Improved Next-Day Wildfire Spread Prediction and the WSTS+ Benchmark

Saad Lahrichi
University of Missouri
Columbia, MO 65201
saad.lahrichi@missouri.edu

Jake Bova
University of Montana
Missoula, MT 59802
jacob.bova@umt.edu

Jesse Johnson
University of Montana
Missoula, MT 59802
jesse.johnson@umt.edu

Jordan Malof
University of Missouri
Columbia, MO 65201
jmdrp@missouri.edu

Abstract

Recent research has demonstrated the potential of deep neural networks (DNNs) to accurately predict wildfire spread on a given day based upon high-dimensional explanatory data from a single preceding day, or from a time series of T preceding days. For the first time, we investigate a large number of existing data-driven wildfire modeling strategies under controlled conditions, revealing the best modeling strategies and resulting in models that achieve state-of-the-art (SOTA) accuracy for both single-day and multi-day input scenarios, as evaluated on a large public benchmark for next-day wildfire spread, termed the WildfireSpreadTS (WSTS) benchmark. Consistent with prior work, we found that models using time-series input obtained the best overall accuracy, suggesting this is an important future area of research. Furthermore, we create a new benchmark, WSTS+, by incorporating four additional years of historical wildfire data into the WSTS benchmark. Our benchmark doubles the number of unique years of historical data, expands its geographic scope, and, to our knowledge, represents the largest public benchmark for time-series-based wildfire spread prediction.

1. Introduction

Wildfires are a global cause of concern that have severe human, economical, and environmental impacts, with the average annual economic burden from wildfires falling between $71.1 billion and $347.8 billion [41]. In order to better manage, mitigate, and prevent wildfires, accurately predicting their spread is essential. In this work, we focus on the problem of next-day wildfire spread prediction, where we are provided with current and/or historical information about a particular wildfire, and then tasked with predicting its spatial extent on the following day.

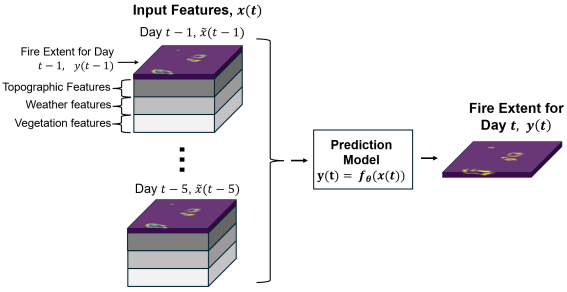

Figure 1. The wildfire prediction models take as input a geospatial map of several variables: vegetation, topography, and weather features, alongside the current day fire mask. We consider two scenarios: one in which the model receives input features from one preceding day, denoted $t-1$, and one in which it receives input from five previous days

A variety of approaches have been investigated to solve this problem, such as those based upon machine learning models [4, 8, 22], or physics-based and observationally-informed models [1, 10, 11]. In this work, however, we focus on a promising emerging class of techniques that utilize high-capacity machine learning models – namely, deep neural networks (DNNs) – to predict wildfire spread using high-dimensional explanatory input data. These input data typically comprise a geospatial raster of the current extent of the fire, as well as explanatory features such as topography, climate, weather, and vegetation indices. Based upon these input data, the model is tasked with producing a geospatial map, or an image, reflecting the spatial extent of the fire on the following day. See Fig. 1 for an illustration.

A variety of DNN-based models have been proposed to solve next-day prediction, including convolutional models [5, 29, 31], attention-based models such as transformers

[40], and spatio-temporal models [3, 32]. One major limitation of most existing work is the lack of standardized evaluation wherein studies often utilize different datasets, model training and evaluation procedures, or compare to few other existing methods. Furthermore, most existing research has focused on next-day prediction where only explanatory data from the current day is input (e.g., only $\tilde{x}(t-1)$ in Fig. 1). However, recent research found that models utilizing a time-series of T previous days of data can achieve greater prediction accuracy [15], suggesting this as an important new direction in next-day wildfire prediction.

Contributions of this Work **Our first contribution** is to perform a rigorous and controlled comparison of many existing approaches, both for single-day ($T = 1$) and the time-series inputs ($T = 5$). We also propose a novel temporal positional embedding. We compare all methods using the same public benchmark dataset, along with the same training, hyperparameter optimization, and evaluation procedures. We perform all experiments on the WildfireSpreadTS (WSTS) benchmark [15] because it is the only public benchmark for time-series wildfire prediction, it sufficiently large to support DNN training and evaluation, and in contrast to all other existing benchmarks, it employs a realistic 12-fold *leave-one-year-out* cross-validation. Our study reveals the best existing modeling strategies, resulting in substantial performance improvements over the current state-of-the-art (SOTA) for the WSTS benchmark.

Our second contribution is to introduce WSTS+, an extended benchmark for next-day wildfire spread, constructed by doubling the number of years of historical wildfire events in WSTS. By adopting WSTS+, we can double the size of our training datasets. We conduct experiments that reveal additional historical training data, either in WSTS or WSTS+, yields little improvement in the accuracy of wildfire models, and we analyze the causes for this.

The rest of the paper is structured as follows: we formulate our problem setting in Sec. 2, Sec. 3 reviews related works, Sec. 4 describes WSTS, Sec. 5 presents preliminaries, Sec. 6 details our experiments, Sec. 7 introduces WSTS+, and Sec. 8 concludes.

2. Problem Setting

In its general formulation, the goal of next-day wildfire spread prediction is to predict a wildfire's spatial extent on some t^{th} day, denoted $y(t)$, given explanatory data from one or more *preceding* days, denoted $x(t)$. We adopt the more specific settings of recent literature [15, 20], as illustrated in Fig. 1, which assume there are T consecutive previous days of explanatory data, so that $x(t) = \{\tilde{x}(t-i)\}_{i=1}^{T}$, and each $\tilde{x}(t)$ comprises a geospatial raster, so that $\tilde{x}(t) \in \mathbb{R}^{H \times W \times C}$, where H, W correspond to spatial dimensions, and C represents the number of explanatory variables, which may include previous fire masks (e.g., $y(t) \subset \tilde{x}(t)$). The fire extent is encoded in a binary geospatial image, $y(t) \in \{0,1\}^{H \times W}$, where a value of one indicates the presence of a fire. Our goal is then to use a dataset of historical wildfire data to infer parameters, θ, of a predictive model of the form $y(t) = f_\theta(x(t))$.

3. Related Works

Next Day Wildfire Segmentation DNN-based segmentation has received growing attention due to its accuracy, enabled by the recent development of large datasets of historical fire data. [20] created Next-Day Wildfire Spread, a large and public dataset for *next-day* spread prediction, and used it to train a custom deep segmentation model. Concurrently, [36] developed SeasFire Cube and trained Unet++ models [44] for medium-term fire prediction, between 8 and 64 days. [23] improved upon the collected data cube and found that the LSTM and ConvLSTM models outperformed the Fire Weather Index (FWI). In FireSight, [18] collected a dataset using remote sensing data from 20 datasets, and trained a 3D UNet model to model short-term fire hazard, between 3 and 8 days. Overall, most work has been done using U-Net architectures and their variants, and many authors [12, 26, 40, 42] have recently reported that attention-based U-Nets achieve greater accuracy. We investigate the SwinUnet[30] in our study as a widely used and therefore representative example of such models.

Next Day Wildfire Prediction with Time-Series In contrast to the existing work discussed above, we also focus on time-series input for spread prediction, which has been cited as an important emerging direction [12, 15, 26]. Historically, time-series modeling has been challenging due to the lack of appropriate public datasets to train and evaluate models for this task. Recently, [15] extended the Next-Day Wildfire Spread dataset from [20] to be suitable for time-series prediction, and achieved their best overall next-day predictions using a time-series model, termed UTAE [14].

Other DL Approaches Aside from a segmentation formulation, researchers have also investigated, for example, reinforcement learning [39], probabilistic cellular automata [16], and synthetic data approaches [27]. We refer readers to [43] for a review of DL for wildfire prediction.

4. The WSTS Benchmark

In this work, we employ the WildfireSpreadTS benchmark [15]. The dataset includes 607 wildfire events across the western United States between 2018 and 2021, totaling 13,607 daily multi-channel images. These 23 channels include data on active fires, weather, topography, and vegetation, resampled to a common resolution of 375 meters,

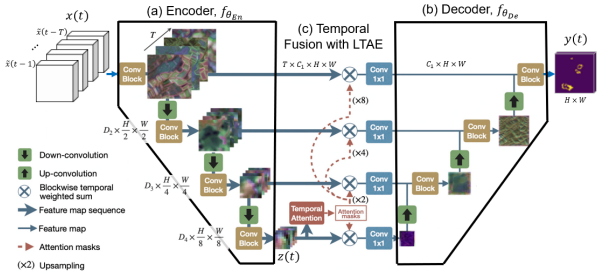

Figure 2. Illustration of the U-Net and UTAE models, adapted from [14] to our wildfire problem: see description in main text.

providing a multi-modal and multi-temporal framework for modeling fire spread. A key feature of this benchmark is a rigorous 12-fold cross-validation evaluation procedure. Each fold of the cross-validation includes all wildfire events from a single year, so that the trained models are always evaluated on wildfire events from a previously unseen year, reflecting real-world use of wildfire prediction models.

5. Preliminaries: the UTAE

The UTAE [14], originally developed for satellite imagery is essentially a U-Net that has been modified to process a time-series of imagery, and was recently found successful for modeling wildfire spread [15]. We propose a novel modification of the time-series positional encodings and therefore discuss the technical details of the UTAE here. The UTAE encodes each entry in the time-series independently using a shared encoder shown in Fig. 2(a), and then fuses the resulting embeddings from each day using a Lightweight Temporal Self-Attention (LTAE) block [13], shown in Fig. 2(c). Given a T-length time-series of input, the encoder produces a series of T embeddings $z(t) = \{\tilde{z}(t-i)\}_{i=1}^{T}$ where $\tilde{z}(t) \in \mathbb{R}^{D_4 \times \frac{H}{8} \times \frac{W}{8}}$ at the output of the last layer of the encoder. Then the LTAE computes an attention mask, $a \in \mathbb{R}^{T \times \frac{H}{8} \times \frac{W}{8}}$, which is utilized to combine the T embeddings. Before computing the temporal attention, LTAE adds a sinusoidal positional embedding, $p(\bar{t})$ to each input embedding, where $\bar{t} \in [1, 365]$ is an integer representing the day of the year, and $p(\bar{t})$ maps $\bar{t}$ to a unique sinusoidal representation. This positional embedding is motivated by the original application of UTAE to agricultural segmentation, where the appropriate segmentation depends heavily upon the day of the year. Once the attention mask is computed, it is then upsampled, and applied to the encoder embeddings output at each resolution to collapse the temporal dimension. After all temporal dimensions are collapsed, a conventional U-Net-like decoder is applied to the collapsed embeddings, as shown in Fig. 2(b).

6. Improving Wildfire Spread Prediction

In this section, we describe our methods for $T = 1$ and $T = 5$ scenarios, respectively, as well as experiments to support them (e.g., ablations). Results for our developed benchmark models are reported in Tab. 2, in terms of Average Precision (AP) using 12-fold leave-one-year-out cross-validation on the WSTS benchmark, following prior work [15]. Also following [15], we report model performance for three feature sets: vegetation features only (Veg), a combination of vegetation and topographic features (Multi), and all features (All), which includes additional weather forecast features. Full experimental details are provided in Sec. 9 in the supplement. Models in Tab. 2 with citations correspond to the three current best models on WSTS, as reported in [15]. All other models reported in Tab. 2 were developed in this work.

6.1. Single-Day Input ($T = 1$)

The current $T = 1$ SOTA utilizes a U-Net architecture with a ResNet-18 encoder, and is denoted Res18-Unet[15]. Therefore we focus our investigation on improving the Res18-Unet[15].

Modeling Improvements We next describe the investigated improvements to the Res18-Unet[15] at a high level. More details can be found in Sec. 9 of the supplement.

(i) Encoders. Better performance may be obtained with larger encoders or those with attention mechanisms. Recent studies have indicated that attention-based models may be superior to convolutional models for wildfire spread [26, 40, 42, 45]. We investigate a ResNet50 [19] encoder, as well as the attention-based SwinUnet-Tiny encoder [6].

(ii) Utilizing Pre-trained Parameters. Utilizing pre-trained weights to initialize training is a well-established technique to improve model accuracy, including in remote sensing [21]. We investigate pre-trained weights for each of the encoders that we consider (i.e., ResNet18, ResNet50, and SwinUnet), while the decoders are trained from scratch.

(iii) Improved Loss Functions. The existing SOTA Res18-Unet [15] is trained using weighted binary cross-entropy loss. However, it has been established that Jaccard/Dice losses are often superior alternatives for segmentation tasks [9], and focal loss has been shown effective for class imbalance [28] (the WSTS benchmark exhibits severe class imbalance), and for wildfire spread in particular [12]. Therefore we investigate and compare the aforementioned losses in our experiments.

(iv) Improved Hyperparameter Optimization. The existing SOTA Res18-Unet [15] was trained by selecting the model with the highest F1 score on the validation, however, all models on WSTS are evaluated utilizing the average precision (AP) metric [15]. We investigate aligning the valida-

tion and testing metrics by using AP for both.

For our experiments, we consider a U-Net with a ResNet-18 encoder (denoted *Res18-Unet*), a ResNet-50 encoder (denoted *Res50-Unet*), and a SwinUnet-Tiny (denoted *SwinUnet*). For each of these models, we perform a grid search over all combinations of learning rates ($[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$), loss functions (BCE, Focal, Dice, Jaccard), and the use of pre-training or not (a binary choice). Following [15], we use a single fold of the 12-fold cross-validation, and only one of the three feature sets (the "All" set) for this optimization. As discussed, in contrast to previous work, we utilize AP during validation to select the best models instead of F1. The focal loss has two hyperparameters: α, set as the inverse frequency of positive class pixels, and γ, set to its default value of two.

Experimental Results We found that pre-training was nearly always beneficial, and that Focal Loss usually yielded substantial improvements compared to our other candidate losses. Therefore, for the WSTS benchmark, we included both pre-training and focal loss in all three of our models: *Res18-Unet*, *Res50-Unet*, and *SwinUnet*. As an ablation study, Tab. 1 reports the performance of our *Res18-Unet* on the full WSTS benchmark, where we progressively remove each of our improvements to assess its impact. Our results indicate that each improvement is highly beneficial, or at least not significantly harmful.

Tab. 2 reports the performance of our three models on the WSTS benchmark, compared to the best existing $T = 1$ model, *Res18-Unet[15]*. Our *Res18-Unet* is identical to the *Res18-Unet[15]*, except for our aforementioned modifications, and obtains substantially higher AP across all input features considered: a 37% improvement on average.

Our other two models, *Res50-Unet* and *SwinUnet*, also substantially outperform the existing *Res18-Unet[15]*. However, despite having approximately twice the number of trainable model parameters of *Res18-Unet*, we find that our model outperforms the two larger models in most cases. *Our Res18-Unet also obtains the highest overall AP (0.468) for the $T = 1$ models when utilizing the "Multi" feature set, establishing a new SOTA on WSTS for $T = 1$.*

Several recent studies have reported that large and/or attention-based models achieve SOTA accuracy for $T = 1$ wildfire spread prediction [26, 40, 42, 45]. However, we find here that with simple improvements and appropriate optimization, *Res18-Unet* outperforms such models. We also suspect that the more rigorous (and potentially more real-world) leave-one-year-out cross-validation adopted by the WSTS benchmark may penalize more complex models for overfitting, compared to the *Res18-Unet*.

In Fig. 3, we qualitatively evaluate our *Res18-Unet* and *Res50-Unet* against the Res18-Unet [15]. Each row corresponds to a fire event, and the columns show the current fire, the next day label, and the predictions of each model. Yellow represents the fire extent, green shows correctly predicted burned areas, and red shows false positives. We observe that the original model tends to overpredict fire spread, leading to multiple red patches where no fire actually occurs. However, the model also underpredicts in areas where the fire spreads, capturing some, but not the full extent of the fire. On the other hand, we observe that our models make consistently more accurate predictions, with far fewer false positives, and slightly better matching green areas.

Table 1. Ablation showing the impact of the successive removal of each of our improvements on a *Res18-Unet* trained on Vegetation features

Model	Test AP	Percent Decrease
Res18-Unet (ours)	0.455 ± 0.092	–
No pretraining	0.456 ± 0.086	-0.22
No focal loss	0.345 ± 0.084	24.18
No AP as validation	0.321 ± 0.078	29.45

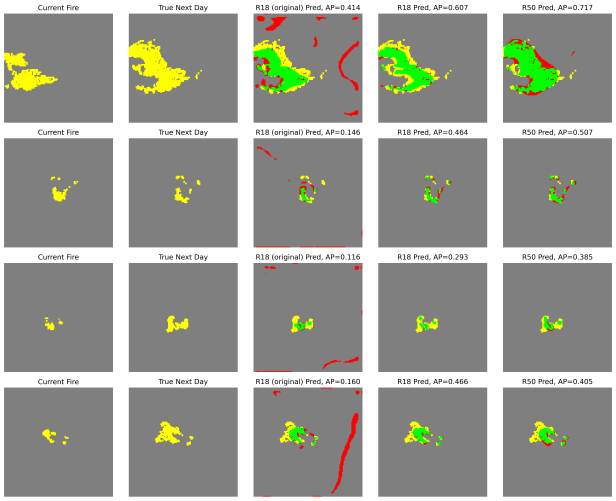

Figure 3. Sample predictions made by the Res18-Unet [15], our *Res18-Unet*, and *Res50-Unet*. The two leftmost columns show the current fire spread $y(t-1)$ and the next-day label $y(t)$. True positive pixels are colored in green, while false positives are colored in red

6.2. Time-Series Input, $T = 5$

Existing models for the time-series scenario generally adopt one of two approaches: (i) a data-level fusion, or (ii) a feature-level fusion. In data-level fusion, the features for each day, $\tilde{x}(t) \in \mathbb{R}^{H \times W \times C}$ are concatenated along the feature dimension into one input tensor, $x(t) = [\tilde{x}(t-1)|, ..., |\tilde{x}(t-T)] \in \mathbb{R}^{H \times W \times CT}$, after which they can be processed in the same manner as single-day input (see

Table 2. Mean test AP $\pm$ standard deviation using vegetation features only (Veg), vegetation, land cover, topography and weather (Multi) and All features, when training with 1 and 5 input days. Models with citations represent accuracy reported on our benchmark from previous publications; all other models reported are developed in this work. Results style: **best**

Fusion Level	Model	Input days	Veg	Multi	All	# Params
-	Res18-Unet[15]	1	0.328 ± 0.090	0.341 ± 0.085	0.341 ± 0.086	14.3M
	Res18-Unet	1	0.455 ± 0.090	$\mathbf{0.468 \pm 0.087}$	$\mathbf{0.460 \pm 0.084}$	14.3M
	Res50-Unet	1	$\mathbf{0.457 \pm 0.089}$	0.459 ± 0.090	0.451 ± 0.093	32.5M
	SwinUnet	1	0.432 ± 0.088	0.437 ± 0.082	0.424 ± 0.090	27.2M
Data	Res18-Unet[15]	5	0.333 ± 0.079	0.344 ± 0.076	0.325 ± 0.108	14.4M
	Res18-Unet	5	$\mathbf{0.472 \pm 0.083}$	$\mathbf{0.469 \pm 0.087}$	$\mathbf{0.460 \pm 0.084}$	14.4M
	SwinUnet	5	0.447 ± 0.087	0.453 ± 0.083	0.435 ± 0.079	27.3M
Feature	UTAE[15]	5	0.372 ± 0.088	0.350 ± 0.113	0.321 ± 0.135	1.1M
	UTAE	5	0.452 ± 0.082	0.459 ± 0.088	0.433 ± 0.099	1.1M
	UTAE(Res18)	5	$\mathbf{0.478 \pm 0.085}$	$\mathbf{0.477 \pm 0.089}$	$\mathbf{0.475 \pm 0.091}$	14.6M

Sec. 2 for problem notation). Therefore, we adopt our best-performing $T = 1$ models from Sec. 6.1, and their hyper-parameter settings, and evaluate them for data-level fusion. As a reference, we also include the *reported* results of the Res18-Unet [15] when it was applied for data-level fusion.

In this context, feature-level fusion implies that we use a shared encoder to first extract features (or embeddings) independently for each day of our input, $\tilde{z}(t) = f_{\theta_{En}}(\tilde{x}(t))$ so that we have a collection of features, $z(t) = \{\tilde{z}(t-i)\}_{i=1}^{T}$, which are utilized as input into a subsequent model for joint processing (i.e., fusion). The current SOTA accuracy on WSTS, both for the time-series setting, and overall, was obtained with a UTAE model [14], as reported in [15]. Furthermore, the UTAE achieved superior accuracy despite having just 1.1M parameters - significantly fewer than many other models considered (e.g., the Res18-Unet has 14.3M). Therefore, we focus our modeling improvements on the UTAE from [15].

Improvements to the UTAE We develop two improved UTAE models, referred to as *UTAE* and *UTAE(Res18)*. We discuss the design of each model next.

Our UTAE Model. Our *UTAE* includes two major improvements over the *UTAE[15]*. The first improvement is to adopt all of the changes investigated for the single-day models from Sec. 6.1. Pursuant to this, following previous work convention, we did a joint search over the following hyperparameters using a single fold of the WSTS benchmark: pre-training (or not), learning rates ($[1e-2, 1e-3, 1e-4, 1e-5]$), and the type of loss (Focal, BCE, Jaccard, and Dice loss). The second improvement is the introduction of a novel positional encoding in the temporal fusion utilized by the *UTAE*. To our knowledge, this modification is novel within the vision and wildfire literature. Specifically, instead of using day-of-year positional encodings, as done in [14, 15], where $\bar{t} \in [1, 365]$, we propose to use a absolute positional encoding that indicates the relative position of each day's set of features within the time-series, so that $\bar{t} \in [1, ..., T]$ for a T-day input. We hypothesize that the features (especially the fire mask) from the most recent day of the fire will be most important for making predictions, and therefore, this relative position information will be much more important than its position in the year. Furthermore, it may be difficult for the models to infer relative positional information from day-of-year encodings, potentially undermining performance.

Our UTAE(Res18) Model. This model is obtained by making one additional improvement to our *UTAE* model. The encoder utilized in the *UTAE[15]* is relatively small (in terms of free parameters). Therefore, in a similar fashion to our investigation in Sec. 6.1, we replace the existing UTAE's encoder with a pre-trained ResNet-18.

Experimental Results Tab. 2 reports the accuracy (in terms of AP) of our time-series models on the WSTS benchmark, categorized by the type of fusion performed: data-level or feature-level. Regarding data-level fusion, our *Res18-Unet* and *Swin-Unet* both substantially outperform the existing *Res18-Unet[15]* across all combinations of input features, with the *Res18-Unet* providing the best overall AP (AP=0.472, on Vegetation features). Regarding feature-level fusion, our two UTAE models (*UTAE* and *UTAE(Res18)*) substantially outperform the existing *UTAE[15]*, which is the current SOTA model on WSTS, both for time-series input ($T > 1$) and overall. Our *UTAE(Res18)* model achieves the highest overall performance for each combination of input features, across both single-day and time-series models. *In particular, our UTAE(Res18) achieves the highest overall AP with the Vegetation (Veg) feature subset, leading to a new overall SOTA*

Table 3. Test AP of UTAE trained on Vegetation features using the original Absolute positional encodings from [15], versus our proposed Relative positional encodings

Pos. Encodings	Absolute	Relative
UTAE	0.419 ± 0.101	$\mathbf{0.452 \pm 0.082}$

performance on WSTS of AP=0.478.

Notably, our results indicate that models receiving time-series input generally outperform those with single-day input. This is especially apparent when comparing data-level fusion models, such as *Res18-Unet* and *SwinUnet*, with their single-day counterparts, since they have few architectural differences. Most existing wildfire spread prediction in the literature has focused on the single-day input; however, our findings here corroborate those from [15] and suggest that time-series modeling is a promising emerging modeling strategy.

Our results also provide evidence that each modeling change is beneficial. As discussed, our *UTAE* included several applicable improvements discussed for our single-day models in Sec. 6.1, as well as our improved temporal encodings described in this sub-section. We therefore conducted an ablation experiment, reported in Tab. 3, to demonstrate that our modified positional encodings provide additional benefits. To show that the pre-trained ResNet-18 encoder is beneficial, we can compare the performance of *UTAE(Res18)* and *UTAE* in Tab. 2: the pre-trained ResNet-18 is the only difference between these two models.

Finally, we observe that increasing the number of input features is not always beneficial, which is consistent with [15], where the best AP was often achieved with the Veg or Multi feature sets rather than the All set. This suggests that the explanatory power of some features are outweighed by the cost of (often significantly) increasing the input dimensionality. We hypothesize that this may be due to the low resolution and/or noise present in some features, such as the weather forecast features, which have a resolution of 27 km, while the fire masks have a resolution of 375 m.

7. The WSTS+ Benchmark

Our results on the WSTS benchmark indicated that relatively simple models performed best, such as those based upon a ResNet-18, rather than models utilizing larger encoders (e.g., ResNet-50) or those utilizing attention (e.g., SwinUnet). This contrasts sharply with the broader vision literature where larger models tend to perform best, given sufficient quantities of training data. Therefore, we hypothesize that collecting more training data would facilitate the use of larger models, yielding superior modeling performance. To investigate this hypothesis, we expand the original WSTS benchmark by curating four additional years of historical wildfire data: 2016, 2017, 2022, and 2023. Our extended dataset, termed WSTS+, contains twice the number of years of historical wildfire data, expands the geographic diversity of the benchmark, and is – to our knowledge – the largest public benchmark for time-series next-day wildfire spread prediction. We visualize the geographic distribution of WSTS+ events in Fig. 4 and find that it much of the new data is in the Western United States, similar to WSTS, but that it includes some unique locations there, and some additional data in the eastern states. Tab. 4 summarizes the differences between both datasets in terms of numbers of years, fire events, total images, and active fire pixels. Further collection details can be found in Sec. 10 of the supplement.

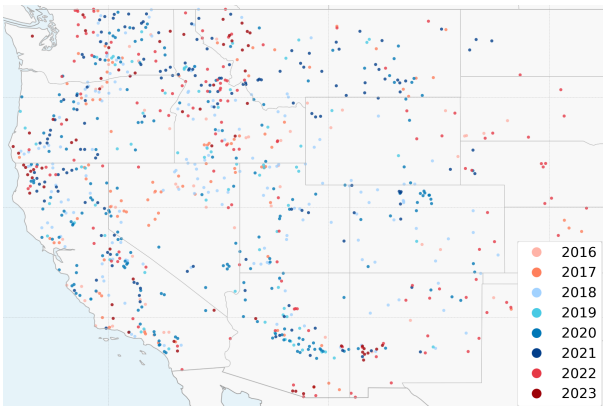

Figure 4. Geographic distribution of the fire events in each year of WSTS (blue) and WSTS+ (red)

Table 4. Comparison between the original WSTS dataset and our extension. We double the number of years and total images and drastically increase the number of fire events and active fire pixels.

Dataset	WSTS	WSTS+	Increase (%)
Years	4 (2018-2021)	8 (2016-2023)	+100
Fire Events	607	1,005	+65.6
Total Images	13,607	24,462	+79.8
Active Fire Px	1,878,679	2,638,537	+40.4

7.1. Benchmarking Models with WSTS+

As compared to WSTS, we propose a new scheme for evaluating models using WSTS+, which exploits its greater size to significantly reduce computational complexity compared to WSTS's 12-fold cross-validation – thereby making the benchmark more accessible to researchers – while maintaining a similar level of real-world rigor. For WSTS+, we propose to divide the available data into four folds that each contain two consecutive years of historical wildfire data. We then evaluate models using four-fold cross-validation, where in each iteration, one fold of data is used for test-

1	(2016, 2017)	(2018, 2019)	(2020, 2021)	(2022, 2023)
2	(2016, 2017)	(2018, 2019)	(2020, 2021)	(2022, 2023)
3	(2016, 2017)	(2018, 2019)	(2020, 2021)	(2022, 2023)
4	(2016, 2017)	(2018, 2019)	(2020, 2021)	(2022, 2023)

Figure 5. New cross-validation folds used for WSTS+. Each pair of consecutive years is used as validation/testing once. Color code: blue: training, orange: validation, green: test

ing, one fold for validation, and two folds for training, as illustrated in Fig. 5. To ensure that the testing and validation sets have the same relative temporal distance to the training set, we always select them so that they are non-consecutive. This results in four-fold cross-validation instead of the twelve-fold cross-validation utilized in WSTS, making it far less computationally intensive. At the same time, this approach doubles the quantity of data in the training and validation sets, ideally allowing researchers to train larger and more sophisticated models. Lastly, because two consecutive years of data are included in the test set, the benchmark still evaluates models under challenging realistic testing conditions.

Table 5. Mean test AP $\pm$ standard deviation using vegetation features only (Veg), vegetation, land cover, topography and weather (Multi) and All features, when training on the WSTS+ dataset Results style: **best**

Model	Days	Veg	Multi	All
Res18-Unet	1	0.349 ± 0.109	0.351 ± 0.105	$\mathbf{0.351 \pm 0.122}$
Res50-Unet	1	0.345 ± 0.096	0.353 ± 0.122	$\mathbf{0.351 \pm 0.122}$
Res18-Unet-LTAE	5	$\mathbf{0.354 \pm 0.113}$	$\mathbf{0.363 \pm 0.129}$	0.350 ± 0.117

7.2. Experimental Results with WSTS+

Using our updated cross-validation scheme, we train our best $T = 1$ models and our best $T = 5$ model on WSTS+ and report the results in terms of mean average precision across all three feature sets in Tab. 5. We see that the performance rank-order of our three models is still similar on WSTS+ as compared to WSTS. However, the overall performance is significantly lower for these models on WSTS+ as compared to WSTS (by roughly 0.1 AP). These results seem to contradict our initial hypothesis that additional training data would enable larger models and improve accuracy. To investigate further, Fig. 6 reports the *per-year* performance for a Res18-Unet trained on either WSTS or WSTS+ (denoted Res18-Unet(WSTS) and Res18-Unet(WSTS+), respectively; see caption for details). These results reveal that both models obtain very similar AP on every testing year, despite Res18-Unet(WSTS+) being trained on twice as many years of data in each fold as Res18-Unet(WSTS). Since the two models perform similarly across all years, the lower overall performance ob-

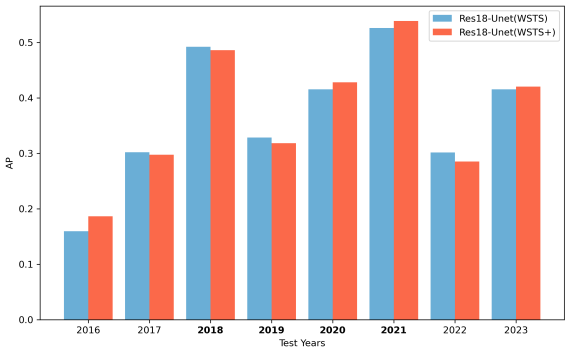

Figure 6. Performance breakdown by test year. Blue bars represent models trained on the original WSTS data, termed Res18-Unet(WSTS), while red bars represent those trained on WSTS+, termed Res18-Unet(WSTS+). The bolded x-axis ticks highlight original test years from WSTS.For Res18-Unet(WSTS+) we simply stratify its performance by year. For Res18-Unet(WSTS), we stratify by year to obtain performance for 2018 to 2021. To obtain performance on the remaining years, we select the cross-validation fold with the best-performing model (as judged by its test fold error) and then report its performance on the newly added WSTS+ years.

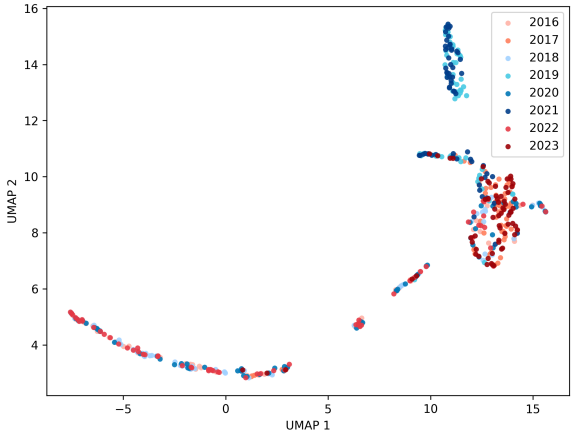

Figure 7. UMAP visualization of the inputs features across years. Each point represents the encoded features at the deepest layer of our best Res18-Unet encoder, with blue indicating original WSTS year and red newly added years in WSTS+

tained on the WSTS+ benchmark in Tab. 5 is likely due to the greater apparent difficulty of the new testing years (2016, 2017, 2022, and 2023), rather than lower predictive accuracy of the models trained on WSTS+.

Domain Shift: A Potential Challenge to Scaling Data-Driven Wildfire Modeling The results in Fig. 6 raise a question: why does significantly increasing the quantity and diversity of the training data in WSTS+ lead to little or no improvement? In Sec. 10 of the supplement, we present evidence that, in pursuit of WSTS+, we accurately reproduced the preprocessing used for WSTS. Therefore we ar-

gue here that this result is likely caused by cross-year *domain shift* [37], wherein the joint probability distribution of the features and targets, denoted $p(x(t), y(t))$, varies across years. There is a large literature about identifying and addressing domain shift (e.g., see [35, 37]), and comprehensively addressing these problems is beyond the scope of this work. However, we seek here to provide evidence that domain shift is present in historical wildfire datasets, provide an initial characterization of it, and discuss the implications of it. There are many different types of domain shift based upon precisely how $p(x(t), y(t))$ changes from training to testing conditions (or across years in our case) [24, 25, 37]). We consider here two widely-studied types of shift: concept and covariate shift.

Concept Shift refers to changes in the conditional distribution $p(y(t)|x(t))$, which in a regression context *generally* implies that the true underlying function $y(t) = f(x(t))$ is changing. Under this hypothesis, we would expect that combining multiple years of data would likely lead to significant reductions in accuracy, since each year exhibits a different underlying relationship. Our results in Fig. Fig. 6 indicate that adding two additional years of training data, as is done in WSTS+, did not significantly impact accuracy, suggesting that significant concept shifts are unlikely. In Tab. 6 we also report the results of an experiment where we train eight Res18-Unet models - one on data from each year - and then test each model on a disjoint test set from each year (see Sec. 11 of the supplement for experimental details). The results indicate that, given a specific single testing year, most models achieve relatively similar accuracy, also suggesting they are each learning similar concepts.

Covariate Shift refers to change in the marginal distribution, $p(x(t))$. Covariate shifts are thought to often have limited negative impact for high-capacity models [37], such as our DNNs here. Under this hypothesis, additional years of training data may be either beneficial or neutral, but not especially detrimental. This hypothesis is therefore consistent with Fig. 6. It is also corroborated by the results of Tab. 6 if we note that, for a specific testing year, the accuracy of most single-year models is similar to that obtained by WSTS models (trained on two years) and WSTS+ models (trained on four years) in Fig. 6. For example, if we take 2018 as the testing year, then the average of single-year models in Tab. 6 is nearly the same as the WSTS and WSTS+ models in Fig. 6. In other words, despite significant differences in the years included in, and total size of the training data, these models usually perform similarly. Notably, this also indicates that the WSTS benchmark did not benefit from additional training data: the WSTS models in Fig. 6 do not perform differently (on average) than the single-year models in Tab. 6. Lastly, as additional qualitative evidence of covariate shift, Fig. 7 presents a umap visualization of the features extracted by our Res18-Unet for each year in WSTS+. The results show that there is significant overlap in the feature distributions, but there are also significant apparent shifts across years as well.

Implications The analysis above suggests that there is significant cross-year covariate shift in wildfire data, which may explain the limited benefits of additional training years, although we emphasize that domain shift is a complex topic and we have only performed a high-level analysis, suitable for the scope of this paper. If covariate shift is indeed present in wildfire data, we anticipate that this is a temporary challenge; as wildfire datasets grows over time, it becomes less likely that each new year will exhibit distinct feature distributions compared to all preceding years.

Table 6. Cross-year results: Rows show the year the model was trained on, while columns show the year the model was tested on.

Year	2016	2017	2018	2019	2020	2021	2022	2023	Avg
2016	**0.350**	0.291	0.490	0.276	0.173	0.544	0.268	0.416	0.351
2017	0.242	**0.300**	0.487	0.288	0.180	0.568	0.301	0.437	0.351
2018	0.265	0.297	**0.576**	0.313	0.194	0.595	0.344	0.465	0.381
2019	0.219	0.259	0.455	**0.329**	0.159	0.530	0.324	0.428	0.338
2020	0.222	0.263	0.501	0.285	**0.220**	0.572	0.295	0.460	0.352
2021	0.253	0.321	0.534	0.330	0.187	**0.649**	0.328	0.465	0.384
2022	0.227	0.249	0.460	0.261	0.163	0.508	**0.390**	0.416	0.334
2023	0.242	0.279	0.483	0.289	0.157	0.568	0.324	**0.582**	0.365
Avg	0.253	0.282	0.498	0.296	0.179	0.567	0.322	0.459	0.357

8. Conclusion

We investigated the problem of next-day wildfire spread prediction, where we systematically compare a variety of (mostly) existing modeling strategies in two scenarios: single-day ($T = 1$) and time-series ($T = 5$) input, as illustrated in Fig. 1. We conducted our experiments on the WildfireSpreadTS (WSTS) benchmark [15] using a realistic 12-fold leave-one-year-out cross-validation. We draw the following conclusions:

- Our study revealed which modeling strategies perform best, resulting in new models that obtain a 37% and a 28% improvement, respectively, over the current WSTS state-of-the-art for single-day and time-series prediction.
- A time-series model obtained the best overall performance, and time-series models usually outperformed comparable single-day models, suggesting time-series models are an important future area of research.
- We introduce WSTS+, an extension of WSTS, that doubles the number of years of historical wildfire events in WSTS, and yields the largest existing public benchmark for *time-series* spread prediction.
- Analysis of WSTS and WSTS+ suggests that there is significant cross-year domain shift in historical wildfire data. Preliminary investigation suggests it is primarily in the form of covariate shift, undermining the benefits of adding training data, but we hypothesize this problem may subside as total available hisorical data grows.

Future work may focus on investigating the nature of domain shift in historical wildfire data and overcoming any associated challenges, potentially enabling larger or more complex models (e.g., high capacity attention-based models) to realize their full potential performance.

Acknowledgments

The authors gracefully acknowledge partial support from the USDA Forest Service, Rocky Mountain Research Station through joint venture agreement 24-JV-11221636-140. Partial support for Johnson was provided by NSF 2242802. Partial support for Bova was provided by NSF 366456.

References

[1] Martin E Alexander and Miguel G Cruz. Evaluating a model for predicting active crown fire rate of spread using wildfire observations. *Canadian Journal of Forest Research*, 36(11): 3015–3028, 2006. 1

[2] Tomàs Artés, Duarte Oom, Daniele De Rigo, Tracy Houston Durrant, Pieralberto Maianti, Giorgio Libertà, and Jesús San-Miguel-Ayanz. A global wildfire dataset for the analysis of fire regimes and fire behaviour. *Scientific data*, 6(1):296, 2019. 1

[3] Andrew Bolt, Carolyn Huston, Petra Kuhnert, Joel Janek Dabrowski, James Hilton, and Conrad Sanderson. A spatio-temporal neural network forecasting approach for emulation of firefront models. In *2022 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 110–115. IEEE, 2022. 2

[4] Karol Bot and José G Borges. A systematic review of applications of machine learning techniques for wildfire management decision support. *Inventions*, 7(1):15, 2022. 1

[5] John Burge, Matthew R Bonanni, R Lily Hu, and Matthias Ihme. Recurrent convolutional deep neural networks for modeling time-resolved wildfire spread behavior. *Fire Technology*, 59(6):3327–3354, 2023. 1

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 3, 1

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1

[8] Khaled Chetehouna, Eddy El Tabach, Loubna Bouazaoui, and Nicolas Gascoin. Predicting the flame characteristics and rate of spread in fires propagating in a bed of pinus pinaster using artificial neural networks. *Process Safety and Environmental Protection*, 98:50–56, 2015. 1

[9] Tom Eelbode, Jeroen Bertels, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE transactions on medical imaging*, 39(11): 3679–3690, 2020. 3

[10] Mark A. Finney. *FARSITE: Fire Area Simulator-model development and evaluation*. 1998. 1

[11] Mark A Finney. An overview of flammap fire modeling capabilities. In *In: Andrews, Patricia L.; Butler, Bret W., comps. 2006. Fuels Management-How to Measure Success: Conference Proceedings. 28-30 March 2006; Portland, OR. Proceedings RMRS-P-41. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 213-220*, 2006. 1

[12] Jack Fitzgerald, Ethan Seefried, James E Yost, Sangmi Pallickara, and Nathaniel Blanchard. Paying attention to wildfire: Using u-net with attention blocks on multimodal data for next day prediction. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 470–480, 2023. 2, 3

[13] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*, pages 171–181. Springer, 2020. 3

[14] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021. 2, 3, 5, 1

[15] Sebastian Gerard, Yu Zhao, and Josephine Sullivan. Wildfirespreadts: A dataset of multi-modal time series for wildfire spread prediction. *Advances in Neural Information Processing Systems*, 36:74515–74529, 2023. 2, 3, 4, 5, 6, 8, 1

[16] Rohit Ghosh, Jishnu Adhikary, and Rezki Chemlal. Fire spread modeling using probabilistic cellular automata. In *Asian Symposium on Cellular Automata Technology*, pages 45–55. Springer, 2024. 2

[17] Luis Giglio, Chris Justice, Luigi Boschetti, and David Roy. Modis/terra+ aqua burned area monthly l3 global 500m sin grid v061. *NASA EOSDIS Land Process. DAAC*, 2021. 1

[18] Julia Gottfriedsen, Johanna Strebl, Max Berrendorf, Martin Langer, and Volker Tresp. Firesight: Short-term fire hazard prediction based on active fire remote sensing data. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[20] Fantine Huot, R Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 2

[21] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 3

[22] Sadegh Khanmohammadi, Mehrdad Arashpour, Emadaldin Mohammadi Golafshani, Miguel G Cruz,

Abbas Rajabifard, and Yu Bai. Prediction of wildfire rate of spread in grasslands using machine learning methods. *Environmental Modelling & Software*, 156:105507, 2022. 1

[23] Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, María Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17):e2022GL099368, 2022. 2

[24] Wouter M. Kouw and Marco Loog. An introduction to domain adaptation and transfer learning, 2019. arXiv:1812.11806 [cs]. 8

[25] Meelis Kull and Peter Flach. Patterns of dataset shift. 8

[26] Bronte Sihan Li and Ryan Rad. Wildfire spread prediction in north america using satellite imagery and vision transformer. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1536–1541. IEEE, 2024. 2, 3, 4

[27] Yanzhi Li, Keqiu Li, LI GUOHUI, Chanqing Ji, Lubo Wang, Die Zuo, Qing Guo, Feng Zhang, Manyu Wang, Di Lin, et al. Sim2real-fire: A multi-modal simulation dataset for forecast and backtracking of real-world forest fire. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[29] Shuwen Liu, Lin Cao, Chuanying Lin, Yuxuan Dai, Xingdong Li, Sanping Li, Shufa Sun, and Dandan Li. Fire spread prediction model based on multi-scale convolutional neural network. *Multimedia Tools and Applications*, pages 1–22, 2024. 1

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 1

[31] Mohammad Marjani, Masoud Mahdianpari, and Fariba Mohammadimanesh. Cnn-bilstm: A novel deep learning model for near-real-time daily wildfire spread prediction. *Remote Sensing*, 16(8):1467, 2024. 1

[32] Dimitrios Michail, Lefki-Ioanna Panagiotou, Charalampos Davalas, Ioannis Prapas, Spyros Kondylatos, Nikolaos Ioannis Bountos, and Ioannis Papoutsis. Seasonal fire prediction using spatio-temporal deep neural networks. *arXiv preprint arXiv:2404.06437*, 2024. 2

[33] MTBS Project, USDA Forest Service/U.S. Geological Survey. MTBS Data Access: Fire Level Geospatial Data, 2017. Last revised. 1

[34] Ozan Oktay. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1

[35] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 8

[36] Ioannis Prapas, Akanksha Ahuja, Spyros Kondylatos, Ilektra Karasante, Eleanna Panagiotou, Lazaro Alonso, Charalampos Davalas, Dimitrios Michail, Nuno Carvalhais, and Ioannis Papoutsis. Deep learning for global wildfire forecasting. *arXiv preprint arXiv:2211.00534*, 2022. 2

[37] Joaquin Quiñonero-Candela, editor. *Dataset shift in machine learning*. MIT Press, Cambridge, Mass, 2010. 8

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[39] William L Ross. Being the fire: A cnn-based reinforcement learning method to learn how fires behave beyond the limits of physics-based empirical models. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. 2

[40] Kamen Shah and Maria Pantoja. Wildfire spread prediction using attention mechanisms in u-net. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6. IEEE, 2023. 2, 3, 4

[41] Douglas Thomas, David Butry, Stanley Gilbert, David Webb, Juan Fung, et al. The costs and losses of wildfires. *NIST special publication*, 1215(11):1–72, 2017. 1

[42] Hongtao Xiao, Yingfang Zhu, Yurong Sun, Gui Zhang, and Zhiwei Gong. Wildfire spread prediction using attention mechanisms in u2-net. *Forests*, 15(10):1711, 2024. 2, 3, 4

[43] Zhengsen Xu, Jonathan Li, Sibo Cheng, Xue Rui, Yu Zhao, Hongjie He, and Linlin Xu. Wildfire risk prediction: A review. *arXiv preprint arXiv:2405.01607*, 2024. 2

[44] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2

[45] Yufei Zou, Mojtaba Sadeghi, Yaling Liu, Alexandra Puchko, Son Le, Yang Chen, Niels Andela, and Pierre Gentine. Attention-based wildland fire spread modeling using fire-tracking satellite observations. *Fire*, 6(8):289, 2023. 3, 4

Improved Next-Day Wildfire Spread Prediction and the WSTS+ Benchmark

Supplementary Material

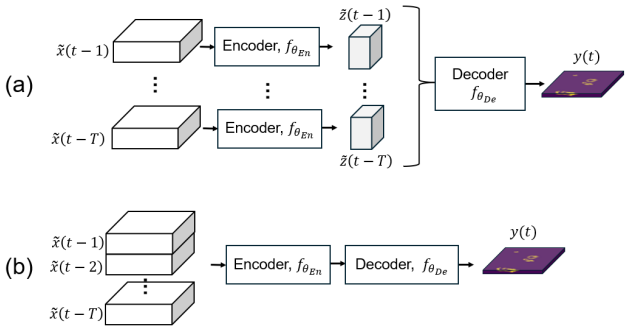

Figure 8. Illustration of (a) feature-level fusion, and (b) data-level fusion as we define it here. Further description is provided in the main text, and mathematical notation is described in Sec. 2

9. Experimental Details

SwinUnet SwinUnet [6] is a pure transformer-based Unet-shaped model that was first proposed for medical imagery segmentation. The model replaces the convolution blocks of the Unet with Swin Transformer blocks [30], including them throughout the encoder, bottleneck, and decoder. They also rely on patch merging and patch expansion layers in the encoder and decoder, respectively, to downsample the input features and then upsample the extracted features and produce the segmentation mask. Finally, they preserve skip connections to concatenate shallow and deep features. The SwinUnet outperformed the Unet [38], ViT [7], Att-Unet [34], and TransUnet [7] on two medical benchmark datasets. Its state-of-the-art performance, ability to learn both global and long-range dependencies, and use of the more efficient Swin blocks make it a good candidate for our task. Since the model was developed for RGB images, we modify the `in_chans` parameter to take in the number of channels of our multi-modal inputs (Veg: 7, Multi: 33, All: 40) instead of 3.

Model pre-training To evaluate the effect of pre-training on the SwinUnet model, we load the `swin-tiny-patch4-window7-224` weights from HuggingFace onto each of our Swin blocks. These weights correspond to a Swin Transformer trained on ImageNet at 224x224 resolution. We zero-pad our input images (128x128) to match the expected input dimensions and benefit from the pre-trained weights. As for the Unet models, we follow [15] and use the `segmentation_models_pytorch` implementation, and set `encoder_weights` to `imagenet`, which loads a model with ImageNet pre-trained weights. Finally, the UTAE pre-training uses the PASTIS weights, released with the original paper [14]. We use the 4th fold checkpoint, as it was the one with the highest performance.

Training details To train our models, we adopt the implementations shared by [15], which can be found in this GitHub repository. The implementation relies on PyTorch Lightning for model creation, training, and testing and Weights & Biases for model logging and metric visualization. All our models use a fixed batch size of 64, the AdamW optimizer, and a fixed optimized learning rate, as described in Sec. 6. Also following [15], we train our models for 10,000 iterations. Increasing the number of iterations to 15,000 and 20,000 did not yield any notable increases in performance. For all runs in Tab. 2, we report the mean test AP averaged over the 12 folds, and the standard deviation. During the hyperparameter search, we only use a single data fold (id = 2), train for 50 epochs, and pick the combination that yields the highest validation AP.

10. WSTS+ Details

10.1. Collection Details

To ensure our added wildfire events are most similar to the original ones, we follow the exact same collection procedure in [15]. Namely, we rely on the Google Earth Engine script found in this repository, to only collect wildfires that are larger than 10 km^2, and we use the GlobFire dataset [2] to identify wildfire events in the United States for 2016 and 2017. However, given GlobFire's temporal availability ends at 2021, we use the MTBS Burned Areas Boundaries Dataset [33] to identify wildfires in 2022 and 2023.

The main differences between both datasets used for fire event *identification* are that GlobFire relies on MODIS [17] as a data source, which has a resolution of 500 meters, while MTBS uses Landsat imagery, which has 30 meter resolution. Furthermore, GlobFire returns burned area maps with start and end dates, while MTBS returns fire perimeters with start dates only. Regardless, we only use the centroid coordinate for both area maps and perimeters to download the fire masks. To account for the lack of fire end dates in MTBS, we collect 30 days of samples after the start date, with an additionnal buffer of 4 days before and after the fire events, similar to [15]. We visualize the distribution of fire events in WSTS+ in Fig. 9.

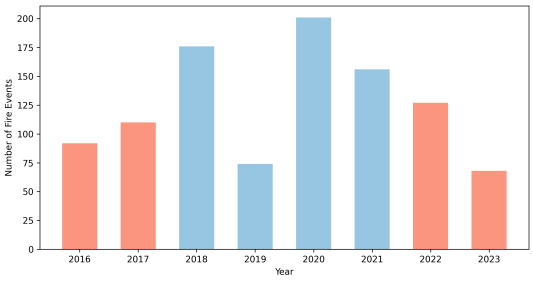

Figure 9. Distribution of fire events in WSTS+ per year

10.2. Quality Assurance

The new data were processed in the exact same way as the original WSTS data. To verify that it was done properly, we first replicated the downloading and processing of the original WSTS data (2018-2021), and measured the differences between our reproduction and the original data. We found that both are quantitatively similar. Specifically, we computed the mean pixel values of each data band for two folds (2018, 2019; and 2020, 2021) and found virtually no difference (max difference was 7.11e-04% of each other). Further, to ensure these differences were not meaningful, we trained a Res18-Unet with T=1 on the Multi feature set (the best performing one from Tab. 2) using our replicated WSTS and the original one. To verify that the results are similar, we show in Tab. 7 the test performance on each individual year and found that they are within a small numerical error of each other.

Upon collecting the additional data in WSTS+, we computed the means and variances of each explanatory feature (e.g., wind speed, humidity, NDVI, EVI2, ERC) as well as the active fire feature across both original years (2018-2021) and newly added ones (2016, 2017, 2022, and 2023), and found that the distributions suggest some distribution shift. Fig. 10 shows kernel density estimates of the yearly distributions of each explanatory feature in the dataset, highlighting varying degrees of cross-year domain shift across years. To validate this hypothesis, we conduct the cross-year experiments described in Sec. 11.

Table 7. Comparison of model performance on WSTS original data versus our replicated WSTS data.

Test Year	Original	Replicated
2018	0.49533	0.48594
2019	0.31190	0.32115
2020	0.42248	0.41793
2021	0.56742	0.56031
Average	0.44928	0.44633

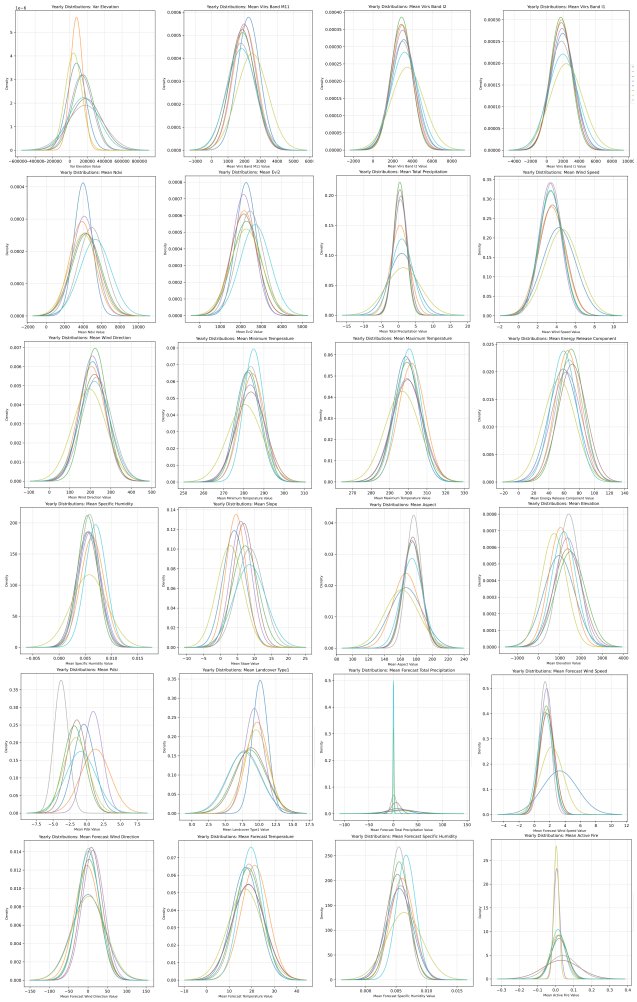

Figure 10. Smoothed probability density functions of the yearly distributions for all explanatory variables used in WSTS+. Each subplot represents a different variable, and the overlapping curves indicate variations in distribution across different years. These patterns reveal varying degrees of cross-year domain shift, with some features (e.g., elevation, temperature) exhibiting stable distributions, while others (e.g., wind speed, landcover type, forecast precipitation) show noticeable year-to-year variability.

11. Cross-year experimental design

We discuss here the experimental design of the results shown in Tab. 6 in the main manuscript. We trained a Res18Unet model on each training dataset listed in Tab. 8. Each year contributed a fixed quantity (and importance) of data samples (338 per year) to a shared validation set. We reached that number by reserving 20% of the data of the year with the least amount of samples (2019 had 1351 total samples) as validation and used that number for all other years, resulting in 2704 validation samples across 8 years, which represented between 8.25% and 16% of the total samples of the remaining 7 years. The training sets contain

$min(2000, |N|)$ where N is the total data available for that year, after removing the validation samples. This ensured the training sets across years had roughly the same amount of data to train on (all years ended up having 2000 samples, except for 2016 and 2019, with 1751 and 1002 samples, respectively). We then evaluated each model's average precision (AP) on each test set in Tab. 8. Notably, we ensured the training and validation sets for each year contain disjoint sets of fire events.

Table 8. Training, validation, and testing set sizes for each year, used for the cross-year train/testing.

Year	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Training Set Size	1385	1697	2000	692	2000	2000	2000	1818
Validation Set Size	2704							
Testing Set Size	338	338	338	338	338	338	338	338