SUPPLEMENTARY MATERIALS OF "CROPNET: AN OPEN LARGE-SCALE DATASET WITH MULTIPLE MODALITIES FOR CLIMATE CHANGE-AWARE CROP YIELD PREDICTIONS"

Anonymous authors
Paper under double-blind review

OUTLINE

This document provided supplementary materials to support our main paper. In particular, Section A offers the details of three data sources. Section B illustrates more examples of Our CropNet dataset. Section C provides details of data collection. Section D presents additional experimental settings and results.

A DATA SOURCES

Our CropNet dataset is crafted from three different data sources, as listed below.

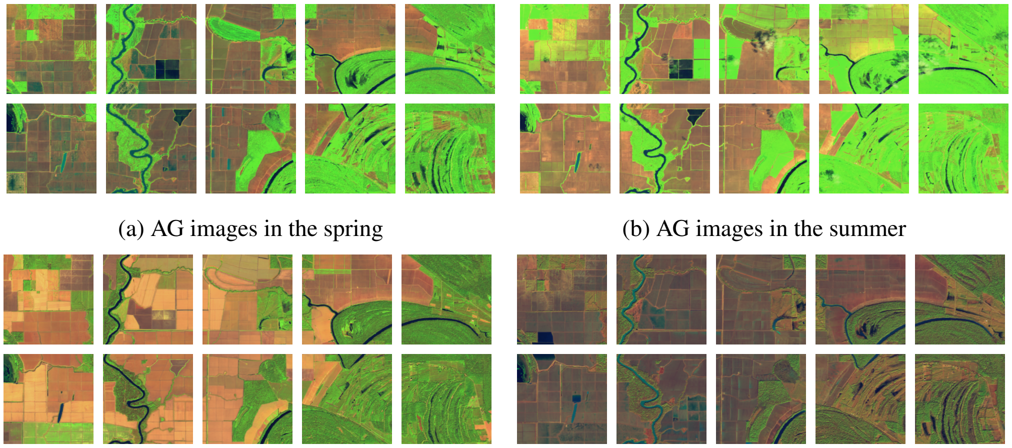
Sentinel-2 Mission. The Sentinel-2 mission Sentinel-2 (2023), launched in 2015, serves as an essential earth observation endeavor. With its 13 spectral bands and high revisit frequency of 5 days, the Sentinel-2 mission provides wide-swath, high-resolution, multi-spectral satellite images for a wide range of applications, such as climate change, agricultural monitoring, *etc*.

WRF-HRRR Model. The High-Resolution Rapid Refresh (HRRR) HRRR (2023) is a Weather Research & Forecasting Model (WRF)-based forecast modeling system, which hourly forecasts weather parameters for the whole United States continent with a spatial resolution of 3km. We take the HRRR assimilated results starting from July 2016 and archived in the University of Utah for use, which provides several crop growth-related parameters, *e.g.*, temperature, precipitation, wind speed, relative humidity, radiation, *etc*.

USDA. The United States Department of Agriculture (USDA) USDA (2023) provides annual crop information for major crops grown in the U.S., including corn, cotton, soybeans, wheat, *etc*., at the county level. The statistical data include the planted areas, the harvested areas, the production, and the yield for each type of crop, dated back to 1850 at the earliest.

B EXAMPLES OF OUR CROPNET DATASET

This section supplements Section 3.3 of our main paper by providing more examples regarding three modalities of data in our CropNet dataset. First, Figures 1 and 2 respectively illustrate Agricultural (AG) and Normalized Difference Vegetation Index (NDVI) images from Sentinel-2 Imagery under different seasons. Second, Figure 3 shows examples from the HRRR Computed Dataset, with the temperature in Spring, Summer, Fall, and Winter depicted respectively in Figures 3a, 3b, 3c, and 3d. Third, Figure 4 provides 2022 crop yield information from the USDA Crop Dataset, with the corn, cotton, soybeans, and winter wheat yields shown in Figures 4a, 4b, 4c, and 4d, respectively. Note that Our CropNet Dataset also includes crop production data for corn, cotton, soybeans, and winter wheat, with examples regarding their 2022 production information illustrated respectively in Figures 5a, 5b, 5c and 5d.



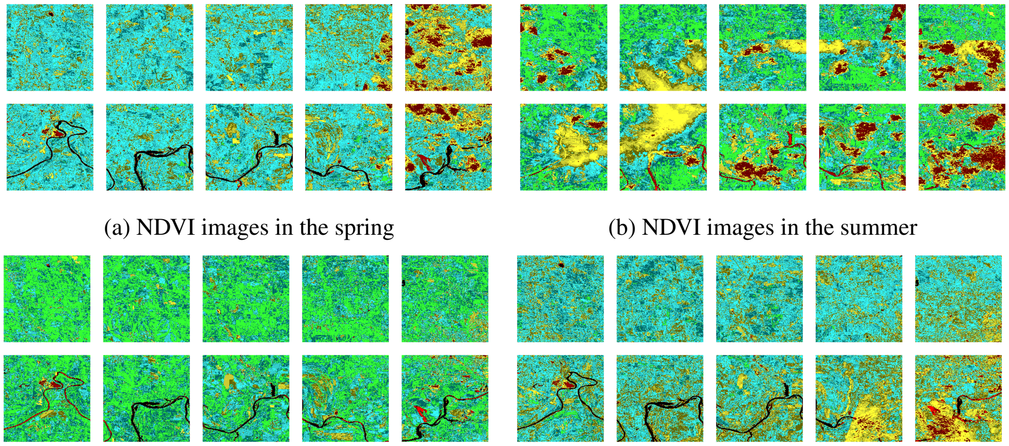
(a) AG images in the spring

(b) AG images in the summer

(c) AG images in the fall

(d) AG images in the winter

Figure 1: Examples of agricultural imagery (AG) from Sentinel-2 Imagery.



(a) NDVI images in the spring

(b) NDVI images in the summer

(c) NDVI images in the fall

(d) NDVI images in the winter

Figure 2: Examples of normalized difference vegetation index (NDVI) from Sentinel-2 Imagery.

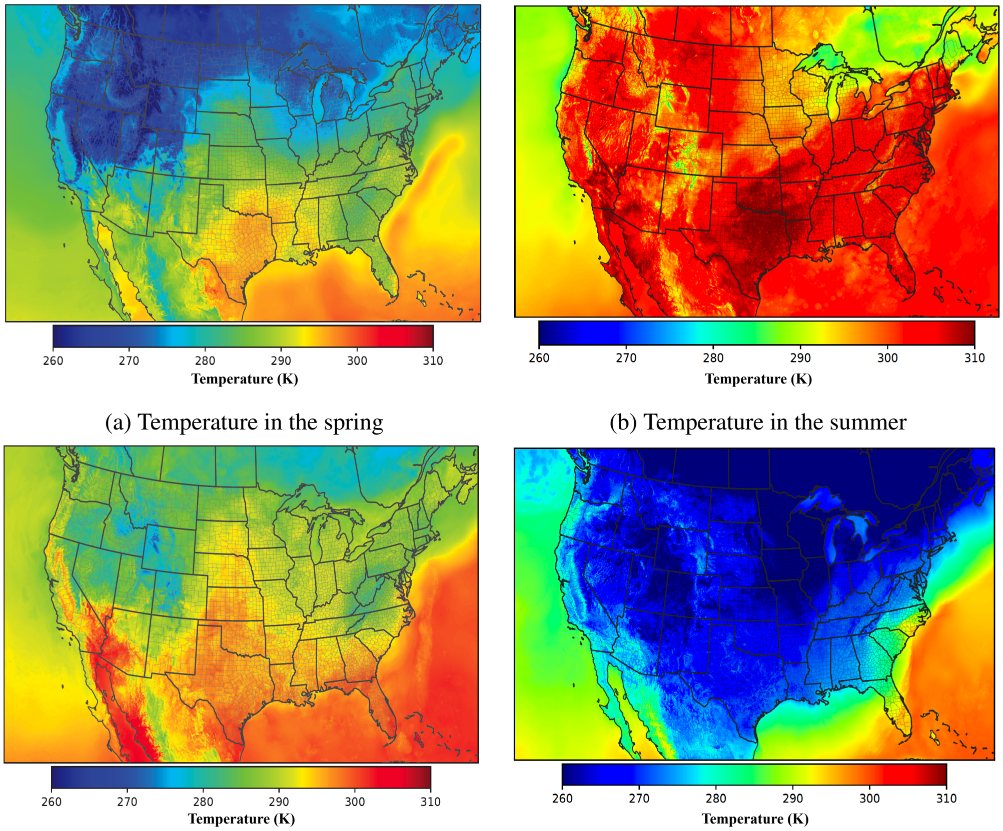
C Details of Data Collection

C.1 Significance of Our Cloud Coverage Setting and Revisit Frequency for Sentinel-2 Imagery

This section supplementss Section 3.3 of the main paper by demonstrating the necessity and importance of our cloud coverage setting (*i.e.*, $\leq 20\%$) and revisit frequency (*i.e.*, 14 days) for Sentinel-2 Imagery. Figures 6 and 7 present examples of Sentinel-2 Imagery under the original revisit frequency of 5 days with and without our cloud coverage setting, respectively. Figure 8 illustrates satellites images under our revisit frequency of 14 days and our cloud coverage setting (*i.e.*, $\leq 20\%$).

From Figure 6, we observed that the cloud coverage may significantly impair the quality of Sentinel-2 Imagery (see Figures 6b, 6d, and 6e). Worse still, the extreme cases of cloud coverage (refer to Figures 6d and 6e) degrade satellite images into noisy representations. This demonstrates the significance of our cloud coverage setting for discarding low-quality satellite images. Unfortunately, under the original sentinel-2 revisit frequency of 5 days, our cloud coverage setting would result in a large proportion of duplicate satellite images, *e.g.*, 50% (*i.e.*, 3 out of 6 satellite images) as depicted in Figure 7 [1]. This is because if the cloud coverage in our requested revisit day exceeds 20%, Processing API Sentinel-Hub (2023) will download the most recent available satellite images,

[1] Figures 7a and 7b depict identical satellite images. Likewise, Figures 7c, 7d, and 7e also display the same satellite images.



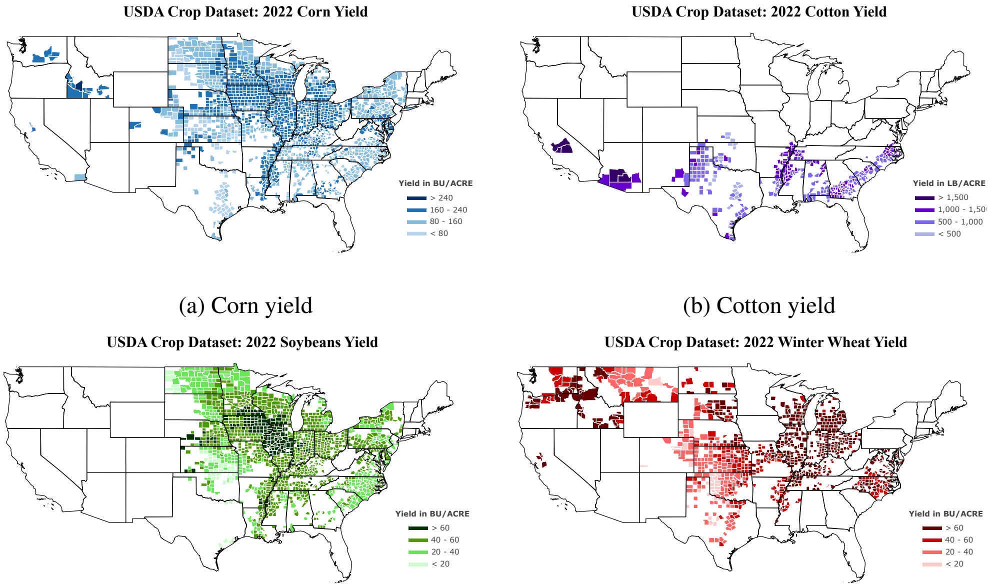
(a) Temperature in the spring

(b) Temperature in the summer

(c) Temperature in the fall

(d) Temperature in the winter

Figure 3: Illustration of the WRF-HRRR Computed Dataset for temperature in (a) the spring, (b) the summer, (c) the fall, and (d) the winter.



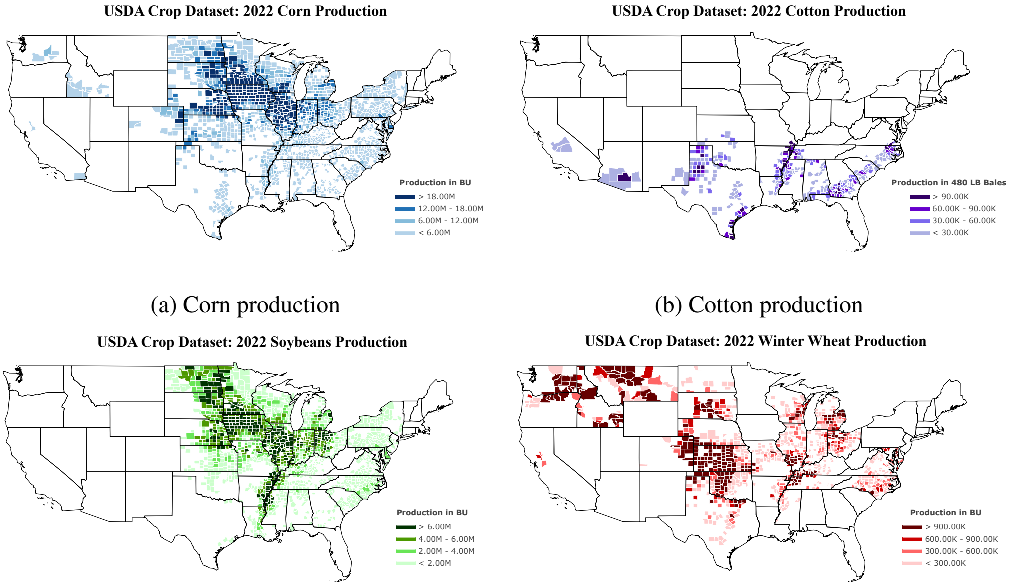
(a) Corn yield

(b) Cotton yield

(c) Soybeans yield

(d) Winter wheat yield

Figure 4: Illustration of USDA Crop Dataset for (a) 2022 corn yield, (b) 2022 cotton yield, (c) 2022 soybeans yield, and (d) 2022 winter wheat yield.



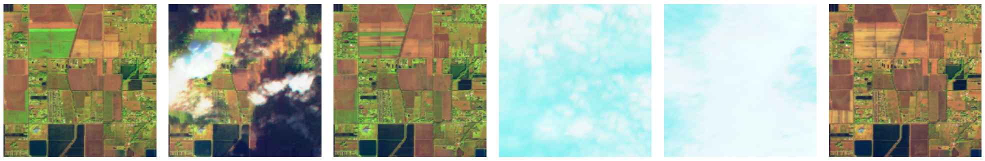
(a) Corn production

(b) Cotton production

(c) Soybeans production

(d) Winter wheat production

Figure 5: Illustration of USDA Crop Dataset for (a) 2022 corn production, (b) 2022 cotton production, (c) 2022 soybeans production, and (d) 2022 winter wheat production.



(a) 2022-12-01 Cloud: 0%

(b) 2022-12-06 Cloud: 35.8%

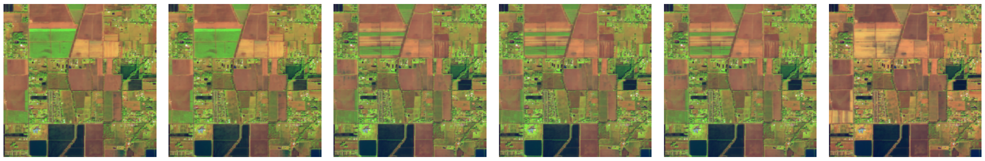
(c) 2022-12-11 Cloud: 0%

(d) 2022-12-16 Cloud: 97.2%

(e) 2022-12-21 Cloud: 100%

(f) 2022-12-26 Cloud: 2.7%

Figure 6: Examples of Sentinel-2 Imagery under the original revisit frequency of 5 days without our cloud coverage setting, with the revisit date and the cloud coverage listed below each image.



(a) 2022-12-01 Original

(b) 2022-12-06 Duplicate

(c) 2022-12-11 Original

(d) 2022-12-16 Duplicate

(e) 2022-12-21 Duplicate

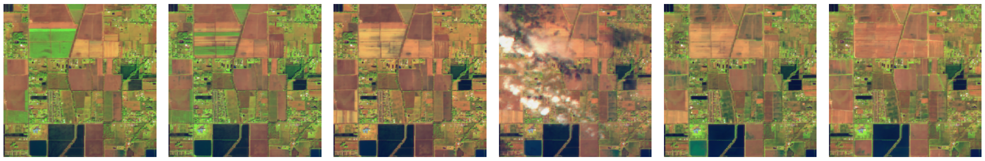
(f) 2022-12-26 Original

Figure 7: Examples of Sentinel-2 Imagery under the original revisit frequency of 5 days and our cloud coverage setting. The revisit date is listed below each image. "Duplicate" (or "Original") indicates whether the satellite image is duplicate (or not) under our cloud coverage setting.

whose cloud coverage satisfies our condition (*i.e.*, $\leq 20\%$). In sharp contrast, extending the revisit frequency from 5 days to 14 days markedly decreases the occurrence of duplicate satellite images. For example, there are no duplicate satellite images observed in Figure 8. Hence, our revisit frequency of 14 days for Sentinel-2 Imagery is necessary as it can significantly improve storage and training efficiency.

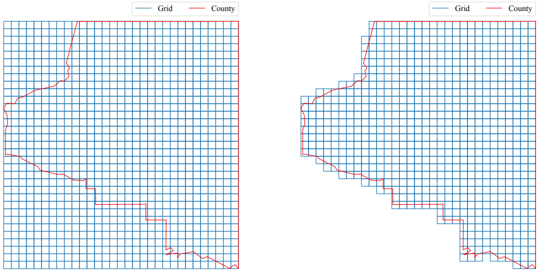
C.2 County Partitioning

In Section 3.3 of our main paper body, we have introduced partitioning one county into multiple high-spatial-resolution grids for precise agricultural tracking. Here, we provide the details for such a partition. A naive way to achieve this is to expand a county's geographic boundary to a rectangle area



(a) 2022-12-01 Original (b) 2022-12-15 Original (c) 2023-01-01 Original (d) 2023-01-15 Original (e) 2023-02-01 Original (f) 2023-02-15 Original

Figure 8: Examples of Sentinel-2 Imagery under our revisit frequency of 14 days and our cloud coverage setting, with the revisit date listed below each image. We would like to highlight that there are no duplicate satellite images observed.



(a) Naive solution (b) Our solution

Figure 9: Difference between the naive solution and our solution. (a) The naive solution leads to a significant number of grids falling outside the county's polygon. (b) By using our solution, the boundaries of grids (*i.e.*, the blue line) align perfectly with the county's boundary (*i.e.*, the red line).

Table 1: Details of WRF-HRRR Computed Dataset

Source	Parameters	Description
WRF-HRRR model	Averaged Temperature	2 metre averaged temperature during a day/month. Unit: K
	Precipitation	Total precipitation. Unit: $\mathrm{kg/m^2}$
	Relative Humidity	2 metre relative humidity. Unit: %
	Wind Gust	Wind gust on the ground. Unit: $\mathrm{m/s}$
	Wind Speed	Wind speed on the ground. Unit: $\mathrm{m/s}$
	Downward Shortwave Radiation Flux	The total amount of shortwave radiation that reaches the Earth's surface. Unit: $\mathrm{W/m^2}$
Computed by us	Maximal Temperature	2 metre maximal temperature during a day/month. Unit: K
	Minimal Temperature	2 metre minimal temperature during a day/month. Unit: K
	Vapor Pressure Deficit (VPD)	The amount of drying power the air has upon the plant. Unit: kPa

by using its maximal and minimal latitude and longitude, and then evenly divide such a rectangle area into multiple grids. Unfortunately, such a partition solution may result in a large number of grids outside the county polygon for some large counties (see Figure 9a). To handle this matter, we develop a novel solution by dropping the grids outside the county's boundary (see Figure 9b). Compared to the naive solution, our solution enjoys two advantages. First, it can significantly reduce the disk space storage size. Take Coconino County in Arizona for example, by employing our solution, its total number of grids degrades from 1023 to 729, which is 0.71x less than that from the naive solution. Second, our solution can evade the negative effect incurred by regions outside the county's boundary on crop yield predictions.

C.3 Details of WRF-HRRR Computed Dataset

Table 1 presents details of meteorological parameters in the WRF-HRRR Computed Dataset, where 6 weather parameters, *i.e.*, averaged temperature, precipitation, relative humidity, wind gust, wind speed, downward shortwave radiation flux, are obtained directly from the WRF-HRRR model, and 3 other parameters, *i.e.*, maximal temperature, minimal temperature, vapor pressure deficit (VPD), are calculated by ourselves. Notably, VPD describes the difference between the amount of moisture in the air and the maximum amount of moisture the air can hold at a specific temperature, which is an important concept in understanding the environmental conditions that affect plant growth and transpiration. Given two meteorological parameters, *i.e.*, the temperature measured in Kelvin T_K

and the relative humidity RH, VPD is calculated by the following equations:

$$
\begin{aligned}
T_C &= T_K - 273.15, \\
VP_{\text{sat}} &= \frac{610.7 \times 10^{(7.5 \times T_C)/(237.3 + T_C)}}{1000}, \\
VP_{\text{air}} &= VP_{\text{sat}} \times \frac{RH}{100}, \\
VPD &= VP_{\text{sat}} - VP_{\text{air}}.
\end{aligned}
\tag{1}
$$

```
{
    "FIPS":"01003",
    "year":2022,
    "county":"BALDWIN",
    "state":"AL",
    "county_ansi":"003",
    "state_ansi":"01",
    "data":{
        "HRRR":{
            "short_term":[
                "HRRR/data/2022/AL/HRRR_01_AL_2022-04.csv",
                "HRRR/data/2022/AL/HRRR_01_AL_2022-05.csv",
                "HRRR/data/2022/AL/HRRR_01_AL_2022-06.csv",
                "HRRR/data/2022/AL/HRRR_01_AL_2022-07.csv",
                "HRRR/data/2022/AL/HRRR_01_AL_2022-08.csv",
                "HRRR/data/2022/AL/HRRR_01_AL_2022-09.csv"
            ],
            "long_term":[
                [
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-01.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-02.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-03.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-04.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-05.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-06.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-07.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-08.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-09.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-10.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-11.csv",
                    "HRRR/data/2021/AL/HRRR_01_AL_2021-12.csv"
                ],
                # The remaining years are hidden for conserving space
                ...
            ]
        },
        "USDA":"USDA/data/Soybeans/2022/USDA_Soybeans_County_2022.csv",
        "sentinel":[
            "Sentinel/data/AG/2022/AL/Agriculture_01_AL_2022-04-01_2022
    -06-30.h5",
            "Sentinel/data/AG/2022/AL/Agriculture_01_AL_2022-07-01_2022
    -09-30.h5"
        ]
    }
}
```

Listing 1: Example of our JSON configuration file.

C.4 Spatial and Temporal Alignment of Our CropNet Dataset

Here, we present an example of our JSON configuration file (see Listing 1) for one U.S. county (*i.e.*, Baldwin in Alabama), to show how satellite images from Sentinel-2 Imagery, daily and monthly weather parameters from the WRF-HRRR Computed Dataset, and the crop information from USDA

Crop Dataset, are spatially and temporally aligned. As presented in Listing 1, “data.sentinel” and “data.HRRR.short_term” respectively represent satellite images and daily meteorological parameters during the crop growing season, “data.HRRR.long_term” indicates monthly weather conditions from previous 5 years, and “data.USDA” provides the crop information for the county. Meanwhile, “FIPS” and “year” respectively indicate the unique FIPS code and the year for the growing season, enabling us to obtain the data for our targeted county in a specific year. In summary, the JSON configuration file allows us to retrieve all three modalities of data over the time and region of interest.

D Supporting Experimental Settings and Results

D.1 Additional Experimental Setups

CropNet Data. Due to the limited computational resources, we are unable to conduct experiments across the entire United States. Consequently, we extract the data with respect to five U.S. states, *i.e.*, Illinois (IL), Iowa (IA), Louisiana (LA), Mississippi (MS), and New York (NY), to exhibit the applicability of our crafted CropNet dataset for county-level crop yield predictions. Specifically, two of these states (*i.e.*, IA and IL) serve as representatives of the Midwest region, two others (*i.e.*, LA and MS) represent the Southeastern region, and the fifth state (*i.e.*, NY) represents the Northeastern area. Four of the most popular crops are studied in this work, *i.e.*, corn, cotton, soybeans, and winter wheat. For each crop, we take the aligned Sentinel-2 Imagery and the daily data in the WRF-HRRR Computed Dataset during growing seasons in our CropNet dataset, respectively for precise agricultural tracking and for capturing the impact of growing season weather variations on crop growth. Meanwhile, the monthly meteorological parameters from the previous 5 years are utilized for monitoring and quantifying the influence of climate change on crop yields.

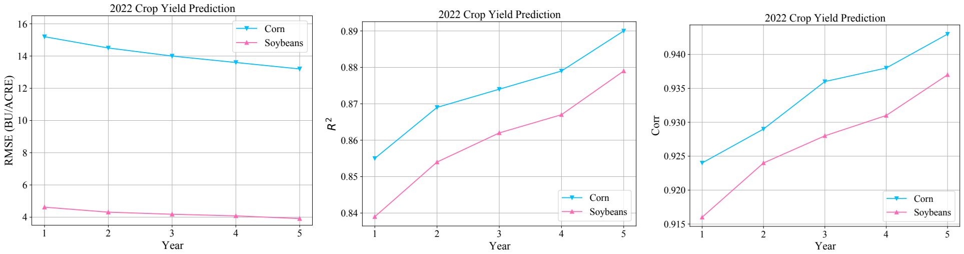
D.2 Significance of Each Modality of Our CropNet Dataset

To show the necessity and significance of each modality data in our CropNet dataset, we examine five scenarios. First, we drop the temporal satellite images (denoted as “w/o temporal images”) by randomly selecting only one day’s imagery data. Second, we discard the high-resolution satellite image (denoted as “w/o high-resolution images”) by using only one satellite image to capture the whole county’s agricultural information. Third, we ignore the effects of weather variations on crop yields by dropping all meteorological data, denoted as “w/o WRF-HRRR data”. Similarly, “w/o short-term data” and “w/o long-term data” represent masking out the daily and monthly meteorological parameters, respectively. We also include prediction results by using all modalities of the CropNet (denoted as “All”) for performance comparison. Note that the USDA Crop Dataset provides the label for crop yield predictions; hence, no ablation study requires.

Table 2 presents the experimental results under the MMST-ViT model Lin et al. (2023). We have four observations. First, discarding the temporal satellite images (*i.e.*,“w/o temporal images”) degrades performance significantly, raising the RMSE value by 8.9 (or 1.81) and lowering the Corr value by 0.073 (or 0.058) for corn (or soybeans) yield predictions. This is due to that a sequence of satellite images spanning the whole growing season are essential for tracking crop growth. Second, “w/o high-resolution images” achieves the worst prediction performance, with a largest RMSE vaue of 27.9 (or 7.8) and a lowest Corr value of 0.810 (or 0.794) for corn (or soybeans) yield predictions. The reason is that high-resolution satellite images are critical for precise agricultural tracking. Third, dropping meteorological parameters (*i.e.*, w/o WRF-HRRR data) makes MMST-ViT fail to capture meteorological effects on crop yields, leading to the increase of RMSE value by 7.4 (or 1.87) and the decease of Corr value by 0.072 (or 0.063) for predicting corn (or soybeans) yields. Fourth, discarding either daily weather parameters (*i.e.*,“w/o short-term data”) or monthly meteorological parameters (*i.e.*,“w/o long-term data”) lowers crop yield prediction performance. The reason is that the former is necessary for capturing growing season weather variations, while the latter is essential for monitoring long-term climate change effects. Hence, we conclude that each modality in our CropNet dataset is important and necessary for accurate crop yield predictions, especially for those crops which are sensitive to growing season weather variations and climate change.

Table 2: Ablation studies for different modalities of the CropNet dataset, with five scenarios considered and the last row presenting the results by using all modalities

Modality	Scenario	Corn			Soybeans		
		RMSE (↓)	R^2 (↑)	Corr (↑)	RMSE (↓)	R^2 (↑)	Corr (↑)
Sentinel-2 Imagery	w/o temporal images	22.1	0.758	0.870	5.72	0.773	0.879
	w/o high-resolution images	27.9	0.656	0.810	7.80	0.631	0.794
WRF-HRRR Computed Dataset	w/o WRF-HRRR data	20.6	0.758	0.871	5.78	0.764	0.874
	w/o short-term data	18.6	0.796	0.892	5.04	0.816	0.903
	w/o long-term data	15.3	0.854	0.924	4.72	0.825	0.908
All	—	13.2	0.890	0.943	3.91	0.879	0.937



(a) RMSE

(b) R^2

(c) Corr

Figure 10: Performance of 2022 corn and soybeans yield predictions by varying the length of long-term meteorological data from one year to five years.

D.3 Impact of Long-Term Meteorological Data

We further quantify the effects of long-term meteorological data by varying its length from one year to five years, for supporting Section 4 of the main paper. Figure 10 depicts the performance results for 2022 corn (*i.e.*, the blue line) and soybeans (*i.e.*, the pink line) yield predictions under the MMST-ViT model. We observed that lengthening the monthly meteorological data from one year to five years can improve the performance of crop yield predictions, lowering the RMSE value by 2.2 (or 0.7) and lifting R^2 and Corr values respectively by 0.035 (or 0.040) and by 0.019 (or 0.021), for corn (or soybeans) yield predictions. The reason is that a longer range of monthly meteorological parameters can better monitor and relate the climate change effect on crop growth.

References

HRRR. The high-resolution rapid refresh (hrrr), 2023. URL `https://home.chpc.utah.edu/~u0553130/Brian_Blaylock/cgi-bin/hrrr_download.cgi`.

Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, Drew Gholson, Nicolas Ashwell, Tri Setiyono, Brenda Tubana, Lu Peng, Magdy Bayoumi, and Nian-Feng Tzeng. Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *International Conference on Computer Vision (ICCV)*, 2023.

Sentinel-2. The copernicus sentinel-2 mission, 2023. URL `ttps://sentinel.esa.int/web/sentinel/missions/sentinel-2`.

Sentinel-Hub. Sentinel hub process api, 2023. URL `https://docs.sentinel-hub.com/api/latest/api/process/`.

USDA. The united states department of agriculture (usda), 2023. URL `https://quickstats.nass.usda.gov`.