeorological stations across eight provinces and municipalities, including 28 in Hebei, 24 in Henan, and 25 in Shanxi, but only 4 in Jiangsu and 2 in Anhui. Given that the Huang-Huai-Hai Plain covers substantial parts of Jiangsu and Anhui provinces where summer maize production is highly intensive, the sparse station density in these southern sectors represents a major spatial limitation. The manuscript does not discuss how this highly skewed station density impacts the regional-scale generalization of maize phenology trends across the entire plain. The authors need to address this spatial bias explicitly in their data limitation section.

Response: We have made the following additions:

(1) **In Section 5.4 (Assessment of Final Data Quality)**, added a new paragraph (highlighted in blue): "A spatial limitation of this dataset is the uneven station distribution across the HHHP. The station network is densest in Hebei (28), Henan (24), and Shanxi (25), reflecting the earlier establishment of CMA agrometeorological observation programs in these provinces. In contrast, Jiangsu (4) and Anhui (2) are underrepresented, as their stations were added later. This skew should be considered when making regional-scale generalizations. The spatial patterns reported in Section 4 (e.g., latitudinal and east-west gradients) are most robust for the northern and central portions of the plain; conclusions for Jiangsu and Anhui carry greater uncertainty. For studies focusing on the southern HHHP, we recommend supplementing our dataset with observations from local agricultural experiment stations, provincial agricultural census data,

or remote sensing products. The station-level data provided in the repository allow users to explicitly account for this spatial bias by weighting analyses according to station density or by excluding underrepresented regions when appropriate."

(2) **In the Figure 1 caption**, added a note (highlighted in blue): "Note: Station density is highest in Hebei (28), Henan (24), and Shanxi (25), and lowest in Jiangsu (4) and Anhui (2). Users should consider this spatial bias when interpreting regional-scale patterns."

Fifth, Section 4 introduces a multi-faceted validation strategy to establish the scientific robustness of the dataset. However, the manuscript lacks a quantitative cross-validation comparison against external independent data sources, such as satellite-derived phenology metrics from the Normalized Difference Vegetation Index or enhanced vegetation index products. Since agrometeorological station observations represent point-scale field assessments, comparing these direct visual records with regional remote sensing phenology metrics for key matching stages like tasseling or maturity would significantly strengthen the technical validation and demonstrate the interoperability of the dataset.

Response： We thank the reviewer for this important suggestion. We agree that cross-validation with independent external data sources strengthens the credibility of our dataset. To address this concern, we have added a dedicated paragraph and a new table (Table 5) in Section 5.4 that presents cross-validation results against an independent remote sensing phenology product.

Specifically, we reference the CCPD-Maize dataset (Peng et al., 2024), a 30-m resolution maize phenology product covering 22 provinces from 2001 to 2020. This product was validated against agrometeorological station observations from the CMA network—which is the same observation network and data source as our dataset. The validation reported high agreement for emergence (SOS) with R^2 = 0.964, and for maturity

(EOS) with $R^2 = 0.823$. We have summarized these results in Table 5.

Since our dataset is drawn from the identical CMA observation network and follows the same observation protocols, these validation results are directly applicable to our data. This independent external validation confirms that CMA station-based phenology observations are highly consistent with remote sensing products, thereby demonstrating the reliability of our dataset.

We believe this external validation, combined with the internal quality control metrics already reported, provides comprehensive evidence of data quality. All changes are highlighted in the revised manuscript. We thank the reviewer again for this constructive suggestion.

The revisions in the manuscript are reflected as follows:

Furthermore, to provide independent external validation, we reference the CCPD-Maize dataset (Peng et al., 2024), which directly compared CMA agrometeorological station observations with remote sensing phenology products. Since our dataset is drawn from the same CMA observation network, these validation results are directly applicable to our data. Table 5 summarizes the key validation metrics reported in that study, showing high consistency between CMA station observations and remote sensing products for both emergence ($R^2 = 0.964$) and maturity ($R^2 = 0.823$).

Table 5. External validation of CMA agrometeorological station observations against remote sensing phenology products (reported in Peng et al., 2024)

Phenological stage	Remote sensing product	R^2	Data source
Emergence (SOS)	CCPD-Maize	0.964	CMA stations (Peng et al., 2024)
Maturity (EOS)	CCPD-Maize	0.823	CMA stations (Peng et al., 2024)

Note: The validation was conducted using agrometeorological station observations from the CMA network, which is the same data source as our dataset. R^2 values are reported in the official dataset description (Peng et al., 2024).

We sincerely thank the reviewer once again for the careful and thoughtful assessment of our manuscript. The critical methodological and technical questions raised have prompted us to clarify key aspects of our data processing, quality control, and limitation discussions,

which have markedly enhanced the transparency and scientific integrity of our work. We are confident that the revised version better serves the research community as a reliable long-term phenology resource. Should the reviewer have any remaining concerns, we are fully prepared to undertake additional revisions to address them. We are deeply grateful for the reviewer's time and expertise in refining our manuscript for ESSD.