*remote sensing*

MDPI

*Article*

# NOAH: A Multi-Modal and Sensor Fusion Dataset for Generative Modeling in Remote Sensing

**Abdul Mutakabbir [1], Chung-Horng Lung [2], Marzia Zaman [3], Darshana Upadhyay [4,5], Kshirasagar Naik [6], Koreen Millard [7,*], Thambirajah Ravichandran [8] and Richard Purcell [5]**

[1] Department of Data Science, Analytics, and Artificial Intelligence, Carleton University, Ottawa, ON K1S 5B6, Canada; mutakabbir@cmail.carleton.ca
[2] Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
[3] Research and Development, Cistel Technology, 30 Concourse Gate, Suite 200, Ottawa, ON K2E 7V7, Canada
[4] Research and Development, TeleAI Corporation, 43-30 Concourse Gate, Ottawa, ON K2E 7V7, Canada
[5] Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 4R2, Canada
[6] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[7] Department of Geography and Environmental Studies, Carleton University, Ottawa, ON K1S 5B6, Canada
[8] Research and Development, Hegyi Geomatics Inc., 102-30 Concourse Gate, Ottawa, ON K2E 7V7, Canada
* Correspondence: koreen_millard@carleton.ca

## Highlights

**What are the main findings?**

- This research produced a multimodal and sensor fusion dataset (NOAH) to be used by researchers for Generative Modeling of soft sensors to generate synthetic "Point -> Region" data.
- The research showed the ability of the FiLM+UNet model to generate synthetic data for TIR and Cirrus bands.

**What are the implications of the main findings?**

- Model-generated or synthetic data can be produced for hypothetical climate change scenarios by varying the ground-based sensor values.
- Using nearby point ground-based sensor data, missing data for a region can be produced

## Abstract

Earth Observation (EO) and Remote Sensing (RS) data are widely used in various fields, including weather, environment, and natural disaster modeling and prediction. EO and RS done through geostationary satellite constellations in fields such as these are limited to a smaller region, while sun synchronous satellite constellations have discontinuous spatial and temporal coverage. This limits the ability of EO and RS data for near-real-time weather, environment, and natural disaster applications. To address these limitations, we introduce **N**ow **O**bservation **A**ssemble **H**orizon (NOAH), a multi-modal, sensor fusion dataset that combines Ground-Based Sensors (GBS) of weather stations with topography, vegetation (land cover, biomass, and crown cover), and fuel types data from RS data sources. NOAH is collated using publicly available data from Environment and Climate Change Canada (ECCC), Spatialized CAnadian National Forest Inventory (SCANFI) and United States Geological Survey (USGS), which are well-maintained, documented, and reliable. Applications of the NOAH dataset include, but are not limited to, expanding RS data tiles, filling in missing data, and super-resolution of existing data sources. Additionally, Generative Artificial Intelligence (GenAI) or Generative Modeling (GM) can be applied for near-real-time model-generated or synthetic estimate data for disaster modeling in remote locations. This can complement the use of existing observations by field instruments, rather than replacing them. UNet backbone with Feature-wise Linear Modulation (FiLM) injection

Check for updates

Academic Editor: Eric Ariel L. Salas

Received: 30 December 2025
Revised: 28 January 2026
Accepted: 29 January 2026
Published: 1 February 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

of GBS data was used to demonstrate the initial proof-of-concept modeling in this research. This research also lists ideal characteristics for GM or GenAI datasets for RS. The code and a subset of the NOAH dataset (NOAH mini) are made open-sourced.

**Keywords:** soft sensors;generative AI; generative modeling; weather modeling; climate change modeling; sensor fusion; multimodal representation; multi-modal dataset; multi-modality images; multi-modality downstream tasks

## 1. Introduction

In recent decades, Earth Observation (EO) through satellite imagery has been used in various fields, including but not limited to monitoring [1,2], modeling [3], forecasting [4–6], and predicting [7,8] weather; detection, mitigation, modeling [9], monitoring [10], management [11] predicting and forecasting [12,13] natural disasters; monitoring land use [14]; and identification of environmental changes [15]. One of the most widely used applications of near-real-time EO with Remote Sensing (RS) is in wildfire detection [16,17], prediction [18–21], management [22], and mitigation [23].

Geostationary RS satellites orbit the Earth at a specified altitude and at the same speed as the rotation of the Earth, hence appear to be stationary. This leads them to capture only a limited Field of View (FoV) rendering them unsuitable for world-scale representation. Nevertheless, with higher operational and deployment costs, a constellation of geostationary satellites can be used at a country scale or continental scale. Further, geostationary satellites can record data at short intervals (10 to 15 min [24]). In most cases, if not all, the low temporal resolution comes at a cost of coarse spatial resolution [25] of 2 to 3 km [24]. SEVIRI/MSG is an example of such a satellite from Africa, which captures data every 15 min, but at a coarse spatial resolution of 3 km [25]. Considering, for example, wildfires, on average, for each large fire (area burned $> 2$ km$^2$) there are seven small fires (area burned $\leq 2$ km$^2$) [26]. A RS image of resolution of 3 km will cover 9 km$^2$ in a single pixel, which is 4.5 times greater than the size of the largest small fire (2 km$^2$) and may not be useful to model or monitor such small fires. Further, coarse resolution is not ideal for local weather monitoring or environmental changes.

RS data products from sun synchronous satellites, on the other hand, record data for the entire Earth as it rotates. They have a higher spatial resolution but a lower temporal resolution compared to geostationary satellites [27]. For example, Landsat 8, a sun synchronous satellite, records at a spatial resolution between 15 m to 100 m, but it takes 16 days for it to record the same point again.

Multispectral-thermal RS is impacted by cloud cover, which accounts for, on average, 68% of the Earth's surface each day [25]. Sensors such as Thermal InfraRed (TIR) cannot reliably penetrate dense clouds, leading to inaccurate or missing data in certain bands. Occlusion of the Earth's surface due to cloud cover results in missing data in spatio-temporal environmental data, which impacts downstream calculations of derived data products [28], such as Land Surface Temperature (LST) and Soil Moisture (SM). These measurements are important for modeling weather patterns and natural disasters. There has been RS research into filling in missing data since 2010 [27] and is through either physics [29–31] or Machine Learning (ML) [8,32,33] approach.

Most raw RS data products also require atmospheric correction or temporal and angular normalization. These corrections require Ground-Based Sensors (GBS) data to validate them based on the measurements on the ground (in-situ measurements). With GBS spread far across, it is not always easy to find multiple GBS in a single RS tile (unless

they are specifically designed for validation). Hence, when datasets are collated, the region covered by a tile should be large enough to account for having multiple GBS spread across it for later validation of data.

Given the challenges, most RS data products independently are not ideal for fine resolution monitoring of every region of the Earth in near-real-time. This limits their application for actively monitoring near-real-time weather, environment, and natural disasters at fine resolutions. GBS, such as weather stations, are not hindered by temporal resolution and cloud cover when recording ground data. However, they are limited by a much narrower (point) spatial coverage than geostationary satellites. This was illustrated in the case of a farmer relying on GBS's weather prediction, who lost all the crop due to frost at night, since weather prediction proved wrong at longer distances [34]. Hence, a combination of both RS and GBS data will be ideal.

To take advantage of and to mitigate the issues of both RS and GBS, we introduce a new dataset called **Now Observation Assemble Horizon** (NOAH).The objective of NOAH is to create a datataset that can be used to build a near-real-time soft sensors using GBS data, which provides current information ("now observation") of the ground, which can then be collated with less frequently changing RS and geospatial data to generate ("assemble") data for an expanded region ("horizon") instead of just a point location.

NOAH is a multi-modal (RS, geospatial, and GBS) dataset that combines information from multiple sensors (anemometers, barometers, thermometers, TIR, and Operational Land Imager (OLI)). The data is collated from fine spatial resolution imagery (30 m) of Landsat [35] and combined with fine resolution (30 m) derived RS data products from Spatialized CAnadian National Forest Inventory (SCANFI) [36–43] and weather GBS data from Environment and Climate Change Canada (ECCC) [44].

Data provided by SCANFI is available for countries in North America through their respective governing bodies. Further, the data provided by SCANFI are derived from Landsat, which is temporally harmonized spectral imagery along with tile-level regional models based on k-nearest neighbours and a random forest imputation method [37]. Since the data provided by SCANFI are derived from Landsat, the topographic, vegetation, and fuel-type information that they provide can be derived for the other parts of the world.

The data provided by ECCC is identical to that provided by the World Meteorological Organization (WMO), which is a specialized agency of the United Nations (UN) and provides open-sourced meteorological data for the world. Further details are given in Section 3.4. In case of regions outside of Canadian boundaries, researchers can substitute the data with the one for their region. The process would be similar as the data features will be similar to the one in NOAH but with a different distribution. Researchers can also apply transfer learning on models trained on NOAH dataset to the data from their respective region.

The NOAH dataset is the primary contribution of the research. The source to generate the dataset and a subset of the dataset have been made open-sourced. Further details are given in Section 3. NOAH dataset covers **8,742,469** $km^2$ of non-overlapping land areas in Canada at **815** distinct locations under **5 modalities (Topography, Vegetation, Fuel, Satellite Images, and GBS weather stations)** at **30 m** spatial resolution, where each modality covers **40,000** $km^2$. We expect prospective researchers to be able to utilize the NOAH dataset to leverage Generative Artificial Intelligence (GenAI) or Generative Modeling (GM) to train foundational models that can generate near-real-time RS data, fill in missing values of existing RS products, reduce the temporal resolution of existing data, or super-resolution/upsample existing data products at coarse resolution.

Soft sensors, also known as virtual sensors, are predictive models that estimate a value based on massive amounts of available data without measuring it directly with a physical

sensor [45]. With the NOAH dataset, researchers can build soft sensors utilizing GenAI or GM, which is capable of generating near-real-time model-generated or synthetic data for disaster modeling in remote locations to complement the use of existing observations by field instruments, rather than replacing them.

This research also provided benchmarking for future researchers to compare their results with. To maintain a consistent benchmark in research with varying computational resources, we evaluate the proof-of-concept models' results on NOAH mini since not all future researchers will have the computing capability to train on the entire NOAH dataset. A UNet + Feature-wise Linear Modulation (FiLM [46]) proof-of-concept architecture was used to benchmark the results.

## 2. Related Work

In this section, we first describe the requirements of the ideal dataset based on related work. Secondly, we assess the related work in terms of the multi-modal RS datasets currently available. Then, we analyze the literature on sensor fusion datasets. Finally, we compare our dataset (NOAH) with other datasets.

### *2.1. Ideal Dataset Requirements for Remote Sensing Soft Sensors with Generative Modeling*

The ideal requirements for RS soft sensors with GM require specific conditions and forms of data. In this subsection, we discuss our work in relation to the ideal dataset needs for GM for RS. A comparison of the ideal requirements with existing datasets is given later in Section 2.4.

This research identified the following points as ideal requirements for any RS soft sensors data source that can be used in GM.

1. *Non-Simulated Data*: Real-world data exhibit imperfections which are not easily reconstructed in simulation, making transfer learning from synthetic to real data a nontrivial task [47]. This, although stated for a 3D object, is true for RS data as well. The simulated data may provide more training data, yet there may be bias or missing imperfections in the data that is being simulated. When this data is used for training models or to apply transfer learning, the models trained will incorporate these inconsistencies. Hence, an ideal dataset for GM of RS should not have simulated data.
2. *Large Dataset Size*: The dataset should be significantly large, since deep learning approaches need more data for training. An improvement in prediction performance with increasing magnitude of the dataset size was found in [48]. Novel approaches and foundational model training especially need a large dataset size.
3. *Multi-Modal Dataset*: EO data obtained from RS satellites typically has missing data due to cloud cover. Cloud cover on average accounts for 68% of surface data each day [25]. Hence, data from different modalities will be needed to find suitable ways of filling in missing values. Further, validation of data collected from RS sources requires in-situ measurements from ground sensors.
4. *Sensor Fusion Datasets*: Multiple sensor types together give a holistic picture of the environment being measured. For example, in Landsat 8 and 9, TIR and OLI sensors are used to get 11 bands of data [35].
5. *Large Temporal Coverage*: More than 96% of the land cover of Canada remains unchanged over a period of 10 years [49]. Hence, to incorporate large-scale and small-scale changes in the dataset, it should have a temporal coverage of either a few years or decades.
6. *Large Coverage Region*: Let us consider that the dataset is for a specific forest/region with its distinct climate, soil and forestry. Then models trained on such a dataset will be biased to the region and may not be generalizable to other regions.

7. *Large Tile Area*: Data collected from RS sources are corrected and validated using GBS. When data is collated from multiple different types of GBS, which are controlled and operated by various organizations, the GBS distribution is uneven across regions. Hence, it would be ideal to have a large tile area to account for multiple unevenly distributed GBS covered in a single tile.
8. *Low Temporal Resolution*: The local changes, such as temperature, rain, wind speed, and wind direction, change more frequently, in a matter of minutes or hours. Hence, farmers relying on this data will need up-to-date information [34]. In case of an ongoing disaster, near-real-time information about the surroundings is vital. Therefore, lower temporal resolution is better.
9. *Fine Spatial Resolution*: There are 7× more small fires (area burned $\leq$ 2 km$^2$) than large fires [26]. Consider an RS satellite has a spatial resolution of 1 km, then it will cover an area of 1 km$^2$. This means that at max 2 pixels in the RS data product will have all of the information of the wildfire. This may not be enough information to model fire spread, as it will be at max 1 additional pixel. Similarly, for other natural disasters as well, it would be recommended to have fine spatial resolution. Additionally, if we have fine spatial resolution, we can build models that can be trained to super-resolution existing coarse-resolution data products.

### 2.2. Multi-Modal Remote Sensing Dataset

In the context of this research, modality refers to the sources from which data is collected, which can either be from RS or even non-RS data. When the modality is greater than 1, it is called multi-modal. For example, NOAH provides topographical data as elevation; GBS data as time series from weather stations; fuel data from forest inventory; land cover, crown cover, biomass from vegetation data; and multi-spectral satellite imagery having a modality of 5. Some available multi-modal datasets, such as [4,50,51], are available, but they are domain-specific (e.g., wildfires spread [50,51], and solar energy forecasting [4]). Sim2Real dataset [51] were built to forecast wildfire spread using GBS data, fuel information, vegetation, and topography. It is the most similar dataset that is closest to NOAH in terms of having a good spatial resolution and five modalities, but most of the data is simulated (1 thousand real images and 1 million simulated images). This may not be ideal for training simulation bias-free models applicable in real-world situations with respect to GM.

Data from multiple modalities has been used in several cases [27,52,53] and GM techniques were applied with fine spatial resolution data by converting the images into pixels to generate derived data products and fill in missing data. The dataset built and used is not provided, while [27] also reviews the currently used techniques and categorizes them. Often, the lack of access to data that is produced by researchers to benchmark leads other researchers to build their own benchmarking dataset, making the process redundant. The comparability and benchmarking challenge due to lack of access to data is a matter of concern [27].

Although there is research on multiple modalities of data for filling in missing values or for domain-specific purposes, the availability of such datasets as open-sourced is limited. When available, it is for a domain-specific task such as wildfire. Hence, we aim to provide a dataset that is first open-sourced and has applications in multiple areas, such as expanding existing data, filling in missing values, super-resolution, weather modeling, and disaster modeling.

### 2.3. Sensor Fusion Remote Sensing Dataset

In the context of this research, sensor fusion refers to the integration of data from multiple sensors in a satellite such as OLI and TIR. Datasets provided in some studies [7,54–60]

can be considered as only sensor fusion datasets since they use multiple bands/sensors from single or multiple satellites. These datasets are designed for a narrow domain-specific task, such as weather prediction [7], fire segmentation [55–57], or cloud removal [60]. They are also limited by the missing information due to cloud cover or temporal gap, which cannot be supplemented from other modalities. Mesogeos [50] and Sim2Real [51] are also sensor fusion datasets but have other modalities of data, hence they are considered as multi-modal sensor fusion datasets.

### *2.4. Comparison of Datasets*

In Table 1, we provide a detailed tabulation of the existing datasets in comparison to NOAH (our dataset) concerning the ideal characteristics mentioned earlier in Section 2.1. In terms of the regions covered in the dataset (column 2 of Table 1), having "Worldwide" coverage is ideal, while a "simulated worldwide" coverage is not ideal. Although our dataset (NOAH) does not provide the best coverage region, it still has a large geographical coverage. In terms of tile area and total area coverage, the value may be as high as possible.

**Table 1.** Tabulation of Datasets in Relation to Ideal GenAI Remote Sensing Dataset with NOAH.

| Name | Coverage Region | Coverage | | | Resolution | | Images | Modality | Domain |
|---|---|---|---|---|---|---|---|---|---|
| | | Tile Area ($km^2$) | Total Area ($km^2$) | Temporal Period | Spatial | Temporal | | | |
| AllClear [60] | Worldwide | 6.5 | 155.5 K | 2022 | Fine (10 m) | NA | 4.3 M | Satellite | Cloud Removal |
| MODIS Thermal Anomaly [55–57] | Worldwide | 1.4 M | 149 M | 2000–2023 | Medium (1 km) | 5 min, 1 day, 8 days | 40 K | Satellite | Fire Segmentation |
| SolarCube [4] | Worldwide | 360 K | 6.84 M | 2018 | Coarse (5 km) | 15 min | 35 K | Derived Data Products, GBS | Solar Energy Forecasting |
| SAMRS [58] | Worldwide | 1–16.7 | ~400 K | NA | NA | NA | 105 K | Satellite (Up-sampled) | Segmentation |
| WorldStrat [54] | Worldwide | 2.5 | 10 K | 2017–2022 | Fine (1.5–60 m) | NA | 3449 | Satellite (Up-sampled) | Super-Resolution |
| ClimSim [59] | Worldwide Simulated | NA | NA | NA | NA | NA | NA | Satellite | Climate Simulation |
| ENS-10 [7] | Worldwide Simulated | NA | NA | 1998–2017 | Coarse (1 km) | 3.5 days | NA | Satellite | Weather Prediction |
| Sim2Real [51] | Worldwide Simulated | 130 | NA | 2013–2023 | Fine (30 m) | 1 h | 1 K (real), 1 M (simulated) | Topography, Vegetation, Satellite, Fuel, and GBS | Wildfire Spread |
| Mesogeos [50] | Mediterranean | 4096 | 9 M | 2006–2022 | Coarse (1 km) | 1 day | 25 K | Vegetation, Climate, and Anthropogenic | Wildfire Spread |
| **NOAH (Our)** | Canada (Expandable to Worldwide) | 40,000 | 8,742,469 | 2013–2025 | Fine (30 m) | 1 h | 234,089 | Topography, Vegetation, Satellite, Fuel, and GBS | Soft Sensors, Super-Resolution, Cloud Removal, GM, etc. |

Each row represents a unique dataset, and the columns specify the characteristics of the dataset. Cells in green represent the ideal characteristic for a specific dataset, while the ones in red represent the non-ideal conditions. The cells in white indicate a neutral characteristic (neither ideal nor non-ideal). It should be noted that the colored representation used in this table is a generic and subjective definition only within the specific context of this research.

In terms of tile area coverage (column 3) and total area coverage (column 4), the higher the value, the better. For both, MODIS Thermal Anomaly has the best value. It should also be noted that these two values, considered independently, do not provide a complete picture. They should be considered along with spatial resolution (column 6). For example, SolarCube has a very large tile area of 360,000 $km^2$, but the spatial resolution is 5 km. This means that the images are of size 120 × 120. When we consider this in relation to NOAH (our dataset), it has a tile area of 40,000 $km^2$, but a spatial resolution of 30 m. This means the images in NOAH have a size of 6667 × 6667. This puts NOAH in second place, after MODIS

Thermal Anomaly for largest spatial coverage. Further, it should be noted that although NOAH only covers Canada, it still has a significantly higher total area covered compared with AllClear, WorldStrat, SAMRS, and SolarCube, which are all considered "Worldwide".

In terms of temporal coverage (column 5) as stated in the ideal requirements, the higher the better. MODIS Thermal Anomaly has the most extensive temporal coverage of over 2 decades. ENS-10 has the second largest coverage, followed by Mesogeos and our dataset (NOAH). It should be noted that datasets apart from NOAH have a fixed temporal coverage. In the case of NOAH, the dataset can be expanded to get more recent information by running the open-sourced code provided. This gives NOAH a greater edge as the data is expanding. Further, the data sources are carefully chosen such that they will reliably produce data in the future (discussed in detail later in Section 3).

NOAH, Sim2Real, WorldStrat, and AllClear have the fine spatial resolution as mentioned in column 6 which was identified as an ideal characteristic. SolarCube has the best temporal resolution (column 7) of 15 min. NOAH (our dataset) and Sim2Real also have an ideal temporal resolution of 1 h. MODIS Thermal Anomaly has a varying temporal resolution in the different versions of the data, hence was not considered ideal.

In terms of the number of images (column 8), AllClear has the highest number of images (4.3 million). Although Sim2Real has 1 million simulated images, it only has 1 thousand real images. Hence, NOAH (our dataset) has the second-highest number of real images (234,089). It should be noted here that, as specified above, the NOAH dataset can expand to a larger size as new data is created by data sources.

In column 9, we see the different modalities of the datasets under consideration. Here NOAH (our dataset) and Sim2Real have 5 modalities which is the highest. Most have single modality such as, ClimSim, ENS-10, MODIS Thermal Anomaly, AllClear, WorldStrat, and SAMRS. Similarly, in column 10, where the domains are discussed, all have a single domain of application, such as wildfire for Mesogeos and Sim2Real. Only NOAH has multiple purposes, with applications in GM. The applications of NOAH are later discussed in Section 5.

To the best of our knowledge, NOAH is the only dataset that can serve multiple purposes, is expandable as data is generated, contains fine spatial resolution, and has an extensive coverage for tile area for validation of the physical accuracy of temperature or energy balance. Datasets like ENS-10, ClimSim, and MODIS Thermal Anomaly only contain satellite imagery modality at coarse resolution showing a clear inferiority to the NOAH dataset. In comparison to Mesogeos, which has three modalities (Vegetation, Climate, and Anthropogenic), NOAH contains five modalities (Topography, Vegetation, Satellite, Fuel, and GBS). The modalities in Mesogeos are specifically selected for wildfire spread and do not contain satellite imagery, which is essential for synthetic surface temperature data generation using generative modeling. Details on how NOAH can be used in such an application are presented in Sections 4.3 and 5. Additionally, Mesogeos contains coarse resolution at 1 km, which has limited application in generating synthetic super-resolution data. In comparison to Sim2Real, which also has five modalities like NOAH, but note that Sim2Real only contains 1 K real images while NOAH has 234,089 real images (which can be further extended as more data is generated). Further, the coverage of the tile area in the Sim2Real dataset is limited to only 130 $km^2$, which is not suitable for validation of the physical accuracy of temperature or energy balance as elaborated later in Section 3.1 compared to 40,000 $km^2$ tile area coverage of NOAH.

When taking a holistic picture of all the datasets considered, NOAH is by far the only dataset that meets all the ideal requirements. As we mentioned earlier, NOAH has a fine spatial resolution, low temporal resolution of 1 h, covers large tile area of 40,000 $km^2$ and total covering 8,742,469 $km^2$ of area. The data in NOAH can be expanded as the data

sources collect more information. Hence, the temporal coverage and number of images will increase as time goes on. The code to generate and expand the NOAH dataset has been made open-sourced.

## 3. Now Observation Assemble Horizon (NOAH) Dataset

NOAH contains real-world RS and GBS data collected from Canadian governmental organizations (ECCC [44], SCANFI [36–43]) with Landsat 8 imagery [35] jointly made available by the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS). The dataset is collected from 815 locations covering 8,742,469 $km^2$ non-overlapping area. Each location contains multi-modal (topography, vegetation, fuel, satellite, and GBS weather) data integrated from multiple sensors, totaling 234,089 satellite data points over the years 2013 to 2025. An example of one location in NOAH is given in Figure 1.

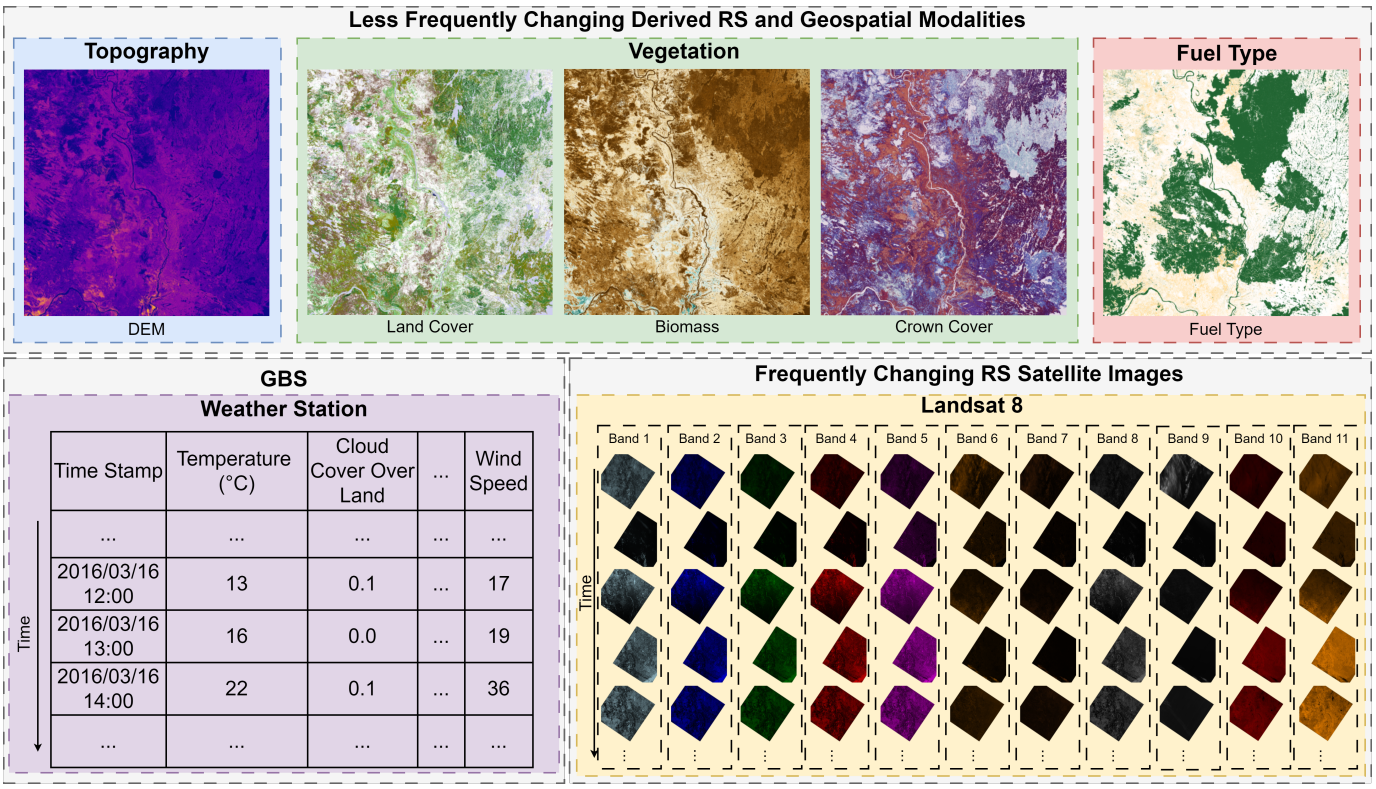



**Figure 1.** Available data for a sample location in NOAH. The topography modality is colored in blue, the vegetation modality is colored in green, the fuel type modality is colored in red, the GBS weather station modality is colored in purple, and the satellite image modality is colored in yellow.

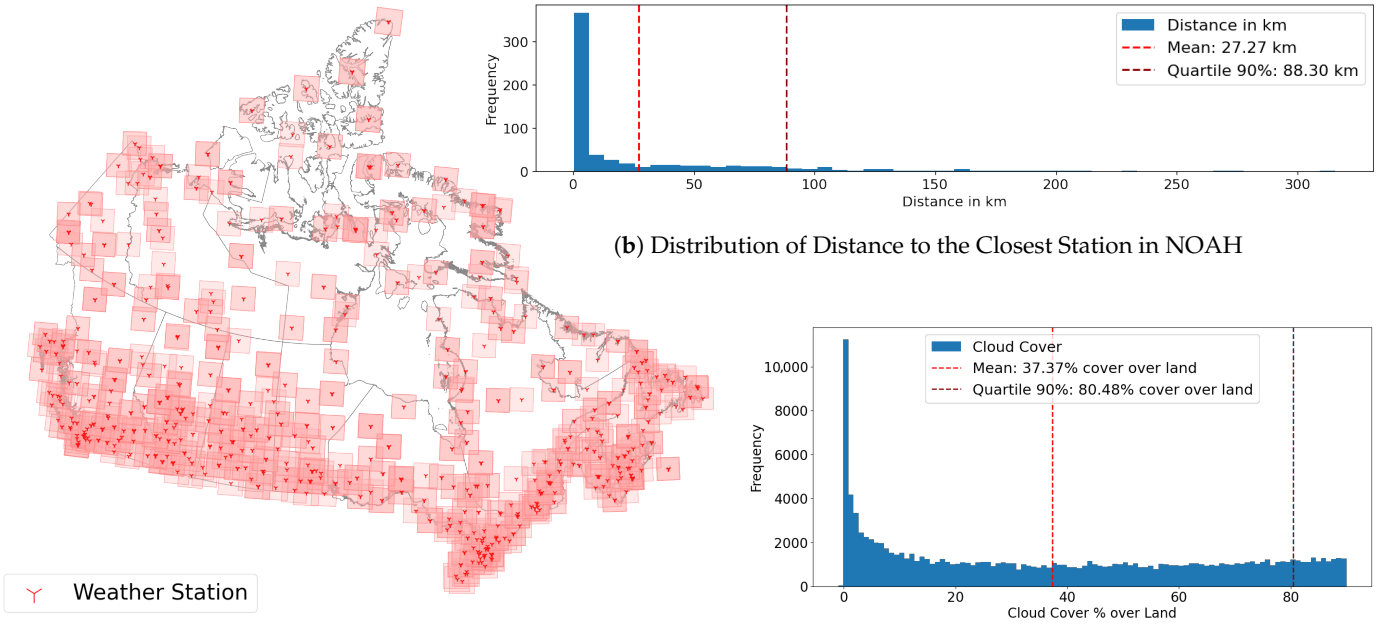
NOAH is an expanding dataset, i.e., the data in NOAH will grow and expand as the data sources provide more information. The new information can be collated automatically using the code to build the dataset. Hence, the code to generate and expand the NOAH dataset has been made open-sourced.

The size of the NOAH dataset exceeds 20 TB. Therefore, it cannot be made completely available to the public in raw format. To facilitate researchers with limited computational resources and to have a consistent benchmark across future research, the NOAH mini dataset was created. NOAH mini dataset is a smaller sample of the complete NOAH dataset and is made publicly available on Hugging Face. NOAH mini contains 10 locations with 10 satellite images for each area. Details on NOAH mini are provided later in Section 3.5.

### *3.1. Region of Interest*

We chose Canada as our regional coverage primarily because there are limited large collated datasets available specifically for Canada that take into account its diverse vegetation. Secondly, RS and geospatial datasets are being produced by SCANFI, which will be updated regularly over 5-year periods, providing reliable, standard, benchmarking sources to produce future extensions of the data.

We considered **815 distinct locations** as seen in Figure 2a, since they have GBS weather stations' recording data **every 1 h**, operated by ECCC. The region covered by an image in NOAH is highlighted with red boxes in Figure 2a. The total non-overlapping area covered by the red boxes accounted for **8,742,469** km$^2$. This is significantly larger than most available datasets [4,58,60].




(**a**) NOAH Spatial Coverage

(**b**) Distribution of Distance to the Closest Station in NOAH

(**c**) Distribution of Cloud Cover Over Land

**Figure 2.** Distribution of Spatial Data, Weather Stations and Cloud Cover Over Land in NOAH.

We wanted to choose a region of interest for each location that would include other ECCC weather stations, so that model-generated or synthetic estimate data can be validated with other in-situ measurements. A histogram of distance to the closest station from a given station is provided in Figure 2b with the mean and 90% quartile labelled. The minimum distance to the closest station for 90% of the GBS weather stations was found to be 88.3 km. Hence, in the NOAH dataset, 100 km distance in each direction (north, south, east, and west) around a given GBS weather station was chosen (i.e., a side of 200 km box), accounting for a total **coverage of 40,000** km$^2$ (200 km $\times$ 200 km) per image was chosen at **30 m resolution**. Having a tile area coverage of 200 km $\times$ 200 km ensures that there is $\geq$90% chance of one additional weather station being present within the tile area, apart from the one in the center of the tile. The purpose of the selection is for validation of the physical accuracy of temperature or energy balance based upon the descriptive statistics provided in Figure 2b, assuming the distribution holds.

It should be noted that not all models may be able to accurately generate 40,000 km$^2$ coverage, but they will at least have additional GBS point source data within the coverage to validate the accuracy of the model's results. Additionally, the use of fine spatial resolution (30 m) in NOAH, along with the large tile area, will allow for training of super-resolution models by downscaling images to coarse resolution. For example, the 6667 $\times$ 6667 images at 30 m spatial resolution per pixel in NOAH can be downscaled to 666 $\times$ 666 images at

300 m spatial resolution per pixel to train models to upscale a 300 m spatial resolution image to 30 m spatial resolution.

Distribution of the cloud cover in the satellite modality of NOAH is given in Figure 2c. 90% of the satellite images in NOAH have less than 88.48% cloud cover over land and on average 37.37%. When RS data is derived from different sources, the Coordinate Reference System (CRS) may be different. Care was taken to have the data converted to the same CRS across all modalities in NOAH.

### 3.2. Modalities of Data

NOAH has 5 modalities (topography, vegetation, fuel types, satellite imagery, and GBS weather) which are highlighted in Figure 1. A description of each of the modalities is given in the following subsections. It should be noted that in the NOAH dataset, due to the point-scale nature of the GBS (weather stations) data modality, fine-scale spatial variability is largely driven by frequently and less frequently changing RS and geospatial modalities, rather than resolved atmospheric gradients.

#### 3.2.1. Topography

Topography is a type of less frequently changing geospatial modality. It contains the elevation of the terrain at any given location in the image. NOAH contains the topographical information derived from the Digital Elevation Model (DEM), which was acquired from SCANFI [37] at 30 m resolution. Other RS topographical information, such as slope and aspect, can be derived from DEM. Since the terrain data in NOAH does not undergo any upsampling, it does not contain any biases from the sampling techniques as may occur in other datasets. A sample is highlighted in blue in Figure 1. There is one DEM image for each of the 815 locations in Figure 2a.

#### 3.2.2. Vegetation

Vegetation data in NOAH consists of land cover, biomass and crown cover derived from SCANFI [36,37,39–43] at 30 m resolution. These datasets are also an example of less frequently changing RS data. Only 8.7% of the land cover changed between 2015 and 2020 in SCANFI [49]. In NOAH, vegetation modality contains three types of information: land cover, biomass, and crown cover, which together should accurately represent the vegetation on the ground. A sample of vegetation data is highlighted in green in Figure 1 with appropriate labels for land cover, biomass and crown cover. There are three images (land cover, biomass, and crown cover) for each of the 815 locations in Figure 2a.

#### 3.2.3. Fuel Types

Fuel types specify the categorization based on combustible materials present on the ground. Fuel types are also acquired from SCANFI [37,38] at 30 m resolution. Fuel type modality may be beneficial when dealing with wildfires and post-fire vegetation growth. A sample of fuel types is highlighted in red in Figure 2a. There is one fuel type image for each of the 815 locations in Figure 2a.

#### 3.2.4. Satellite Imagery

Satellite imagery in NOAH was acquired from Landsat 8. A total of 234,089 Landsat 8 images [35] were acquired at 30 m resolution from 2013 to 2025 for a region covering 40,000 $km^2$ centered on top of a known GBS weather station from 815 distinct locations in Figure 2a. Each image has 11 bands of information coming from different sensors in the Landsat 8 satellite. The Landsat 8 images were extracted from Google Earth Engine (GEE) for Landsat 8 Collection 2, Tier 1, Level-1 (LANDSAT/LC08/C02/T1). Since these images are already scaled for reflectance and temperature, no conversions or imputations

were performed for the band values in the satellite imagery modality. A sample of Landsat 8 images centered on a known GBS over time is highlighted in yellow in Figure 1. It should be noted that although the images in Figure 1 are colored based on their bands or metadata, in NOAH, they will be available as a single channel without color. The number of images for each of the 815 locations in Figure 2a varies based on the Landsat 8 coverage over the location. Each satellite image in the NOAH dataset is linked to its respective topography, vegetation, fuel types, and GBS modalities data by the unique location. Hence, all the satellite images from a specific location of the 815 possible locations will have the same topography, vegetation, and fuel types data associated with it for the specific location, as the derived RS and geospatial data modalities change less frequently.

#### 3.2.5. Ground-Based Sensors

In NOAH, the weather station data provided by ECCC [44] was used as its Ground-Based Sensor (GBS). The weather station data are recorded at 1-hour intervals from 1840 to 2025 with varying operation coverage based on weather stations. For NOAH, only 815 stations that were in operation between 2013 to 2025 and providing hourly readings were taken. The locations are given in Figure 2a as a red marker. Each weather station records six features: wind direction; wind speed; pressure; dew point; temperature; and humidity. A sample of the reading for a station is highlighted in purple in Figure 1. 24 h of historical information is incorporated from these weather stations into a single data point in NOAH, along with additional metadata such as location, cloud cover, and elevation.

It should be noted that the GBS data recorded at a 1-h interval is for a point source only, while the other modalities (topography, vegetation, fuel types, and satellite imagery) are not point sources. Each GBS reading in the NOAH dataset is linked to its respective topography, vegetation, fuel types, and satellite imagery modalities by the unique location of the weather station. Similar to the satellite imagery modality, all point source data readings for a unique GBS weather station location will be linked to topography, vegetation, fuel types, and satellite imagery data for that unique location of the respective weather station.

### *3.3. Dataset Collation Workflow*

The workflow of the creation of the NOAH dataset is given in Figure 3. This workflow can be used to further extend the NOAH dataset as new data is provided by the data sources for the 5 modalities increases. The code for creating and expanding the NOAH dataset has been made publicly available.

First, we identified GBS and RS data sources for the modalities. Then, metadata from GBS and RS data sources were collected to understand the data and their distribution. Then, for each GBS and RS source, the data within the region of interest were downloaded from the respective GBS and RS data providers.

Once the data were downloaded from GBS and RS sources, preprocessing needed to be conducted. First, the RS satellite imagery modality was centered to its respective GBS point source location for each tile. Then each satellite imagery modality tile was cropped to a square area of 40,000 $km^2$ around the GBS point source location. CRS was made consistent across all the RS product tiles to *EPSG:3979* since it is suitable for a wide east–west region like Canada. Then, it was ensured that topography, vegetation, fuel types, and satellite imagery modalities were at 30 m spatial resolution. Finally, missing data for tiles were identified, such as cloud cover and no-data regions in the tile. A special mask value was created for the missing data, and the mask value was used to fill in the missing data.

Once the preprocessing was done, the GBS and RS products were collated into a single NOAH dataset. The data from different modalities were collated together by the unique location of the center of the RS or geospatial image, and the location of the weather station.

For each of the 815 unique locations, there is one of each less frequently changing derived RS and geospatial modality (topography, vegetation, and fuel type) image centered to the point location of the GBS, multiple frequently changing RS satellite imagery centered to the point location of the GBS, and the hourly point source readings of the GBS as illustrated in Figure 1.

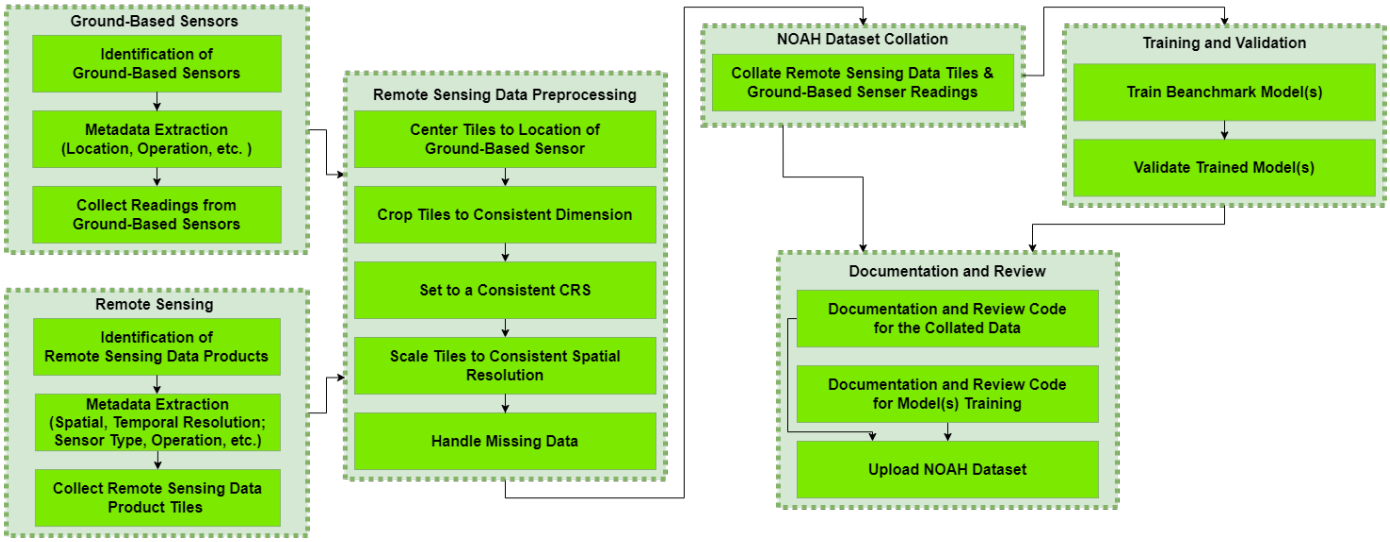



**Figure 3.** Workflow for the Collation of NOAH Dataset.

The NOAH dataset was then used to train and validate the models. Both its dataset generation code and model training code were validated and documented. Finally, the NOAH dataset was uploaded. It should be noted that the collated dataset size was >20 TB, hence only a subset of the dataset (NOAH mini) has been made public. However, the entire code to collate NOAH and train the models is made publicly available.

### *3.4. Future Dataset Growth Potential*

All the collated data was open-source data. The sources of data used to collate the data in NOAH are from standard Canadian and United States governmental agencies. These organizations will continue to record data over time maintaining their specified standards. These standards were incorporated while collating data in NOAH. This gives future growth potential to the NOAH dataset. Further, only Landsat 8 was used; the data can be extended to Landsat 9 as both use the same sensors and spatial resolution to capture RS images.

ECCC is the official Canadian meteorological authority providing Canadian weather observations to the WMO, which is a specialized agency of the UN. Hence, the GBS data from ECCC used in this study is similar to the data collected by the WMO, which records data for the entire world. The weather data collected by WMO is open-sourced and reliable.

### *3.5. NOAH Mini*

Each extracted Landsat 8 image at 40,000 $km^2$ is larger than 100 MB. With 234K+ images extracted in NOAH, the total size exceeded well over 20 TB. Therefore, not all researchers will have the computational resources to train on the complete NOAH dataset. Hence, a smaller version of NOAH, i.e., NOAH mini, was made available on Hugging Face for review and benchmarking. The dataset is documented with Croissant [61] and GeoCroissant metadata. Benchmarking on NOAH mini will ensure consistency irrespective of computational needs. Researchers who wish to train on the complete NOAH dataset can use the open-sourced code provided to generate and expand the complete NOAH dataset.

NOAH mini satellite images cover 40 $km^2$ instead of 40,000 $km^2$ as in NOAH. Further, only 10 distinct locations of GBS were chosen for NOAH mini instead of 815 locations

for NOAH. The GBS locations were chosen at random. Care was taken during random sampling to not include weather station locations with overlapping spatial coverage to ensure that the locations were far apart from each other. For each of the 10 GBS locations, 10 satellite imagery modalities were randomly sampled at the GBS weather station locations where there was minimal or no cloud cover. The Topography, Vegetation, and Fuel Types modality data for NOAH mini dataset were collected by centering a 40 $km^2$ square area on the 10 randomly sampled GBS locations.

The train, test, and validation split of the NOAH mini dataset was done based on time. For each location, the first six time steps were put in the training set, the next two in testing and the last two in validation. Since different time steps are taken in train, test, and validation sets, models trained to generate synthetic data for individual bands in the satellite imagery modality for bands, such as B9 (Cirrus), B10 (TIR), and B11 (TIR) should not be overfit as their values will vary significantly over time.

A sample of the satellite imagery modality comparing the NOAH dataset with the NOAH mini dataset is given in Figure 4. Each row represents a random location of the GBS, and each column is a unique band in Landsat 8. The red-highlighted region in each cell represents the NOAH mini dataset tile of tile area 40 $km^2$, while the entire cell represents the NOAH dataset tile of tile area 40,000 $km^2$.

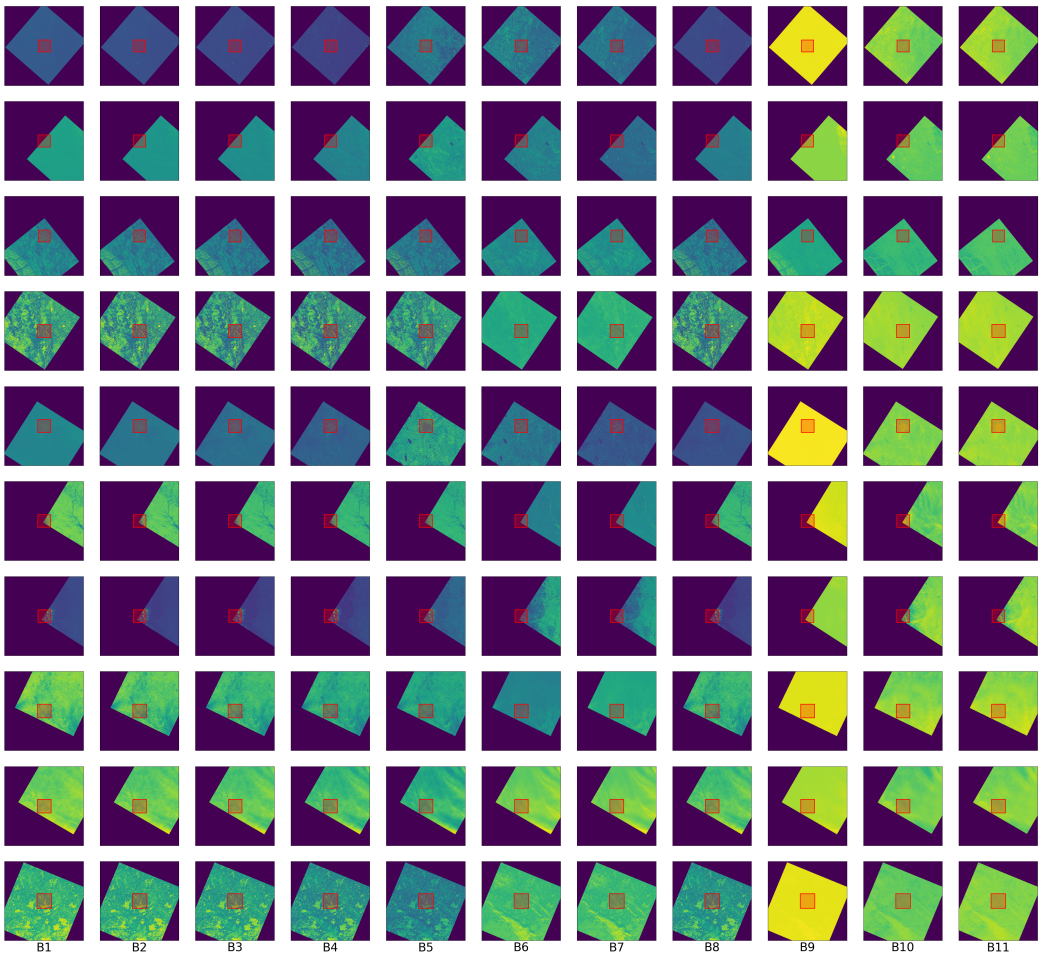



**Figure 4.** Comparison of one sample from each of the 10 distinct randomly selected locations in NOAH mini dataset for all eleven bands in satellite imagery modality with the NOAH dataset. Each row is a distinct location, while each column is a distinct band in the satellite imagery modality for Landsat 8. Each cell is a sample of the NOAH dataset, while the red region highlighted in the cell is a sample from the NOAH mini dataset.

To validate the physical accuracy, the NOAH dataset is ideal due to its large tile area coverage of 40,000 km$^2$ as seen in Figure 4. The large coverage area allows multiple in-situ GBS measurements from the weather station present within the tile area, as depicted in Figure 2b to validate the physical accuracy. Since the tile area of the NOAH mini dataset is significantly smaller than that of the NOAH dataset, and the GBS locations are widely spaced, validating the physical accuracy of temperature or energy balance on the NOAH mini dataset is not recommended. Further, researchers and organizations with large computation resources should use the complete NOAH dataset to train models and use the NOAH mini's validation set to report their benchmarking results. We recommend that researchers and organizations with limited computation resources train, test, and validate the models using the respective training, testing, and validation sets provided in the NOAH mini dataset to start with.

## 4. Experimentation

This section describes the experimentation and benchmark results for the GenAI models trained in the NOAH mini dataset. The NOAH mini dataset was used to maintain consistency with future benchmarks. NOAH mini was benchmarked on its ability to generate all 11 bands of Landsat 8 data tiles. The benchmark models were trained on two *RTX 3090* with 24 GB VRAM in each. Additionally, 125 GB RAM is needed to optimally perform the data preprocessing.

### *4.1. Baseline Modeling and Hyperparameter*

A 5-channel UNet model was used to extract the feature map from the less frequently changing RS and geospatial data products (topography, land cover, biomass, crown cover, and fuel types). Further, FiLM was used to inject 24-h historic weather GBS data from the time of creation of the respective Landsat 8 tile being generated by modulating the intermediate feature maps in the UNet convolution encoder block. Finally, the feature map with FiLM injection underwent transpose convolution in the generator block to get the generated band data. A visual insight is given in Figure 5.

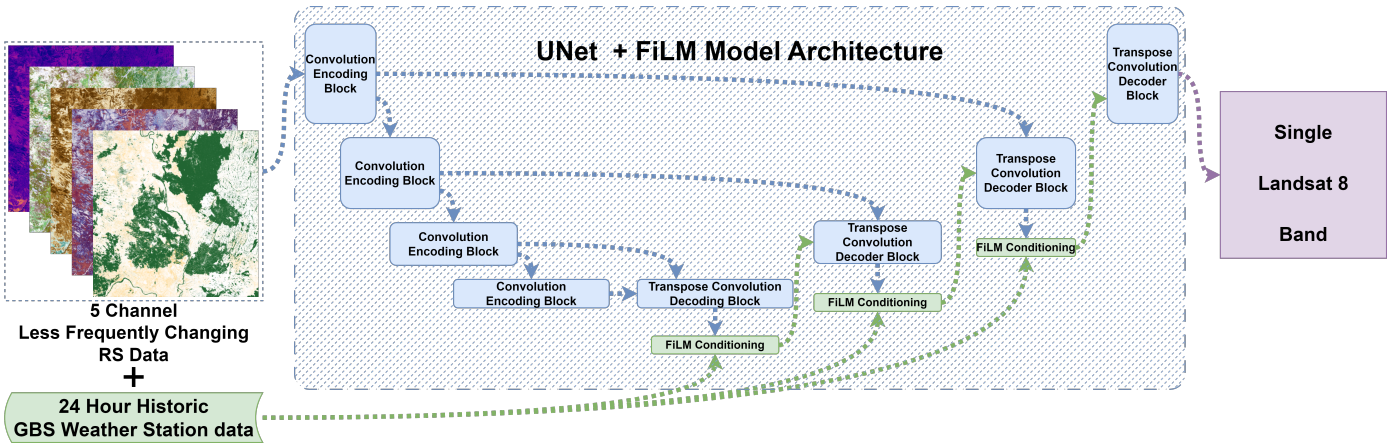



**Figure 5.** NOAH benchmark UNet + FiLM model architecture. The input to the UNet + FiLM model architecture is the 5-channel, less frequently changing remote sensing data and the 24 hour historic GBS weather station data shown at the left. The UNet + FiLM model architecture is shown in the blue box with a blue dotted background. In the UNet + FiLM model architecture, the convolution and the transpose convolution block are shown in blue, the FiLM conditioning block is shown in green, and the flow of outputs is represented with arrows. The model-generated synthetic data for a single Landsat 8 band image is shown in purple on the right.

FiLM is a conditioning mechanism which can modulate modalities based on an external signal (GBS data) using affine transformations. Hence, instead of hard-coding fusion rules or having attention within less frequently changing input RS modalities, FiLM lets a neural network decide how one single (GBS data point) should influence other less frequently changing RS input modalities (topography, vegetation, and fuel type) to produce frequently changing RS modalities (i.e., the individual bands in the satellite modality). This paper adopts UNet + FiLM, as their combination has been proven effective for multi-modality architectures in [46,62,63]. The applications of FiLM + UNet in cardiac medical image segmentation were studied in [62], while the application in image separation was studied in [63].

The objective of this paper is to demonstrate the concept of UNet and FiLM, rather than finding an optimal solution to the target problem. Various possible combinations can be explored in future for performance evaluation or finding an optimal solution. For example, backbone architectures such as Pix2Pix [64], Earthformer [65], or Vision Transformers (ViT) [66] could have been used instead of UNet to generate feature maps, but were not, to account for limited computational resources at training time. Similar to varying the backbone architectures, changes in the loss function and validation on different samples may vary the outcome of the results. Finding the ideal loss function and impact of sampling techniques is also beyond the scope of the proof-of-concept modeling this research.

To maintain consistency, every model was trained for 100 epochs with a learning rate of 0.00001 on Adam optimizer. The batch size was limited to two to account for the limited compute training capability. Models were measured on Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR). MSE is the difference between the actual and target pixel value, and a lower value is better. SSIM is measured between 0 and 1, with a higher value being better. PSNR measure reconstruction and compression, in which larger values are ideal.

### *4.2. Modeling Metrics*

In this subsection, we go over the three metrics (MSE, SSIM, PSNR) used in this research for the proof-of-concept modeling.

#### 4.2.1. Mean Squared Error

Mean Squared Error (MSE) is the average of pixel-wise differences between the generated and original images. The lower the value, the better the results are in terms of the pixel difference between the original and generated images. It should be noted that the MSE alone does not imply a better perceptually generated image with respect to the original image, as it does not measure the structural similarities between the two images. A slight shift or translation between the generated and original images will produce a huge MSE. A mathematical formulation of MSE is given in Equation (1):

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (I_{(i,j)} - \hat{I}_{(i,j)})^2 \quad (1)$$

where $I_{(i,j)}$ is the original pixel value at location $(i, j)$, $\hat{I}_{(i,j)}$ is the generated pixel value at location $(i, j)$, $M$ is the height of the original and generated images, and $N$ is the width of original and generated images. It should be noted that the height and width of the original and generated images should be the same.

#### 4.2.2. Structural Similarity Index Measure

Structural Similarity Index Measure (SSIM) checks the human visual structural perception between the generated image and the original image, instead of measuring pixel

errors. SSIM does so by decomposing the measured difference between the generated and original images into three parts (luminance, contrast, and structure). A detailed mathematical derivation and its complete formulation, along with the selection of constants, are presented in [67]. A simplified mathematical derivation of SSIM is as follows:

The luminance ($l$) is defined in [67] as follows:

$$l(I, \hat{I}) = \frac{(2 \times \mu_I \times \mu_{\hat{I}}) + C_1}{\mu_I^2 + \mu_{\hat{I}}^2 + C_1}$$

where $\mu_I$ is the pixel sample mean of the original images, $\mu_{\hat{I}}$ is the pixel sample mean of the generated images, and the constant $C_1$ is included to avoid instability when $(\mu_I^2 + \mu_{\hat{I}}^2)$ is close to zero. Hence, $C_1$ is defined in [67] as follows:

$$C_1 = (k_1 \times L)$$

where $k_1 = 0.01$ by default and $L$ is the dynamic range of the pixel values.

The contrast ($c$) is defined in [67] as follows:

$$c(I, \hat{I}) = \frac{(2 \times \sigma_I \times \sigma_{\hat{I}}) + C_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2}$$

where $\sigma_I$ is the sample variance of the original images, $\sigma_{\hat{I}}$ is the sample variance of the generated images, and the constant $C_2$ is included to avoid instability when $(\sigma_I^2 + \sigma_{\hat{I}}^2)$ is close to zero. Hence, $C_2$ is defined in [67] as follows:

$$C_2 = (k_2 \times L)$$

where $k_2 = 0.03$ by default and $L$ is the dynamic range of the pixel values.

The structure (s) is defined in [67] as follows:

$$s(I, \hat{I}) = \frac{\sigma_{I\hat{I}} + C_3}{(\sigma_I \times \sigma_{\hat{I}}) + C_3}$$

where $\sigma_I$ is the sample variance of the original image, $\sigma_{\hat{I}}$ is the sample variance of the generated image, and the constant $C_3$ is included to avoid instability when $(\sigma_I \times \sigma_{\hat{I}})$ is close to zero. $C_3$ is defined as $\frac{C_2}{2}$

The resulting similarity index SSIM is defined in [67] as follows in Equation (2):

$$SSIM(I, \hat{I}) = l(I, \hat{I})^{\alpha} \times c(I, \hat{I})^{\beta} \times s(I, \hat{I})^{\gamma} \quad (2)$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters used to adjust the relative importance of the three components.

#### 4.2.3. Peak Signal to Noise Ratio

Peak Signal to Noise Ratio (PSNR) measures how close a generated image is to the original image on pixel by pixel basis. The higher the PSNR value, the closer the generated image is to the original image. PSNR uses MSE between the generated and original image, but does not correlate to human perception of the image, as it does not measure the structure between the two images. A mathematical formulation of PSNR is given in Equation (3):

$$PSNR = 10 \log_{10} \left( \frac{MAX^2}{MSE} \right) \quad (3)$$

where $MAX$ is the maximum possible pixel value and $MSE$ is the mean square error between the generated and original image.

### *4.3. Benchmark Results*

The results in this subsection are for the UNet + FiLM architecture described in Section 4.1. The MSE, SSIM, PSNR are reported for every band generated for Landsat 8 using SSIM loss in Table 2 for the NOAH mini validation dataset. It can be seen in Table 2 that the performance of the models is poor. This is primarily attributed to the use of NOAH mini instead of the full NOAH dataset, as NOAH mini only has 60 data points to train on. Training the model on the complete NOAH dataset requires significant computational resources. Not all prospective researchers in the field may have access to such resources. Therefore, we benchmarked the results on the NOAH mini dataset, as most researchers could work with this dataset size. Prospective researchers who wish to use the complete NOAH dataset for training can do so and report their metrics via benchmarking done on the NOAH mini. There should be a notable increase in the model performance once it is trained on full NOAH with 234,089 data points (which can be further extended as more data is generated). Furthermore, it should also be noted that increased dataset size alone may not resolve limitations related to physical consistency, station representativeness, or loss function choice.

In Table 2, the bands that performed better for a given metric are highlighted in green, and the ones that performed the best are in bold font. Overall, bands B9 to B11 had better SSIM and MSE compared to other bands. B9 is used to detect high to thin cloud cover. B10 and B11 are TIR bands which can be used to map the surface temperature. This is encouraging as the higher positive results for B9, B10, and B11 show significant potential for GM for producing model-generated or synthetic surface temperature data, weather modeling, and to filling in missing data. It should be noted here that the positive result is regarding image similarity, measured in terms of SSIM and PSNR and not physical accuracy of temperature or energy balance.

**Table 2.** Tabulation of NOAH mini Validation Dataset Results across Landsat 8 Bands.

| Generated Landsat 8 Band | UNet + FiLM | | |
|---|---|---|---|
| | **MSE** | **SSIM** | **PSNR** |
| Band 1 (B1): Ultra Blue, Coastal Aerosol | $1.68 \pm 0.33$ | $0.23 \pm 0.08$ | $1.02 \pm 0.05$ |
| Band 2 (B2): Blue | $1.53 \pm 0.43$ | $0.15 \pm 0.06$ | **$4.17 \pm 1.02$** |
| Band 3 (B3): Green | $1.16 \pm 0.02$ | $0.13 \pm 0.05$ | $3.46 \pm 1.58$ |
| Band 4 (B4): Red | $1.21 \pm 0.20$ | $0.09 \pm 0.07$ | $2.48 \pm 1.11$ |
| Band 5 (B5): Near InfraRed | $1.45 \pm 0.68$ | $0.06 \pm 0.03$ | $0.71 \pm 0.09$ |
| Band 6 (B6): Shortwave InfraRed (1.57–1.65 µm) | $2.96 \pm 1.85$ | $0.13 \pm 0.09$ | $-3.41 \pm 1.08$ |
| Band 7 (B7): Shortwave InfraRed (2.11–2.29 µm) | $1.80 \pm 0.97$ | $0.04 \pm 0.03$ | $-2.91 \pm 0.41$ |
| Band 8 (B8): Panchromatic | $1.36 \pm 0.72$ | $0.23 \pm 0.16$ | $4.13 \pm 1.39$ |
| Band 9 (B9): Cirrus | $0.89 \pm 0.12$ | **$0.58 \pm 0.02$** | $0.97 \pm 0.21$ |
| Band 10 (B10)-TIR (10.60–11.19 µm) | **$0.70 \pm 0.08$** | $0.44 \pm 0.01$ | $2.63 \pm 0.61$ |
| Band 11 (B11)-TIR (11.50–12.51 µm) | $0.83 \pm 0.35$ | $0.45 \pm 0.01$ | $2.73 \pm 1.93$ |

The first column specifies the bands, and the subsequent three columns represent the MSE, SSIM, and PSNR metrics. The highest metric in each column is in bold. The cells that show better metrics in each column compared to others in the respective column are colored in green.

### *4.4. Ablation Study Results*

The results of the ablation study are given in Table 3. MSE, SSIM, and PSNR metrics are given for the validation dataset for each band while removing a single source of less frequently changing RS data product. For example, in columns 2 to 4, only MSE, SSIM, and PSNR values are given for the validation set when biomass was removed as an input. The SSIM cells are highlighted in either green or red to indicate either a gain or a loss in

SSIM when compared to the benchmark results presented in Table 2. For example, in B11, the removal of biomass, fuel type, and land cover caused a decrease in the performance of the model's ability to generate B11 validation data, while the removal of crown cover and topography caused an increase in the performance of the model's ability to generate B11. Only SSIM is highlighted, as SSIM was the loss function for model training. The cells in each row for a specific band generation are also put in bold to indicate either the highest gain from the baseline or the highest loss from the baseline. For example, in B11, the highest gain in SSIM was with the removal of topography, and the highest loss was with the removal of land cover.

**Table 3.** Ablation Study of NOAH mini with Validation Dataset Results across Landsat 8 Bands.

| Bands | w/o Biomass | | | w/o Crown Cover | | | w/o Fuel Types | | | w/o Land Cover | | | w/o Topography | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | SSIM | PSNR | MSE | SSIM | PSNR | MSE | SSIM | PSNR | MSE | SSIM | PSNR | MSE | SSIM | PSNR |
| B1 | 1.25 | 0.18 | −2.80 | 0.68 | 0.22 | −1.77 | 1.25 | **0.03** | −6.92 | 0.97 | 0.09 | −5.19 | 0.77 | 0.08 | −1.91 |
| B2 | 0.85 | **0.25** | −1.05 | 0.94 | **0.09** | −3.82 | 1.13 | 0.10 | −0.25 | 1.09 | 0.21 | −1.75 | 1.05 | 0.14 | 2.08 |
| B3 | 1.18 | **0.14** | 1.60 | 0.88 | 0.13 | −3.21 | 1.17 | **0.11** | −2.22 | 1.18 | 0.12 | −0.94 | 0.76 | 0.12 | −4.56 |
| B4 | 0.85 | 0.17 | −1.76 | 1.17 | **0.01** | −4.83 | 0.98 | **0.20** | −0.55 | 1.70 | **0.20** | −0.29 | 0.76 | 0.17 | −1.45 |
| B5 | 1.65 | **0.00** | −6.47 | 1.42 | 0.04 | −3.90 | 3.05 | 0.03 | −3.04 | 2.13 | **0.08** | −2.65 | 1.30 | 0.06 | −3.60 |
| B6 | 4.14 | 0.04 | −3.96 | 2.59 | **0.01** | −5.49 | 1.13 | 0.11 | −2.18 | 1.49 | 0.07 | −5.95 | 3.76 | 0.03 | −3.49 |
| B7 | 1.86 | 0.01 | −5.95 | 1.91 | **0.00** | −7.75 | 1.76 | **0.00** | −10.37 | 1.22 | **0.00** | −6.28 | 1.11 | **0.00** | −2.57 |
| B8 | 1.70 | 0.10 | −0.30 | 0.99 | 0.10 | −0.86 | 1.10 | 0.14 | −1.75 | 1.28 | 0.07 | −0.98 | 1.19 | **0.06** | −1.43 |
| B9 | 1.71 | **0.19** | −0.96 | 0.93 | 0.39 | −0.53 | 1.02 | 0.37 | 0.44 | 1.12 | 0.49 | −2.02 | 0.78 | 0.57 | 0.64 |
| B10 | 0.87 | 0.34 | −1.01 | 0.76 | 0.44 | 1.13 | 0.69 | **0.45** | 2.48 | 0.72 | 0.34 | −1.58 | 1.56 | **0.11** | −5.01 |
| B11 | 0.68 | 0.42 | 3.11 | 0.67 | 0.52 | 0.21 | 0.96 | 0.41 | 3.27 | 0.67 | **0.37** | 2.17 | 0.62 | **0.55** | 1.85 |

The first column specifies the bands, and the subsequent groups of three columns represent the MSE, SSIM, and PSNR metrics without (w/o) one less frequently changing modality. Each row represents the ablation studies MSE, SSIM, and PSNR metrics of an individual band generation with one less frequently changing derived remote sensing and geospatial modalities. The highest loss and the highest gain for each band generation w/o one less frequently changing modality is put in bold in each row. The cells for the SSIM loss are colored either green or red to indicate a gain or loss in the SSIM metric when compared to the baseline presented in Table 2. Only the SSIM values are highlighted as the models were trained using SSIM loss only.

In Table 3, we observe that for band B1 (ultra blue, coastal aerosol), the model's ability to produce similar model-generated synthetic data for B1 is inferior to the baseline of $0.23 \pm 0.08$ in Table 2 if fuel types, land cover, or topography are removed. Based on empirical observation, we notice that biomass and crown cover do not play any significant role in the model's ability to generate synthetic data for B1. The largest loss was noted on the removal of fuel types. Fuel types may signal the model about the combustible particles in the atmosphere which impact B1, hence the highest impact. Similarly, land cover will provide information on the terrain climatic features and vegetation type, which may signal the model to generate better data for B1.

For B2 (blue) in Table 3, we observed that the removal of crown cover had the most impact on the model's ability to produce similar model-generated synthetic data for B2 compared to the baseline of $0.15 \pm 0.06$ in Table 2. Fuel types and topography also showed some loss, but did not deviate from the baseline. This was likely due to water and other blue light reflecting data being missing from crown cover, fuel types, or topography modalities. Removing the biomass modality showed an improvement in the model's ability to generate synthetic data for B2. We presume that having biomass information provides no additional information for the model to generate synthetic data for blue light and may further increase noise when trained on a small scale of the NOAH mini dataset. Similarly, land cover also showed an improvement but did not deviate from the baseline.

In the case of B3 (green), we see no deviations from the baseline of $0.13 \pm 0.05$ in Table 2 on the removal of any of the modalities in case of model-generated synthetic data for B3 in Table 3. We presume that since the green light information is being provided by

multiple modalities, such as fuel types and land cover, removing just one does not impact the model's ability to produce model-generated synthetic data for green light in B3.

In Table 3, band B4 (red) showed an improvement in producing similar model-generated synthetic data for red light from the baseline of $0.09 \pm 0.07$ in Table 2 on the removal of biomass, fuel types, land cover, or topography. The greatest improvement was seen with the removal of fuel types or land cover. Since no red light information is being provided by these modalities, we presume that having them in training at a small scale of the NOAH mini dataset added noise, which was discarded on the removal of the modalities. We also noticed a small deviation in reduced performance for the generation of synthetic red light data from the baseline in the case of the removal of crown cover. The deviation will still fall within the margin of error, and the performance of the baseline models itself was very poor for B4. Hence, it is difficult to derive any conclusive evidence as results are largely inconclusive.

For band B5 (near infrared) in Table 3, we observed that the removal of biomass caused a decrease in performance for the production of model-generated synthetic data for B5 from the baseline of $0.06 \pm 0.03$ in Table 2. Other modalities showed no deviation from the baseline. Although the removal of land cover showed the most improvement in the generation of synthetic B5 data, the values are still within the baseline. Since the performance of the baseline models itself was poor for B5, it is difficult to derive any conclusion as results are largely inconclusive for model-generated synthetic data for B5.

Band B6 (shortwave infrared 1.57–1.65 µm) in Table 3 demonstrated reduced performance in the production of similar model-generated synthetic data for B6 from the baseline of $0.13 \pm 0.09$ in Table 2 on the removal of crown cover or topography, with the highest loss being on the removal of crown cover. Similar to B4 and B5, the baseline results for B6 were not ideal. Hence, it would be difficult to derive conclusions as results are largely inconclusive.

In the case of B7 (shortwave infrared 2.11–2.29 µm) in Table 3, we noticed that removal of any of the modality caused a decrease in the performance in the model's ability to produce similar model-generated synthetic data for B7 from the baseline of $0.04 \pm 0.03$ in Table 2. The SSIM metric for all was 0.00 except for biomass, which was 0.01. It would be difficult to derive conclusions as the results are largely inconclusive for band B7. Similar to bands B4, B5, and B6, the baseline results for B7 were not ideal.

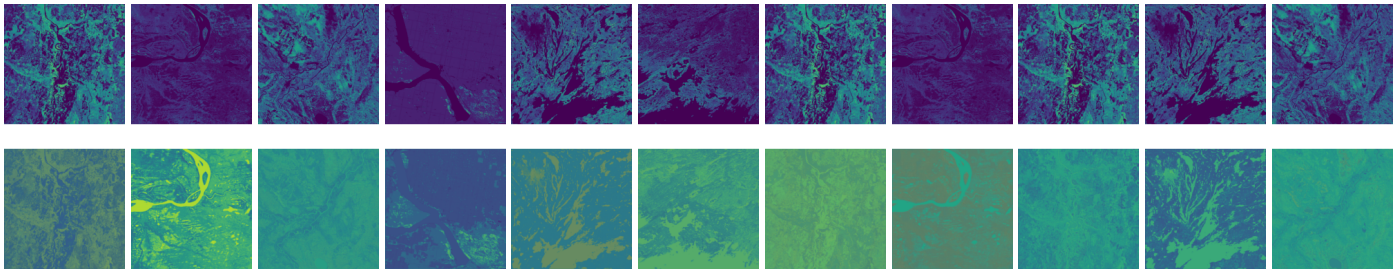
For band B8 (Panchromatic) in Table 3, we noticed that removal of land cover or topography showed a reduced performance in the production of similar model-generated synthetic data for B8 compared to the baseline value of $0.23 \pm 0.16$ in Table 2. B8 records sunlight reflected from a broad visible–near-infrared range and is known to be bad for vegetation indices. We also see that there was no impact on the removal of vegetation-related modalities such as biomass or crown cover. Although land cover, which is a vegetation modality, showed a deviation from the baseline metrics, the deviation was very minimal. Further, the removal of fuel type data showed the least change from the baseline values for the model-generated synthetic data for B8. The greatest loss in model-generated synthetic data for B8 was due to the removal of topography. Slope and elevation can create bright spots in B8 data for sun-facing slopes, and this may impact the model generation of synthetic data for B8.

B9 (Cirrus) in Landsat is designed for the detection of thin clouds. In Table 3, we see that the removal of any of the modalities except topography showed a reduction in the model's synthetic data generative capability compared to the baseline of $0.58 \pm 0.02$ in Table 2 for band B9. The deviations from the baseline for the removal of individual modalities were also significant. Further investigation will be needed to determine why the modalities impact the generation of synthetic data for B9.

For band B10 (TIR 10.60–11.19 µm), removal of biomass, land cover, or topography showed a decrease in the similarity of the model's ability to generate synthetic data in Table 3 compared to the baseline value of $0.44 \pm 0.01$ in Table 2. Since TIR is related to surface temperature, the loss of elevation, slope data from topography, vegetation data from land cover, or tree volume (diameter and height) data from biomass may hinder the model's ability to produce similar model-generated synthetic data for B10.

In band B11 (TIR 11.50–12.51 µm), removal of biomass, fuel types, or land cover showed a decrease in the similarity of the model's ability to generate synthetic data in Table 3 compared to the baseline value of $0.45 \pm 0.01$ in Table 2. Similar to B10, loss of tree volume data from biomass or vegetation data from land cover may hinder the model's ability to produce similar model-generated synthetic data for B11. Further, removal of the crown cover or topography improved the model's ability to generate synthetic data for band B11. This may be due to the B11 not being the ideal band to measure surface temperature independent of other bands, unlike B10. Additionally, B11 is sensitive to water vapor in the atmosphere and is contaminated by stray light, which may impact the production of model-generated synthetic data.

A sample of the generated bands from the trained models on the validation data can be seen in Figure 6. The first row presents the real images captured from Landsat 8 and the second shows the generated images. The columns are different bands of Landsat 8. Since SSIM loss was used during training, it can be seen that the images generated can easily capture the structure of the terrain for the respective band.



(**a**) B1 (**b**) B2 (**c**) B3 (**d**) B4 (**e**) B5 (**f**) B6 (**g**) B7 (**h**) B8 (**i**) B9 (**j**) B10 (**k**) B11

**Figure 6.** Real and Generated Images from Bands 1 to 11 of Landsat 8 in Each Column. (The first row is the real band values, and the second is the generated band values).

## 5. Discussion

The objective of this paper is to introduce NOAH, a multi-modal and sensor fusion dataset that enables the creation of soft sensors for near-real-time model-generated synthetic data estimation. To demonstrate this concept, we adopt UNet and FiLM, which have proven effective for multi-modal data fusion, including images [46,62,63]. One of our future research directions will be to identify the optimal architecture and model size for this problem, as we can explore various combinations and evaluate their performance on different loss functions to determine the best approach.

The primary aim of the research is to provide a proof-of-concept approach to provide model-generated or synthetic band data at a large scale. Further, providing the code to future researchers to generate large real-world (non-simulated) multi-modal sensor fusion datasets will help empower them in this direction. The open-sourced code allows for the regeneration of identical NOAH datasets and the NOAH mini dataset. The code also gives researchers the ability to expand the NOAH dataset as new data is produced by the ECCC, SCANFI, and Landsat. We anticipate the following positive societal applications of our research:

(i) *Training multi-modal sensor fusion foundational models with in-situ and RS data.*

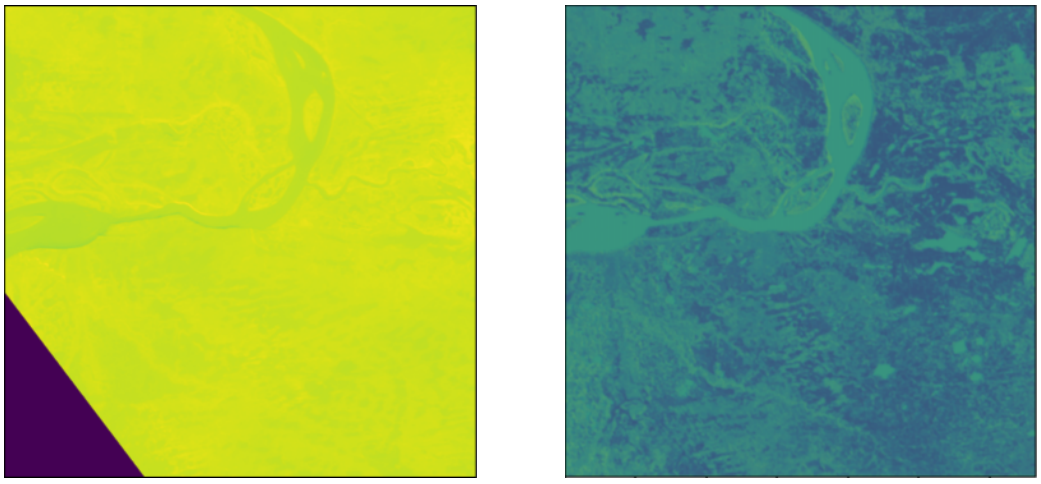
(ii) *Expanding existing Landsat 8 or 9 data tiles to get higher coverage*: Most Landsat tiles produced by USGS and NASA are not perfect squares. GM, as demonstrated in this research, will be able to expand the existing tile data to get higher coverage. An example of expanding existing data can be seen in Figure 7, where the missing data in the lower left corner of Figure 7a was filled in by the respective band model of UNet + FiLM architecture. The model-generated or synthetic filled-in output can be seen in Figure 7b. GM can also provide the ability to produce model-generated or synthetic missing data when multiple tiles are combined.

(iii) *Filling in missing data due to cloud cover with sensor fusion from GBS source*: Similar to how band data can be expanded, GM using sensor fusion from GBS source can be applied to the tiles masked with cloud cover to produce model-generated or synthetic missing data.

(iv) *Generation of near-real-time data for lacking RS coverage during times of disaster*: Sun synchronous satellites do not have near-real-time coverage for all locations. Using GM, we can expand the existing data of nearby regions to cover the regions where disasters like wildfires occur to produce model-generated or synthetic information such as surface temperature and weather.

(v) *Generate new hypothetical satellite band data by varying in-situ measurements to see the impact of climate change*: As seen from Table 2, bands B9, B10, and B11 produce better model-generated or synthetic data capability using multi-modal sensor fusion. Since in-situ measurements from GBS are provided as input to the model in the form of GBS data, one can vary the input of GBS value to see the resulting hypothetical change in generated output surface temperature and weather values. It should be noted that varying GBS inputs produce model-consistent hypothetical scenarios, not physically based climate simulations. The model-generated or synthetic data from varying GBS can then be applied to visualize climate change using GM or GenAI.

(vi) *Generate super-resolution models*: Since the NOAH dataset uses fine spatial resolution images, prospective researchers can train super-resolution models by providing a down-sampled NOAH satellite image modality as input to the model during training and have the model predict the fine spatial resolution NOAH satellite image modality.



(**a**) Captured Image with Missing Data (**b**) Generated Image Trained on NOAH mini

**Figure 7.** Expanding/Filling in Missing Data using UNet + FiLM Model Trained on NOAH mini.

A negative societal impact of models trained for such tasks could be felt in the insurance sector, where the generated data may be used in the calculation of insurance premium amounts.

It should be noted that fine resolution (high resolution) remote sensing image analysis is challenging as highlighted in [68,69]. This is even more challenging when dealing with multiple modalities and large tile areas. For this reason, although insights were derived from the ablation study as to why a decrease or an increase in the similarity of the produced model-generated synthetic data was noticed, a detailed investigation needs to be conducted as part of the future work to identify all the causes.

With the complete NOAH dataset, one can use the large tile area coverage of 40,000 $km^2$ to validate the physical accuracy of temperature or energy balance using in-situ measurements from the weather station GBS present within the tile area, apart from the one in the center of the tile, as presented in the descriptive statistics in Figure 2b. The validation of physical accuracy of temperature or energy balance can also be done on the NOAH mini dataset, but due to its limited sample size and limited overlap with other weather station GBS, it is not ideal to conduct such validation on only the NOAH mini dataset. Instead of the validation of physical accuracy of temperature or energy balance, the use of similarity metrics is recommended to validate the results for only the NOAH mini dataset.

## 6. Limitations

While we propose NOAH to advance GenAI or GM for RS, there are a few data limitations. The primary limitation is that the region to be covered needs to have a GBS weather station there, to collate it with other less frequently changing RS and geospatial data products to generate bands of satellite images. This limits its coverage to a certain extent. The secondary limitation lies in the sources from which the data are being collected. If there are inaccurate readings recorded from either the RS satellites or the GBS sensors, those inaccurate readings will be incorporated in NOAH. Also, if the less frequently changing RS data products from SCANFI have inaccurate measurements, the inaccurate data will be passed on to NOAH. Although the aforementioned limitations are common in collated data sources, including NOAH, we hope this will be mitigated in NOAH since we provide multi-modal data from different sources. Inaccuracies in the modality of one source can be caught in the modality of other sources.

## 7. Conclusions

This study introduces NOAH, a novel multi-modal (remote sensing (topography, vegetation, satellite, fuel) and GBS (wind direction, wind speed, pressure, dew point, temperature, and humidity)) sensor fusion (anemometers, barometers, thermometer, SAR, TIR, OLI) dataset collated from 815 distinct locations across Canada. It complements the use of in-situ measurements (GBS data) to generate new or missing remote sensing data. NOAH has applications in GM or GenAI, such as to produce missing data, expand existing data, fill in missing data, generate super-resolution data products and produce model-generated or synthetic data for hypothetical climatic conditions. It can also produce model-generated or synthetic near-real-time data for disaster modeling when RS satellite coverage is unavailable.

The data in NOAH is extensively researched and collated from standard, open-source, and well-documented data sources (ECCC, SCANFI, Landsat). NOAH dataset can be expanded as new data is produced by data sources. The code for generating and expanding the dataset is made open-sourced. NOAH has a non-overlapping spatial coverage of 8,742,469 $km^2$ with 234,089 Landsat 8 images covering 815 distinct locations from 2013 to 2025 at 30 m spatial resolution. An option to expand NOAH to Landsat 9 has been made available in the open-source code for future researchers. NOAH has future growth potential as the data sources provide more data, and the code to expand it is also available on GitHub. The size of NOAH exceeded 20 TB; therefore, NOAH mini is made publicly

available on Hugging Face. All code used to collate, generate, expand, and benchmark the data is made publicly available on GitHub.

In the proof-of-concept models built on the NOAH dataset, models are trained to output Landsat band images based on GBS data and less frequently changing RS data products (topography, vegetation, fuel). Therefore, they can expand to regions not covered by satellite (Landsat) data. An example is given in Figure 7. With the ability to produce model-generated or synthetic data for regions not covered by satellite overpass when having point source information on the ground, users of the model will be able to generate derived RS data products more frequently by reducing the temporal coverage.

The models benchmarked on the NOAH mini showed less-than-ideal results. This was expected as the sample size of 93 is small for the NOAH mini dataset compared to 234,089 samples in the NOAH dataset. We would also like to note that increased dataset size alone may not resolve limitations related to physical consistency, station representativeness, or loss function choice.

Even with the less-than-ideal results, bands B9 Cirrus, B10 TIR, and B11 TIR have the best results. This shows potential for GM to produce model-generated or synthetic data for missing values, surface temperature, and to model weather. It should be noted that the potential is in reference to the image similarity measured using SSIM and PSNR, not the physical accuracy of temperature or energy balance.

The choice of NOAH mini for benchmarking was made so that a standard set of data would be available for all future researchers to benchmark on, irrespective of their computational resource restrictions. Further, we aimed for a consistent training time for all bands for benchmarking. This would have caused some bands to overfit and others to underfit, which was highlighted in the ablation study in Section 4.3. We expect that future researchers will be able to use the code and data to train foundation models to generate specific bands, fill in missing values, expand existing coverage, and generate near-real-time Landsat data, which can be used to make derived RS products for disaster modeling and EO.

**Author Contributions:** Conceptualization, A.M., C.-H.L., M.Z., and K.N.; methodology, A.M.; software, A.M.; validation, A.M., C.-H.L., M.Z., D.U., K.N., K.M., T.R., and R.P.; formal analysis, A.M., C.-H.L., M.Z., K.N., K.M., and T.R.; investigation, A.M., C.-H.L., M.Z., K.N., K.M., T.R., and R.P.; resources, A.M.; data curation, A.M.; writing—original draft preparation, A.M.; writing—review and editing, A.M., C.-H.L., M.Z., D.U., K.N., K.M., T.R., and R.P.; visualization, A.M.; supervision, C.-H.L., M.Z., and K.N.; project administration, C.-H.L., M.Z., and K.N.; funding acquisition, C.-H.L., K.N., and M.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partly funded by an Alliance Missions grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada grant number ALLRP 570503-2021.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data presented in this study were derived from the following resources available in the public domain: Weather Station https://dd.weather.gc.ca/today/climate/observations/ [44] accessed on 10 December 2025, Landsat 8 https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1 [35] accessed on 10 December 2025, Fuel Types https://open.canada.ca/data/dataset/4e66dd2f-5cd0-42fd-b82c-a430044b31de [38] accessed on 10 December 2025, Crown Cover https://ftp.maps.canada.ca/pub/nrcan_rncan/Forests_Foret/SCANFI/v1/SCANFI_att_closure_SW_2020_v1.2.tif [37] accessed on 10 December 2025, Biomass https://ftp.maps.canada.ca/pub/nrcan_rncan/Forests_Foret/SCANFI/v1/SCANFI_att_biomass_SW_2020_v1.2.tif [37] accessed on 10 December 2025, Topography https://ftp.maps.canada.ca/pub/nrcan_rncan/Forests_Foret/SCANFI/v1/SCANFI_att_height_SW_2020_v1.2.tif [37]

accessed on 10 December 2025 and Land Cover https://open.canada.ca/data/en/dataset/11990a35-912e-4002-b197-d57dd88836d7 [41] accessed on 10 December 2025, are provided by Environment and Climate Change Canada (ECCC), National Aeronautics and Space Administration (NASA), United States Geological Survey (USGS), Google Earth Engine (GEE), Natural Resources Canada (NRCan), and Spatialized CAnadian National Forest Inventory (SCANFI). The NOAH mini dataset prepared in this study is available at Hugging Face NOAH-mini repository at https://huggingface.co/datasets/mutakabbirCarleton/NOAH-mini. The authors confirm that the complete processed NOAH dataset is available from the corresponding author upon reasonable request.

**Acknowledgments:** The research produced is part of ongoing collaborative work between Carleton University, University of Waterloo, and Dalhousie University with industry partners Cistel Technology and Hegyi Geomatics International Inc. Additional support was received from Research Computing Services at Carleton University.

**Conflicts of Interest:** Author Marzia Zaman was employed by the company Cistel Technology. Author Darshana Upadhyay was employed by the company TeleAI Corporation. Author Thambirajah Ravichandran was employed by the company Hegyi Geomatics Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| CRS | Coordinate Reference System |
| DEM | Digital Elevation Model |
| ECCC | Environment and Climate Change Canada |
| EO | Earth Observation |
| FiLM | Feature-wise Linear Modulation |
| FoV | Field of View |
| GBS | Ground-Based Sensor |
| GEE | Google Earth Engine |
| GenAI | Generative Artificial Intelligence |
| GM | Generative Modeling |
| LST | Land Surface Temperature |
| ML | Machine Learning |
| NOAH | Now Observation Assemble Horizon |
| OLI | Operational Land Imager |
| PSNR | Peak Signal-to-Noise Ratio |
| RS | Remote Sensing |
| SAR | Synthetic Aperture Radar |
| SCANFI | Spatialized CAnadian National Forest Inventory |
| SM | Soil Moisture |
| SSIM | Structural Similarity Index Measure |
| TIR | Thermal InfraRed |
| UN | United Nations |
| USGS | United States Geological Survey |
| WMO | World Meteorological Organization |

## References

1. Ma, R.H.; Wang, Y.H.; Lee, C.Y. Wireless Remote Weather Monitoring System Based on MEMS Technologies. *Sensors* **2011**, *11*, 2715–2727. [CrossRef]
2. Thies, B.; Bendix, J. Satellite based remote sensing of weather and climate: Recent achievements and future perspectives. *Meteorol. Appl.* **2011**, *18*, 262–295. [CrossRef]

3. Schmude, J.; Roy, S.; Trojak, W.; Jakubik, J.; Civitarese, D.S.; Singh, S.; Kuehnert, J.; Ankur, K.; Gupta, A.; Phillips, C.E.; et al. Prithvi wxc: Foundation model for weather and climate. *arXiv* **2024**, arXiv:2409.13598.
4. Li, R.; Xie, Y.; Jia, X.; Wang, D.; Li, Y.; Zhang, Y.; Wang, Z.; Li, Z. SolarCube: An Integrative Benchmark Dataset Harnessing Satellite and In-situ Observations for Large-Scale Solar Energy Forecasting. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; pp. 3499–3513. Available online: https://proceedings.neurips.cc/paper_files/paper/2024/file/06477eb61ea6b85c6608d42a222462df-Paper-Datasets_and_Benchmarks_Track.pdf (accessed on 10 December 2025)
5. Mitra, A.K. Use of Remote Sensing in Weather and Climate Forecasts. In *Social and Economic Impact of Earth Sciences*; Gahalaut, V.K., Rajeevan, M., Eds.; Springer Nature: Singapore, 2023; pp. 77–96. [CrossRef]
6. Wang, W.; Bieker, J.; Arcucci, R.; Quilodran-Casas, C. Data Assimilation using ERA5, ASOS, and the U-STN Model for Weather Forecasting over the UK. In Proceedings of the NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning, New Orleans, LA, USA, 10–16 December 2023. Available online: https://www.climatechange.ai/papers/neurips2023/61 (accessed on 10 December 2025)
7. Ashkboos, S.; Huang, L.; Dryden, N.; Ben-Nun, T.; Dueben, P.; Gianinazzi, L.; Kummer, L.; Hoefler, T. ENS-10: A Dataset For Post-Processing Ensemble Weather Forecasts. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 21974–21987. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/89e44582fd28ddfea1ea4dcb0ebbf4b0-Paper-Datasets_and_Benchmarks.pdf (accessed on 10 December 2025)
8. Oskarsson, J.; Landelius, T.; Lindsten, F. Graph-based Neural Weather Prediction for Limited Area Modeling. In Proceedings of the NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning, New Orleans, LA, USA, 10–16 December 2023.
9. Perrakis, D.D.; Cruz, M.; Alexander, M.; Hanes, C.; Thompson, D.; Taylor, S.; Stocks, B. Improved Logistic Models of Crown Fire Probability in Canadian Conifer Forests. *Int. J. Wildland Fire* **2023**, *32*, 1455–1473. [CrossRef]
10. Martino, L.; Ulivieri, C.; Jahjah, M.; Loret, E. Remote Sensing and GIS Techniques for Natural Disaster Monitoring. In *Space Technologies for the Benefit of Human Society and Earth*; Olla, P., Ed.; Springer: Dordrecht, The Netherlands, 2009; pp. 331–382. [CrossRef]
11. Teodoro, A.; Duarte, L. Chapter 10-The role of satellite remote sensing in natural disaster management. In *Nanotechnology-Based Smart Remote Sensing Networks for Disaster Prevention*; Denizli, A., Alencar, M.S., Nguyen, T.A., Motaung, D.E., Eds.; Micro and Nano Technologies; Elsevier: Amsterdam, The Netherlands, 2022; pp. 189–216. [CrossRef]
12. Anand, V.; Miura, Y. PreDisM: Pre-Disaster Modelling with CNN Ensembles for At-Risk Communities. In Proceedings of the NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning, Online, 6–14 December 2021. Available online: https://www.climatechange.ai/papers/neurips2021/53 (accessed on 10 December 2025).
13. Ballard, T.; Erinjippurath, G.; Cooper, M.W.; Lowrie, C. Widespread Increases in Future Wildfire Risk to Global Forest Carbon Offset Projects Revealed by Explainable AI. In Proceedings of the ICLR 2023 Workshop on Tackling Climate Change with Machine Learning, Kigali, Rwanda, 1–5 May 2023.
14. Wang, Y.; Sun, Y.; Cao, X.; Wang, Y.; Zhang, W.; Cheng, X. A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 311–334. [CrossRef]
15. Perbet, P.; Guindon, L.; Côté, J.F.; Béland, M. Evaluating deep learning methods applied to Landsat time series subsequences to detect and classify boreal forest disturbances events: The challenge of partial and progressive disturbances. *Remote Sens. Environ.* **2024**, *306*, 114107. [CrossRef]
16. Allison, R.S.; Johnston, J.M.; Craig, G.; Jennings, S. Airborne Optical and Thermal Remote Sensing for Wildfire Detection and Monitoring. *Sensors* **2016**, *16*, 1310. [CrossRef]
17. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A Review on Early Forest Fire Detection Systems Using Optical Remote Sensing. *Sensors* **2020**, *20*, 6442. [CrossRef]
18. Mutakabbir, A.; Lung, C.H.; Ajila, S.A.; Naik, K.; Zaman, M.; Purcell, R.; Sampalli, S.; Ravichandran, T. A Federated Learning Framework based on Spatio-Temporal Agnostic Subsampling (STAS) for Forest Fire Prediction. In Proceedings of the 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), Osaka, Japan, 2–4 July 2024; pp. 350–359. [CrossRef]
19. Mutakabbir, A.; Lung, C.H.; Ajila, S.A.; Zaman, M.; Naik, K.; Purcell, R.; Sampalli, S. Forest Fire Prediction Using Multi-Source Deep Learning. In *Big Data Technologies and Applications BDTA*; Lecture Notes of the Institute for Computer Science, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2023; Volume 555, pp. 135–146.
20. Mutakabbir, A.; Lung, C.H.; Ajila, S.A.; Zaman, M.; Naik, K.; Purcell, R.; Sampalli, S. Spatio-Temporal Agnostic Deep Learning Modeling of Forest Fire Prediction Using Weather Data. In Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 26–30 June 2023; pp. 346–351. [CrossRef]
21. Mutakabbir, A.; Lung, C.H.; Naik, K.; Zaman, M.; Ajila, S.A.; Ravichandran, T.; Purcell, R.; Sampalli, S. Spatio-Temporal Agnostic Sampling for Imbalanced Multivariate Seasonal Time Series Data: A Study on Forest Fires. *Sensors* **2025**, *25*, 792. [CrossRef]

22. Boroujeni, S.P.H.; Razi, A.; Khoshdel, S.; Afghah, F.; Coen, J.L.; O'Neill, L.; Fule, P.; Watts, A.; Kokolakis, N.M.T.; Vamvoudakis, K.G. A Comprehensive Survey of Research towards AI-Enabled Unmanned Aerial Systems in pre-, active-, and post-Wildfire Management. *Inf. Fusion* **2024**, *108*, 102369. [CrossRef]
23. Vazquez, D.A.Z.; Qiu, F.; Fan, N.; Sharp, K. Wildfire Mitigation Plans in Power Systems: A Literature Review. *IEEE Trans. Power Syst.* **2022**, *37*, 3540–3551. [CrossRef]
24. Chuvieco, E.; Aguado, I.; Salas, J.; García, M.; Yebra, M.; Oliva, P. Satellite Remote Sensing Contributions to Wildland Fire Science and Management. *Curr. For. Rep.* **2020**, *6*, 81–96. [CrossRef]
25. Li, Z.L.; Wu, H.; Duan, S.B.; Zhao, W.; Ren, H.; Liu, X.; Leng, P.; Tang, R.; Ye, X.; Zhu, J.; et al. Satellite Remote Sensing of Global Land Surface Temperature: Definition, Methods, Products, and Applications. *Rev. Geophys.* **2023**, *61*, e2022RG000777. [CrossRef]
26. Natural Resources Canada. Canadian Wildland Fire Information System | Canadian National Fire Database (CNFDB). 2021. Available online: https://cwfis.cfs.nrcan.gc.ca/ha/nfdb (accessed on 10 December 2025) .
27. Mo, Y.; Xu, Y.; Chen, H.; Zhu, S. A Review of Reconstructing Remotely Sensed Land Surface Temperature under Cloudy Conditions. *Remote Sens.* **2021**, *13*, 2838. [CrossRef]
28. Paszkuta, M. Impact of Cloud Cover on Local Remote Sensing—Piaśnica River Case Study. *Oceanol. Hydrobiol. Stud.* **2022**, *51*, 283–297. [CrossRef]
29. Hedley, J.; Russell, B.; Randolph, K.; Dierssen, H. A physics-based method for the remote sensing of seagrasses. *Remote Sens. Environ.* **2016**, *174*, 134–147. [CrossRef]
30. Shen, H.; Li, X.; Cheng, Q.; Zeng, C.; Yang, G.; Li, H.; Zhang, L. Missing Information Reconstruction of Remote Sensing Data: A Technical Review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 61–85. [CrossRef]
31. Thomas, C.; Ranchin, T.; Wald, L.; Chanussot, J. Synthesis of Multispectral Images to High Spatial Resolution: A Critical Review of Fusion Methods Based on Remote Sensing Physics. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1301–1312. [CrossRef]
32. Ebel, P.; Xu, Y.; Schmitt, M.; Zhu, X.X. SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5222414. [CrossRef]
33. Zi, Y.; Xie, F.; Song, X.; Jiang, Z.; Zhang, H. Thin Cloud Removal for Remote Sensing Images Using a Physical-Model-Based CycleGAN with Unpaired Data. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1004605. [CrossRef]
34. Deznabi, I.; Kumar, P.; Fiterau, M. Zero-Shot Microclimate Prediction with Deep Learning. In Proceedings of the NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning, New Orleans, LA, USA, 10–16 December 2023. Available online: https://www.climatechange.ai/papers/neurips2023/41 (accessed on 10 December 2025).
35. U.S. Geological Survey (USGS). Landsat 8 OLI/TIRS Collection 2 Level-2 Data; Accessed via Google Earth Engine (LANDSAT/LC08/C02/T1), 2020. Available online: https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C02_T1 (accessed on 10 December 2025).
36. Guindon, L.; Manka, F.; Correia, D.L.; Villemaire, P.; Smiley, B.; Bernier, P.; Gauthier, S.; Beaudoin, A.; Boucher, J.; Boulanger, Y. A new approach for spatializing the Canadian National Forest Inventory (SCANFI) using Landsat dense time series. *Can. J. For. Res.* **2024**, *54*, 793–815. [CrossRef]
37. Guindon, L.; Villemaire, P.; Correia, D.L.; Manka, F.; Sam, L.; Smiley, B. *SCANFI: Spatialized CAnadian National Forest Inventory Data Product*; Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre: Quebec, QC, Canada, 2023. [CrossRef]
38. LaCarte, S. *Canadian Forest Fire Danger Rating System (CFFDRS) Fire Behaviour Prediction (FBP) Fuel Types 2024, 30 M*; Government of Canada, Natural Resources Canada, Canadian Forest Service: Ottawa, ON, Canada, 2024. Available online: https://open.canada.ca/data/dataset/4e66dd2f-5cd0-42fd-b82c-a430044b31de (accessed on 10 December 2025) .
39. Latifovic, R. *2010 Land Cover of Canada*; Government of Canada; Natural Resources Canada; Canada Centre for Remote Sensing. 2017. Available online: https://open.canada.ca/data/en/dataset/c688b87f-e85f-4842-b0e1-a8f79ebf1133 (accessed on 10 December 2025)
40. Latifovic, R. *Canada's Land Cover (ver. 2015)*; Natural Resources Canada, General Information Product, 119e, Natural Resources Canada. 2019. Available online: https://ostrnrcan-dostrncan.canada.ca/entities/publication/99dc3562-a6de-494a-9093-00f490f6df0a (accessed on 10 December 2025).
41. Latifovic, R. *Land Cover of Canada-Cartographic Product Collection*; Government of Canada; Natural Resources Canada; Canada Centre for Remote Sensing. 2019. Available online: https://open.canada.ca/data/en/dataset/11990a35-912e-4002-b197-d57dd88836d7 (accessed on 10 December 2025)
42. Latifovic, R. *2015 Land Cover of Canada*; Government of Canada; Natural Resources Canada; Canada Centre for Remote Sensing. 2023. Available online: https://open.canada.ca/data/en/dataset/4e615eae-b90c-420b-adee-2ca35896caf6 (accessed on 10 December 2025)
43. Latifovic, R. *2020 Land Cover of Canada*; Government of Canada; Natural Resources Canada; Canada Centre for Remote Sensing. 2024. Available online: https://open.canada.ca/data/en/dataset/ee1580ab-a23d-4f86-a09b-79763677eb47 (accessed on 10 December 2025)

44. Environment and Climate Change Canada. *Government of Canada Historical Climate Data*; Canadian Centre for Climate Services: Gatineau, QC, Canada. Available online: https://dd.weather.gc.ca/today/climate/observations/ (accessed on 10 December 2025)
45. Shang, C.; Yang, F.; Huang, D.; Lyu, W. Data-driven soft sensor development based on deep learning technique. *J. Process Control* **2014**, *24*, 223–233. [CrossRef]
46. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
47. Gondal, M.W.; Wuthrich, M.; Miladinovic, D.; Locatello, F.; Breidt, M.; Volchkov, V.; Akpo, J.; Bachem, O.; Schölkopf, B.; Bauer, S. On the Transfer of Inductive Bias from Simulation to the Real World: A New Disentanglement Dataset. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32. Available online: https://proceedings.neurips.cc/paper_files/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf (accessed on 10 December 2025)
48. Dawson, H.L.; Dubrule, O.; John, C.M. Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. *Comput. Geosci.* **2023**, *171*, 105284. [CrossRef]
49. Mutakabbir, A.; Lung, C.H.; Zaman, M.; Naik, K.; Purcell, R.; Sampalli, S.; Ravichandran, T. Vegetation Land Cover and Forest Fires in Canada: An Analytical Data Visualization. In Proceedings of the 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), Toronto, ON, Canada, 8–11 July 2025.
50. Kondylatos, S.; Prapas, I.; Camps-Valls, G.; Papoutsis, I. Mesogeos: A Multi-Purpose Dataset for Data-Driven Wildfire Modeling in the Mediterranean. In Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023. Available online: https://papers.nips.cc/paper_files/paper/2023/file/9ee3ed2dd656402f954ef9dc37e39f48-Paper-Datasets_and_Benchmarks.pdf (accessed on 10 December 2025)
51. Li, Y.; Li, K.; Li, G.; Wang, Z.; Ji, C.; Wang, L.; Zuo, D.; Guo, Q.; Zhang, F.; Wang, M.; et al. Sim2Real-Fire: A Multi-Modal Simulation Dataset for Forecast and Backtracking of Real-World Forest Fire. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 37, pp. 1428–1442. Available online: https://papers.nips.cc/paper_files/paper/2024/file/02e978a2cc9a1d0d4376a7deb01db612-Paper-Datasets_and_Benchmarks_Track.pdf (accessed on 10 December 2025)
52. Abowarda, A.S.; Bai, L.; Zhang, C.; Long, D.; Li, X.; Huang, Q.; Sun, Z. Generating Surface Soil Moisture at 30m Spatial Resolution using Both Data Fusion and Machine Learning toward Better Water Resources Management at the Field Scale. *Remote Sens. Environ.* **2021**, *255*, 112301. [CrossRef]
53. Zhao, W.; Duan, S.B. Reconstruction of Daytime Land Surface Temperatures under Cloud-Covered Conditions using Integrated MODIS/Terra Land Products and MSG Geostationary Satellite Data. *Remote Sens. Environ.* **2020**, *247*, 111931. [CrossRef]
54. Cornebise, J.; Oršolić, I.; Kalaitzis, F. Open High-Resolution Satellite Imagery: The WorldStrat Dataset –With Application to Super-Resolution. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 25979–25991. Available online: https://papers.nips.cc/paper_files/paper/2022/file/a6fe99561d9eb9c90b322afe664587fd-Paper-Datasets_and_Benchmarks.pdf (accessed on 10 December 2025)
55. Giglio, L.; Justice, C. MOD14. NASA LP DAAC. 2015. Available online: https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod14-006 (accessed on 10 December 2025).
56. Giglio, L.; Justice, C. MOD14A1. NASA LP DAAC. 2015. Available online: https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod14a1-006 (accessed on 10 December 2025).
57. Giglio, L.; Justice, C. MOD14A2. NASA LP DAAC. 2015. Available online: https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod14a2-006 (accessed on 10 December 2025).
58. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 8815–8827. Available online: https://papers.nips.cc/paper_files/paper/2023/file/1be3843e534ee06d3a70c7f62b983b31-Supplemental-Datasets_and_Benchmarks.pdf (accessed on 10 December 2025)
59. Yu, S.; Hannah, W.; Peng, L.; Lin, J.; Bhouri, M.A.; Gupta, R.; Lütjens, B.; Will, J.C.; Behrens, G.; Busecke, J.; et al. ClimSim: A Large Multi-Scale Dataset for Hybrid Physics-ML Climate Emulation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 22070–22084. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/file/45fbcc01349292f5e059a0b8b02c8c3f-Paper-Datasets_and_Benchmarks.pdf (accessed on 10 December 2025)
60. Zhou, H.; Kao, C.H.; Phoo, C.P.; Mall, U.; Hariharan, B.; Bala, K. AllClear: A Comprehensive Dataset and Benchmark for Cloud Removal in Satellite Imagery. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 37, pp. 53571–53597. Available online: https://proceedings.neurips.cc/paper_files/paper/2024/file/60095e1d7ebb292dbba93c4d8f7b2463-Paper-Datasets_and_Benchmarks_Track.pdf (accessed on 10 December 2025)

61. Akhtar, M.; Benjelloun, O.; Conforti, C.; Foschini, L.; Gijsbers, P.; Giner-Miguelez, J.; Goswami, S.; Jain, N.; Karamousadakis, M.; Krishna, S.; et al. Croissant: A Metadata Format for ML-Ready Datasets. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 37, pp. 82133–82148. Available online: https://papers.neurips.cc/paper_files/paper/2024/file/9547b09b722f2948ff3ddb5d86002bc0-Paper-Datasets_and_Benchmarks_Track.pdf (accessed on 10 December 2025)
62. Khan, A.; Asad, M.; Benning, M.; Roney, C.; Slabaugh, G. Compositional Segmentation of Cardiac Images Leveraging Metadata. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 9489–9498. [CrossRef]
63. Meseguer-Brocal, G.; Peeters, G. Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations. In Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 4–8 November 2019. Available online: https://archives.ismir.net/ismir2019/paper/000017.pdf (accessed on 10 December 2025)
64. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Honolulu, HI, USA, 21–26 July 2017. Available online: https://openaccess.thecvf.com/content_cvpr_2017/papers/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.pdf (accessed on 10 December 2025)
65. Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y.; Li, M.; Yeung, D.Y. Earthformer: Exploring Space-Time Transformers for Earth System Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/a2affd71d15e8fedffe18d0219f4837a-Paper-Conference.pdf (accessed on 10 December 2025)
66. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
67. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
68. Li, X.; Xu, F.; Zhang, J.; Zhang, H.; Lyu, X.; Liu, F.; Gao, H.; Kaup, A. Frequency-Guided Denoising Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5400217. [CrossRef]
69. Li, X.; Xu, F.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. A Frequency Decoupling Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5607921. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.