Comparative Analysis of Ensemble and Deterministic Models for Fire Weather Index (FWI) System Forecasting

Shu Chen,[a] Piyush Jain,[b] Elizabeth Ramsey,[c] Jack Chen,[d] Mike Flannigan, [a]

[a] *Department of Natural Resource Science, Thompson Rivers University, Kamloops, British Columbia, Canada*

[b] *Northern Forestry Centre, Canadian Forest Service, Natural Resources Canada, Edmonton, Alberta, Canada*

[c] *Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada*

[d] *Air Quality Research Division, Environment and Climate Change Canada, Ottawa, Ontario, Canada*

Corresponding author: piyush.jain@nrcan-rncan.gc.ca

File generated with AMS Word template 2.0

Early Online Release: This preliminary version has been accepted for publication in *Weather and Forecasting*, may be fully cited, and has been assigned DOI 10.1175/WAF-D-25-0069.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Accurate fire weather forecasting is essential for effective wildfire management, particularly in regions increasingly affected by extreme fire activity such as British Columbia and Alberta, Canada. This study evaluates the predictive performance of three ensemble forecasting systems–the Ensemble Prediction System (ENS), the Global Ensemble Forecast System (GEFS), and the Canadian Global Ensemble Prediction System (GEPS)–and one deterministic model (High Resolution Forecast, HRES) –in forecasting components of the Canadian Fire Weather Index (FWI) System with 1-15 days lead time during the 2021-2023 wildfire seasons. Using ERA5 reanalysis as reference datasets, forecast skill was assessed using Mean Absolute Error (MAE), Continuous Ranked Probability Score (CRPS), and Precision-Recall Area Under the Curve (PR-AUC) metrics. Results show that ENS consistently demonstrates superior performance across all FWI components and weather inputs, with lower MAE and CRPS values across all the forecast lead times. A Super Ensemble combining all ensemble members from ENS, GEFS, and GEPS further improves long-range forecast reliability. Although deterministic forecasts outperform individual ensemble members, they are generally surpassed by ensemble-mean and ensemble-median forecasts at lead times greater than five days. The skill of deterministic forecasts also declines more rapidly with lead time and fails to quantify forecast uncertainty, despite their higher spatial resolution. These findings highlight the operational benefits of incorporating ensemble forecasts into fire management decision-making. This study also emphasizes the importance of overwintering adjustments and ensemble size in forecast skill and provides insights for improving fire weather prediction systems.

SIGNIFICANCE STATEMENT

Accurate fire danger forecasts support timely wildfire response and planning. This study evaluates the performance of three leading ensemble weather forecasting systems in predicting fire weather conditions across western Canada. It also compares the ensemble forecasts with the deterministic forecasts; the latter being more commonly used in operational fire management. The results show that ensemble-based fire weather forecasts can provide more reliable predictions, especially under high-risk conditions. By highlighting the strengths of ensemble systems, this work supports improvements in fire weather forecasting practices and helps inform operational decision-making in wildfire management.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

1. Introduction

The impact of wildfires in western North America has intensified in recent years, as area burned, number of large fires, and burn severity increase, posing significant threats to both the environment and society (Marlon et al. 2009; Barbero et al. 2015; Schoennagel et al. 2017; Hanes et al. 2019; Holsinger et al. 2022; Wang et al. 2025). In Canada, 2023 marked a record-breaking year, with almost 15 million hectares burned, surpassing the historical annual mean by over 7 times (Jain et al. 2024). Wildfires are influenced by numerous factors, including land use, vegetation, weather, topography, and human activities (Mermoz et al. 2005; Thompson and Spies 2009; Gralewicz et al. 2012; Pausas and Keeley 2021). Among these, daily weather conditions, such as high temperature, low humidity, strong winds and reduced precipitation, are critical factors that significantly influence the occurrence, spread, and impact of wildfires (Flannigan and Harrington 1988; Carvalho et al. 2008; Holden et al. 2018). Climate change has been a significant contributor to the increasing extent of wildfires (Flannigan and Van Wagner 1991; McKenzie et al. 2004; Barbero et al. 2015; Schoennagel et al. 2017), with higher temperatures being linked to increased wildfire occurrence and larger burned areas (Gillett et al. 2004; Flannigan et al. 2005; Balshi et al. 2009; Kirchmeier-Young et al. 2024).

Recognizing the critical role of weather conditions in wildfire behavior and the increasing impact of climate-driven extremes, fire weather forecasts are essential for predicting wildfire risks and developing effective management strategies. The Canadian Forest Fire Weather Index (FWI) System (Van Wagner, 1987) is one of the most widely used systems for assessing fire danger based on weather conditions. It evaluates fire potential by incorporating six components, comprising three fuel moisture codes: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), and Drought Code (DC), that describe the moisture content of different fuel layers; and three fire behavior indices: Initial Spread Index (ISI), Build-Up Index (BUI), and Fire Weather Index (FWI), that estimate the potential for fire ignition, spread, and intensity. Among these components, the FWI is an effective indicator of overall fire activity and is widely used for public fire danger alerts, while the moistures codes and fire behavior indices provide critical information in operational fire behavior predictions (Stocks et al. 1989). These components are calculated using daily noon weather inputs: temperature, relative humidity, wind speed, and 24-hour accumulated precipitation. While

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

originally developed for Canadian forests, the simplicity and effectiveness of the FWI System have made it a global standard for assessing fire danger (Di Giuseppe et al. 2020).

Although the FWI System provides a comprehensive framework for assessing fire danger, its effectiveness depends on the accuracy of the meteorological inputs. Advancements in numerical weather prediction (NWP) models have significantly enhanced fire weather forecasting, allowing longer-term fire weather forecasts (Mölders 2010). However, NWP models remain subject to two primary sources of error: initial condition errors and model errors. Initial condition errors stem from uncertainties and the limited availability of observational data, leading to inaccuracies in the state of the atmosphere. Model errors stem from the approximations used in representing complex atmospheric processes such as surface and boundary layer process, radiation and moisture processes (Buizza 2006). These uncertainties accumulate over time, leading to growing forecast errors, particularly at longer lead times, reducing the reliability of weather predictions.

NWP models are generally categorized into deterministic and ensemble models. Deterministic models provide a single forecast based on a specific set of initial conditions, offering the best-guess scenario for future weather at short lead times (Parker 2010). However, these models do not account for inherent atmospheric uncertainty and their accuracy declines rapidly as the forecast lead time increases. Beyond approximately 5 days, small errors in initial conditions can amplify due to the chaotic nature of the atmosphere, resulting in significant forecast uncertainty and reduced reliability (Buizza et al. 2005; Leutbecher and Palmer 2008). To address these limitations, ensemble models generate multiple forecasts by perturbing initial conditions and/or varying model parameterizations to capture a range of possible atmospheric trajectories (Gneiting and Raftery 2005). This probabilistic approach enhances representation of forecast uncertainty, particularly at extended lead times (Leutbecher & Palmer, 2008).

In the context of wildfire management, such probabilistic forecasting is especially important for end-users who need to assess the likelihood for extreme weather events and range of potential scenarios. Multi-day (e.g., 2-14 days) forecasts have been shown to enhance operational preparedness and planning. Ensemble forecasts offer more comprehensive framework for situational awareness, enabling more informed decision-making across short-, medium-, and long-range timescales (Boychuk et al. 2020). Furthermore, ensemble predictions at longer lead times extend forecasting capabilities beyond deterministic limits, improving multi-day preparedness and long-range risk assessment.

Previous studies have demonstrated the potential of ensemble NWP models in fire weather forecasting. Di Giuseppe et al. (2020) evaluated the performance of European Ensemble Prediction System (ENS) forecasts for FWI using both deterministic and probabilistic verification metrics, showing skill up to 10 days ahead. Similarly, Durão et al. (2022) found that ENS ensemble forecasts effectively captured extreme fire danger conditions in Portugal up to 72 hours prior to ignition events. In Canada, Boychuk et al. (2020) used the North American Ensemble Forecast System (NAEFS) to extend FWI forecasts up to 15 days in Ontario. Despite progress, comparative analyses of ensemble forecasts for fire weather, including evaluations of multiple models within the Canadian FWI System, are lacking. While there are several studies assessing ensemble model performance (Lin et al. 2016; Zhou et al. 2017; Richardson et al. 2020), comparisons specific to fire weather forecasts are crucial as the FWI System involves complex interactions among weather inputs and relies on previous moisture codes, which can accumulate forecast errors over time.

In this study, we evaluate and compare the predictive performance of three global ensemble NWP models in forecasting FWI components over British Columbia (BC) and Alberta (AB). These two Canadian provinces have recently experienced severe wildfire seasons driven by extreme weather conditions. For instance, BC's 2023 wildfire season saw over 2.84 million hectares burned, far exceeding the previous record of around 1.35 million hectares set in 2018 (Daniels et al. 2025). Similarly, Alberta's 2023 fire season surpassed historical records, with 2.2 million hectares burned, exceeding the previous record of around 1.4 million hectares set in 1981 (Hanes et al. 2019; Beverly and Schroeder 2025). These regions provide a good example for evaluating fire weather forecasting models due to their diverse landscapes, vulnerability to extreme fire weather conditions, and growing wildfire risk.

In addition to comparing ensemble model skill, this study also assesses whether ensemble forecasts outperform deterministic forecasts, the latter being more commonly used for predicting fire risk. By systematically evaluating ensemble forecast skill and uncertainty, this research aims to explore the potential advantages of ensemble models in the context of operational fire management.

2. Data

Multiple datasets were used in this study as inputs to the FWI System calculation, including NWP forecasts from ensemble and deterministic models, as well as a reanalysis dataset for verification. This study focuses on the period from April to September for the years 2021 to 2023. This choice of timeframe was driven by two primary considerations. First, the fire season in the study region predominantly occurs within these months, making it the most relevant period for evaluating fire weather forecasts. Second, the challenge of downloading and processing the large meteorological datasets constrains the study period. Additionally, the years 2021 and 2023 experienced particularly intense wildfire activity across BC and AB, further underscoring the importance of evaluating fire weather forecasts during these years. The study region covered the geographical range of 48°N-60°N, 109°W-139°W (as depicted in Fig. 6). The selected models provide global weather forecasts with varying spatial resolutions and lead times, supporting fire weather prediction and operational decision-making. Table 1 summarizes the key characteristics of these models and reanalysis, while the following subsections provide further details on data sources and processing.

Table 1. Numerical weather prediction models and reanalysis datasets used in this study.

Model	Repository	Resolution	Time Step	Lead Time	Members
ENS	MARS	0.5° × 0.5°	0–90h 1h, 93–144h 3h, 150–360h 6h	Up to 15 days	50
GEFS	AWS Open Data Registry	0.5° × 0.5°	0–240h 3h, 246–384h 6h	Up to 16 days	30
GEPS	CaSPAr-Globus web service	0.35° × 0.35°	0–384h 1h	Up to 16 days	20
HRES	MARS	0.1° × 0.1°	0–90h 1h, 93–144h 3h, 150–240h 6h	Up to 10 days	Single Forecast
ERA5	Copernicus Climate Data Store	0.25° × 0.25°	Hourly	Daily	-

a. Numerical Weather Prediction (NWP) Forecasts

The Ensemble forecasts used in this study were obtained from three major global ensemble forecasting systems: the Integrated Forecast System Ensemble Prediction System (ENS) by the European Centre for Medium-Range Weather Forecasts (ECMWF 2022), the Global Ensemble Forecast System (GEFS) by the National Centers for Environmental Prediction (NCEP 2024), and the Canadian Global Ensemble Prediction System (GEPS) by the Meteorological Service of Canada (MSC 2024). Each forecasting system provides probabilistic forecasts, representing a range of possible atmospheric trajectories. Additionally, deterministic forecasts were sourced from the Integrated Forecast System High Resolution Forecast (HRES) (ECMWF 2022). These forecasting systems were selected based on their operational reliability, global coverage, and applicability in fire weather forecasting.

The ENS consists of 50 ensemble members, each with slightly perturbed initial conditions and slightly altered model physics. This study used the "Set III - Atmospheric Model Ensemble 15-day Forecast (ENS)" dataset (ECWMF 2024). Data was retrieved from the ECMWF’s Meteorological Archival and Retrieval System (MARS). GEFS consists of 30 ensemble members that use different initial conditions and stochastic physics to capture forecast uncertainty (Zhou et al. 2022). GEFS data was accessed through the AWS Open Data Registry (NOAA 2024). GEPS forecasts are based on an ensemble of 20 perturbed weather forecasts that incorporate variations in initial conditions and model physics to represent forecast uncertainty. GEPS data was accessed from the Canadian Surface Prediction Archive (CaSPAr) (Mai et al. 2020). Note that GEPS data was unavailable for May 26, 2022. To ensure a fair comparison, this date was excluded from the corresponding analysis of ENS and GEFS forecasts. This exclusion does not affect the overall consistency, as the calculation for May 27, 2022, proceeded using ERA5 moisture codes from May 26. HRES is the highest-resolution operational model from ECMWF, offering a spatial resolution of 0.1°×0.1°. Unlike ENS, which provides a range of probabilistic forecasts through multiple ensemble members, HRES issues a single deterministic forecast, often referred to as a “best-guess” prediction, based on the most likely atmospheric state. In this study, the HRES dataset used is “Set I - Atmospheric Model high resolution 10-day forecast (HRES)” (ECMWF 2024).

b. Verification Data

The ERA5 reanalysis dataset is available hourly and a spatial resolution of approximately 31 km or 0.28125° on a reduced Gaussian grid (output at a regular latitude-longitude grid of 0.25°) (Hersbach et al. 2020). It provides a continuous and comprehensive global weather record, while the 4-D-Var data assimilation technique ensures that ERA5 are a physically consistent blend of observations and model data. While ERA5 reanalysis data may contain biases relative to direct observations, it offers a greater spatial and temporal coverage compared to in-situ station data and demonstrates good agreement with direct observations (Hersbach et al. 2020). Di Giuseppe et al. (2020) and McElhinny et al. (2020) also

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

demonstrated that ERA5 data are a good proxy for in-situ station observations in FWI calculations.

For this study, ERA5 data from 1991 to 2023 were obtained from Copernicus Climate Change Service's Climate Data Store (Copernicus Climate Change Service 2023). The 1991-2020 period was used to calculate the climatology of both meteorological inputs and FWI System components, while data from 2021-2023 served as the validation baseline for forecasting systems. ERA5 data served not only as a benchmark for evaluating NWP-based FWI forecasts but also as an initial condition for FWI System forecasts, facilitating comparison across the different forecasting systems.

Although ERA5 reanalysis data is used as the primary reference for forecast evaluation, we also include a supplementary validation using ground-based station observations to assess consistency (Appendix A).

3. Methods

a. Fire Weather Index (FWI) System

The Canadian Fire Weather Index (FWI) System is an empirical model that quantifies the effects of weather on forest fuel moisture and corresponding fire behavior (Van Wagner, 1987; Lawson & Armitage, 2008). The FWI System includes complex interactions between the daily weather inputs: temperature, relative humidity, wind speed and 24-hour precipitation. As illustrated in Fig. 1, the FWI System consists of six components: three moisture codes that assess the dryness of various fuel layers and three fire behavior indices that estimate fire spread potential and fuel availability. These components interact to provide an integrated assessment of fire weather conditions, which is essential for fire management decision-making.

The three fuel moisture codes represent different layers of fuel and their drying/wetting processes. Moisture codes for the current day are derived by integrating the day's weather inputs with the previous day's fuel moisture codes, thus tracking the drying and wetting of fuels (Flannigan et al. 2016). The Fine Fuel Moisture Code (FFMC) estimates the moisture content of surface forest floor, which dries quickly and affects fire ignition probability, with higher values indicating drier conditions and increased ignition potential (Wotton 2009). The Duff Moisture Code (DMC) represents moisture levels in moderately compact organic material beneath the surface, responding to weather over days to weeks. The calculation of

DMC uses a complex set of equations including the simulation of drying and wetting processes of the duff layer. High DMC values suggest increased fuel availability for combustion (Van Wagner, 1987). The Drought Code (DC) reflects deep, compact organic layers, reacting to seasonal precipitation trends rather than short-term weather fluctuations. Higher DC values indicate prolonged dry conditions, making deep-burning fires more likely (De Groot 1998). Each moisture code has differences in their water capacity and drying times under equilibrium conditions. Specifically, time lags represent the time it takes for each fuel layer to lose two-thirds of its free moisture content with normal weather conditions (temperature 21°C, relative humidity 45%). The drying time lags for the three fuel moisture codes are 2/3 of a day (FFMC), 12 days (DMC) and 53 days (DC), respectively (Van Wagner 1987; De Groot 1998; Lawson and Armitage 2008).

The three fire behavior indices are derived from the moisture codes and provide estimates of fire spread and intensity. The Initial Spread Index (ISI) quantifies the expected fire spread rate based on FFMC and wind speed, with higher values indicating faster-moving fires. The Build-Up Index (BUI) integrates DMC and DC to estimate the total fuel available for combustion, where dry conditions in both layers lead to a higher BUI. Finally, the Fire Weather Index (FWI) combines ISI and BUI to assess overall fire danger, making it a critical indicator for fire management operations. Higher FWI values correspond to increased potential fire intensity and greater difficulty in suppression efforts. For a more detailed explanation of the system's mathematical equations and its interpretation, refer to Van Wagner (1987) and Wotton (2009).

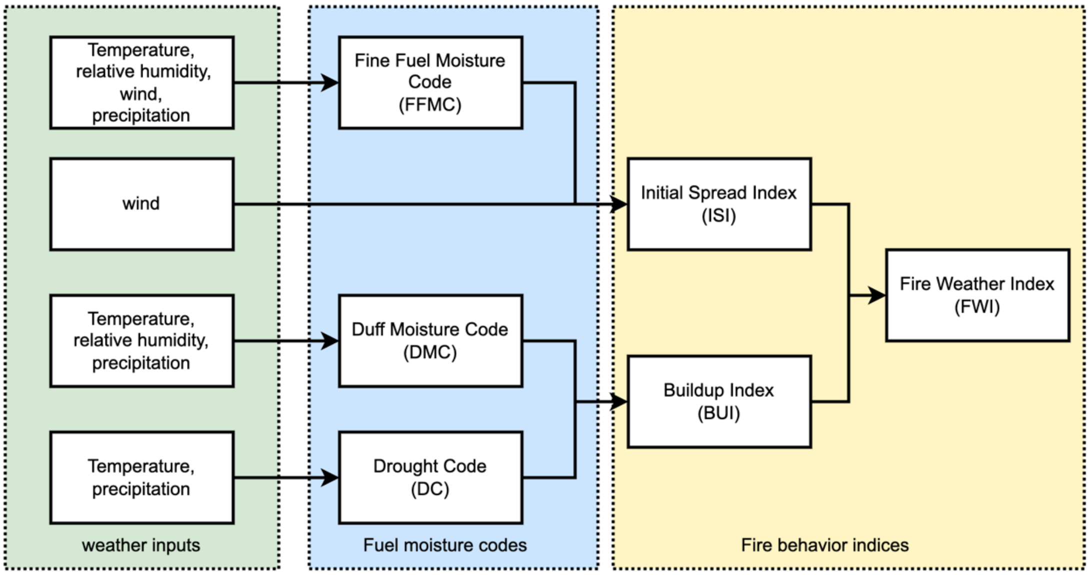

Temperature, relative humidity, wind, precipitation
wind
Temperature, relative humidity, precipitation
Temperature, precipitation
weather inputs
Fine Fuel Moisture Code (FFMC)
Duff Moisture Code (DMC)
Drought Code (DC)
Fuel moisture codes
Initial Spread Index (ISI)
Buildup Index (BUI)
Fire Weather Index (FWI)
Fire behavior indices

Fig. 1. Canadian Fire Weather Index (FWI) System. Adapted from Van Wagner (1987).

1) Pre-processing

Pre-processing of NWP outputs is required for use with the FWI System, which relies on local noon 2-meter temperature, 2-meter relative humidity, 10-meter wind speed values, previous 24-hour accumulated precipitation ending at local noon (Van Wagner, 1987; Lawson & Armitage, 2008).

Since FWI System components are calculated at 12:00 local time, while NWP forecasts are initialized in UTC, a conversion to local time was applied to the forecast data using time zone information obtained from a global shapefile dataset (Natural Earth, 2024) to get the required time offset from UTC for each location in the NWP spatial datasets. Additionally, we created a land mask from the ECMWF land-sea mask to indicate whether each grid point is over land or water (ECMWF 2024). We excluded grid points over water from further processing for computational efficiency.

In this study, Python packages xarray (Hoyer and Hamman 2017) and numpy (Harris et al. 2020), were used in processing the NWP meteorological inputs. All NWP meteorological inputs were loaded as four-dimensional data arrays: $\mathrm{N}_{\mathrm{member}} \times \mathrm{N}_{\mathrm{time}} \times \mathrm{N}_{\mathrm{lon}} \times \mathrm{N}_{\mathrm{lat}}$, where $\mathrm{N}_{\mathrm{member}}$ is the number of ensemble members (e.g., 50 for ENS), $\mathrm{N}_{\mathrm{time}}$ is the number of forecast time steps, $\mathrm{N}_{\mathrm{lon}}$ and $\mathrm{N}_{\mathrm{lat}}$ represent the number of longitude and latitude grid points over the study domain (e.g., 60 and 25 at 0.5° resolution for ENS, 300 and 120 at 0.1°

resolution for HRES), enabling efficient handling of the ensemble members, time steps, and spatial coordinates.

To align the original format of NWP variables with the FWI System's needs, we carried out the following preprocessing steps. Temperature, relative humidity and wind speed were derived or computed as required for the FWI System following standard meteorological procedures. For ENS and GEPS, relative humidity was not directly available and therefore derived from temperature and dew point temperature following the method of Alduchov and Eskridege (1996). Wind speed was calculated from 10-meter u and v wind components. Precipitation values were accumulated from previous local noon to today's local noon for each forecasted day. Maximum daily temperatures were also computed from the hourly air temperature for each forecasted day, where each day is defined by 0 to 23 hours local standard time for the local time zone. As discussed in section 3.a.2.ii, maximum daily temperature was used to calculate a proxy for snow on/off conditions for the FWI calculation as per Wotton and Flannigan (1993).

When NWP model outputs did not align with local noon for a given temporal resolution and time zone offset, it was necessary to first interpolate the data. For instance, we applied linear interpolation to convert the 3-hourly ENS time steps into hourly, ensuring the extraction of local noon values. We also applied the interpolation to hourly values for 6-hourly forecast periods. While such temporal interpolation introduces a potential source of forecast error, previous sensitivity analyses of weather elements in the FWI System (Lawson and Armitage 2008) have shown that moderate variations in temperature, humidity, wind speed, and precipitation typically result in only minor differences in FWI indices, which are not operationally significant. This indicates that the impact of our interpolation approach is likely negligible.

In addition to preprocessing NWP model outputs, ERA5 reanalysis data underwent similar preprocessing to ensure consistency in the subsequent FWI System calculations. However, no temporal interpolation was required as ERA5 provides data at an hourly resolution. After preprocessing, we merged the processed variables into a consolidated dataset, aligned temporally to correspond with the forecast days in local time.

2) FWI FORECAST

FWI System components were calculated for each ensemble member of the NWP forecast models following a structured workflow: initialization of moisture codes, application of overwintering adjustments, and daily calculation of FWI System components.

(i) Initialization of FWI System Moisture Codes

At the start of each forecast cycle, appropriate moisture code values were required to initialize the FWI System. Since moisture states from previous day's forecasts can introduce inconsistencies (e.g., one single poor forecast can propagate over time that leads to large errors), this study used ERA5-derived moisture codes and overwintering status to initialize the first forecast day (lead time = 1). These values were extracted from a continuous ERA5-based FWI computation (Section 3.a.4), ensuring a consistent baseline across forecasting systems.

For lead time ⩾ 2 days, each NWP model used the moisture codes from its previous lead time instead of ERA5 values (e.g., lead time = 3 used those from lead time = 2), ensuring internal forecast consistency while maintaining the ERA5-derived initialization for lead time = 1.

(ii) Overwintering Adjustments

To improve spring FWI initialization, we incorporated an overwintering adjustment based on McElhinny et al. (2020), a step often been overlooked in prior ensemble fire weather forecasting studies (Boychuk et al. 2020; Di Giuseppe et al. 2020; Durão et al. 2022). This adjustment used each model's daily maximum temperature to determine overwintering status, i.e., whether a given location is assumed to be snow-covered and therefore inactive with respect to fuel moisture updates in the FWI System (Lawson and Armitage 2008). The FWI System was activated after three consecutive snow-free days or daily maximum temperatures above 12°C and deactivated after three days of snow cover or maximum temperatures below 5°C (Van Wagner 1987; Wotton and Flannigan 1993). When a location transitioned out of overwintering, FFMC and DMC were reset to 85 and 6, respectively, representing moist spring conditions (Van Wagner 1987). DC was initialized using a precipitation-based adjustment that account for moisture carry-over from the previous fire season and accumulated winter precipitation (Lawson and Armitage 2008):

$$Q_s = a\, Q_f + b\left(3.94\, r_w\right) \tag{1}$$

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

here, Q_s is the starting spring moisture equivalent of DC value, Q_f is the final moisture equivalent of DC at the end of the previous fire season, r_w represents total accumulated winter precipitation, and two user-defined coefficients: a full carry-over fraction a (1.0) and a wetting efficiency b (0.75) (Van Wagner 1987; McElhinny et al. 2020). For locations still in overwintering, precipitation was accumulated for future DC recalibration, and all moisture codes remained unchanged.

(iii) Daily Computation of FWI System Components

Following initialization and overwintering adjustments, daily FWI System components were computed using the standard equations from Van Wagner (1987), based on each model's daily temperature, relative humidity, wind speed, and precipitation. These calculations were performed for each ensemble member using parallelized processing through Python's xarray and Dask (Rocklin 2015) libraries for efficient computation.

3) ERA5-DERIVED FWI SYSTEM COMPONENTS

The ERA5-derived FWI System components were calculated using the same equations as those applied to the NWP models, with the primary difference being the initialization process. Unlike the NWP models, which only used forecasts from April to September of year 2021 to 2023, the entire year was used to calculate FWI from the ERA5 dataset. To ensure temporal continuity, daily moisture codes and overwintering masks were initialized using values from the previous day. The overwintering adjustment followed the same procedure outlined in Sections 3.a.2. At the beginning of 2021, moisture codes and overwintering mask were initialized using the final values from the end of 2020.

b. Verification

1) SCORE METRICS

To assess the performance of NWP models in predicting FWI System components, we employ multiple verification metrics that capture both deterministic and probabilistic forecast skill. The evaluation focuses on three key aspects: deterministic accuracy, probabilistic reliability, and event-based predictive capability.

Mean Absolute Error (MAE) is used to quantify the deterministic accuracy of the forecasted FWI System components and their corresponding weather inputs. It is computed as:

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

$$MAE=\frac{1}{N_T\cdot N_S}\sum_{t=1}^{N_T}\sum_{s=1}^{N_s}\left|F_{t,s}-O_{t,s}\right| \tag{2}$$

where $F_{t,s}$ represents the forecasted value at a given time t and spatial location s, $O_{t,s}$ represents the observed value from ERA5 at the same location and time. N_T is the number of forecast initialization times (or forecast days), and N_S is the number of spatial grid points included in the evaluation. In this study, we compute MAE in two ways: (1) by averaging the MAE of individual ensemble members, where the MAE is calculated for each ensemble member against ERA5 across the all the grid points and forecast initialization days, and (2) by calculating the MAE of ensemble-mean/median-derived FWI System components before comparing against ERA5. The latter approach evaluates the overall predictive skill of the ensemble mean or median, while the former provides insight into the spread and variability within the ensemble.

To provide a baseline for forecast performance evaluation, we also calculated a climatological mean absolute error (MAE). For each calendar day and location, the long-term daily climatological mean for each FWI component and its meteorological inputs was derived from ERA5 data over the 1991-2020 period. For each forecast validation date during 2021-2023, this daily climatology was compared with the corresponding ERA5 verification value, and the absolute differences were averaged over all grid points and dates. This climatological MAE represents the typical error when using a static, climatology-based estimate instead of a dynamic forecast.

For the probabilistic forecast assessment, the continuous ranked probability score (CRPS) was employed as defined by Hersbach (2000). It is computed as:

$$CRPS = \frac{1}{N_T\cdot N_S}\sum_{t=1}^{N_T}\sum_{s=1}^{N_S}\int_{-\infty}^{\infty}\left[F_{t,s}(x)-H\left(x-O_{t,s}\right)\right]^2dx \tag{3}$$

where $F_{t,s}(x)$ is the cumulative distribution function of the ensemble forecast at time t and location s, and $O_{t,s}$ is the corresponding observed value from ERA5. The Heaviside function $H\left(x-O_{t,s}\right)$ equals 1 if $x\geq O_{t,s}$, and 0 otherwise. N_T and N_S denote the total number of forecast initialization times and spatial grid points, respectively. The overall CRPS was computed by averaging over all times and locations, consistent with the MAE approach. CRPS evaluates the probabilistic skill of ensemble forecasts by comparing the cumulative

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

probability distributions of forecasted values against observations. It accounts for both forecast accuracy and uncertainty by comparing the entire distribution of forecasted values (i.e., the ensemble spread) against the observed value. Unlike metrics that only evaluate central tendencies (e.g., the mean), CRPS penalizes forecasts that are either biased or overly dispersed, thus providing a single score that reflects both the closeness of forecasts to observations and the variance of the forecast distribution (Gneiting and Raftery 2007). Lower CRPS values indicate better probabilistic forecast performance, as it captures both forecast accuracy and reliability, where reliability refers to how well the ensemble forecast distribution represents the actual variability in observations (Hersbach 2000). In the special case of a single ensemble member, CRPS and MAE are mathematically equivalent. In this study, CRPS values from the ENS, GEFS and GEPS were compared to evaluate their forecasting capabilities across various lead times and FWI System components.

To assess the models' ability to predict extreme fire weather conditions, we applied Precision-Recall (PR) analysis to both ensemble and deterministic forecasts of FWI. The PR curve evaluates forecast performance by examining the trade-off between precision (the proportion of correctly predicted extreme events among all predicted extremes) and recall (the proportion of correctly predicted extreme events among all actual extremes). In our study, extreme fire weather is defined as FWI $\geq$ 19, this threshold being associated with high fire spread potential (Podur and Wotton 2011). The Area Under the Precision-Recall curve (PR-AUC) was used as the evaluation metric, where higher values (i.e., closer to 1) indicate better model skill in identifying extreme fire weather conditions while minimizing false alarms. We selected PR-AUC instead of more commonly used ROC-AUC (Hanley and McNeil 1982) because PR-AUC is more informative and sensitive in situations with imbalanced datasets. As highlighted by Saito and Rehmsmeier (2015), ROC curves can provide an overly optimistic assessment of model performance under class imbalance, whereas PR curves more directly reflect the performance on the minority class, making them more appropriate for evaluating rare events like high-risk fire weather days (in our study, only approximately 10% of all events had FWI$\geq$ 19).

2) VERIFICATION PREPARATION

To compute the verification metrics, we employed the Python package xskillscore (2024). To directly compare NWP FWI System forecasts with ERA5 FWI System components using xskillscore, we ensured consistency in both time and spatial scales. For intercomparison among

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

the three ensemble systems (ENS, GEFS, GEPS), the verification dataset ERA5 was bilinearly regridded to a common spatial resolution of 0.5° × 0.5° to ensure consistency. GEPS forecasts, originally provided on a rotated 0.35° grid, were first transformed to a geographic coordinate system and then regridded to 0.5° × 0.5°. For comparisons involving ensemble forecasts (ENS) versus the higher-resolution deterministic model (HRES), both ENS (i.e. 0.5°×0.5°,) and HRES (i.e. 0.1°×0.1°) were regridded to 0.25° × 0.25°, the same resolution as the ERA5 data used for validation, to maintain resolution consistency. All regridding was performed using xESMF with bilinear interpolation (Zhuang et al. 2022).

4. Results

This section presents an evaluation of ensemble and deterministic models in predicting FWI System components over BC and AB during the wildfire seasons from April to September in 2021-2023. Section 4a. compares the performance of the three ensemble systems: ENS, GEFS, and GEPS, highlighting differences in forecast quality across various FWI System components. Section 4b. focuses on a comparative analysis between the ensemble system ENS and the deterministic model HRES, assessing their relative strengths in FWI forecasting.

a. Inter-model Comparison across Different Ensemble Models

1) MEAN ABSOLUTE ERROR (MAE)

Fig. 2 and Fig. 3 illustrate the MAE of FWI System components and their corresponding weather inputs across forecast lead times for ENS, GEFS, and GEPS, with climatology-derived benchmarks included for comparison.

Across all FWI System components, ENS generally exhibits the lowest MAE, indicating superior predictive performance compared to GEFS and GEPS. This trend is broadly consistent across all lead times. A minor exception is observed for ensemble-mean-derived FFMC, where GEFS slightly outperforms ENS at lead times around days 10-11. The overall pattern is also reflected in the corresponding weather input performance (Fig. 3), where ENS generally outperforms both GEFS and GEPS in temperature, wind speed, relative humidity, and 24-hour accumulated precipitation. However, GEFS provides more accurate ensemble-mean temperature forecasts after day 10 and relative humidity forecasts around days 10-11, which explains its relative advantage in FFMC during that window. The relative performance of GEFS and GEPS varies across different FWI components. GEFS performs slightly better

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

than GEPS for FFMC, DMC, and BUI but shows inferior performance for ISI and FWI. Analyzing the weather inputs, GEPS outperforms GEFS in temperature and relative humidity across all lead times. While GEFS initially provides better wind speed forecasts, but GEPS surpasses it after lead time day 6. Since the FWI System is not a simple linear combination of weather inputs, the complex interactions among variables influence the relative performance of the models. Additionally, because ERA5 reanalysis fuel moisture conditions for FWI calculations were used to initialize the first forecast day for all three ensemble models, the differences in DC (which has a strong dependence on antecedent conditions) performance are minimal across all three models.

The ensemble-mean-derived FWI System components outperform individual ensemble members across all models and lead times. This aligns with the expectation that averaging across ensemble members smooth out individual forecast variability and reduces the impact of outlier predictions, leading to more accurate forecasts overall. As shown by the solid lines in Fig. 2 and Fig. 3, the MAE of ensemble-mean forecasts is consistently lower than that of individual members throughout the forecast lead times, demonstrating the advantage of using ensemble mean predictions for operational forecasting.

For most FWI System components, ensemble forecasts maintain predictive skill beyond 15 days, as their MAE remains lower than climatology. DC exhibits the most substantial advantage, with significantly lower MAE than climatology. This can be attributed to DC's long-term memory effect, as it is primarily driven by accumulated precipitation deficits over weeks to months. Unlike short-term components (e.g., FFMC, ISI), DC has a longer time-lag (53 days), meaning errors in daily weather inputs do not immediately result in large forecast deviations. Another possible contributing factor is that climatology is based on 1991-2020 conditions, whereas the model forecasts cover 2021-2023. Since 2021 and 2023 were characterized by drier and warmer conditions, the long-term climatology likely overestimates moisture availability. Consequently, model forecasts align better with the dry conditions at that time, leading to significantly lower MAE than climatology.

A notable exception is ISI, which does not converge toward climatology over longer lead times. This discrepancy also extends to FWI, since FWI is partially dependent on ISI. The likely explanation is that ISI is derived from both FFMC and wind speed, making it highly sensitive to short-term variations in weather inputs. Errors in either variable can propagate and amplify through ISI calculations over time, preventing MAE from stabilizing at climatological values beyond 7-10 days.

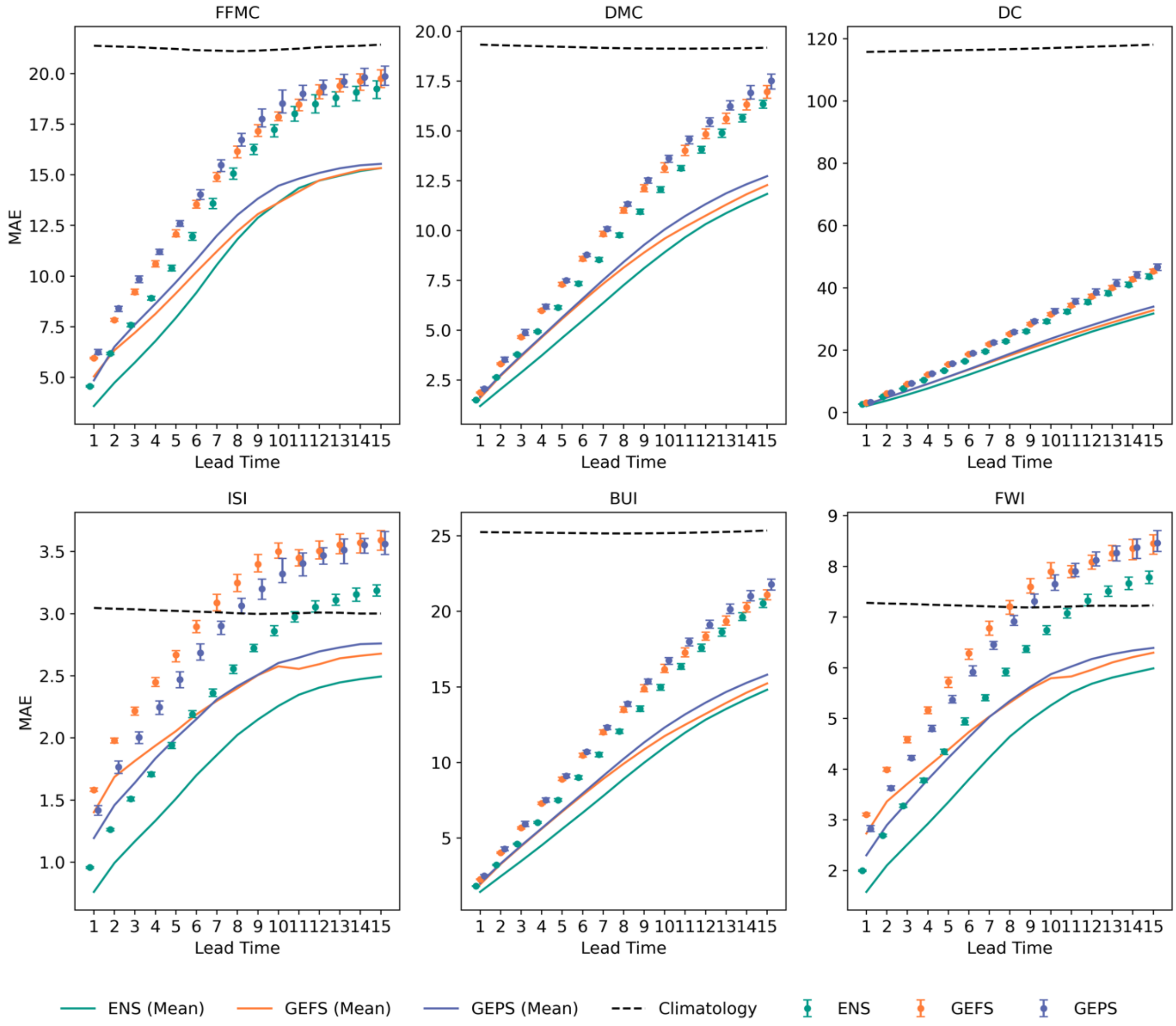


Fig. 2. Mean Absolute Error (MAE) of FWI System components for ENS, GEFS, and GEPS verified against those from ERA5, across forecast lead times (in day). MAE is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). Dots indicate the average MAE across individual ensemble members, error bars represent the 5th-95th percentile confidence intervals, solid lines denote MAE for ensemble-mean-derived FWI components, and dashed lines show the climatological MAE, calculated by comparing ERA5 daily climatological mean values (1991-2020) with the ERA5 verification data during 2021-2023.

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-25-0069.1.

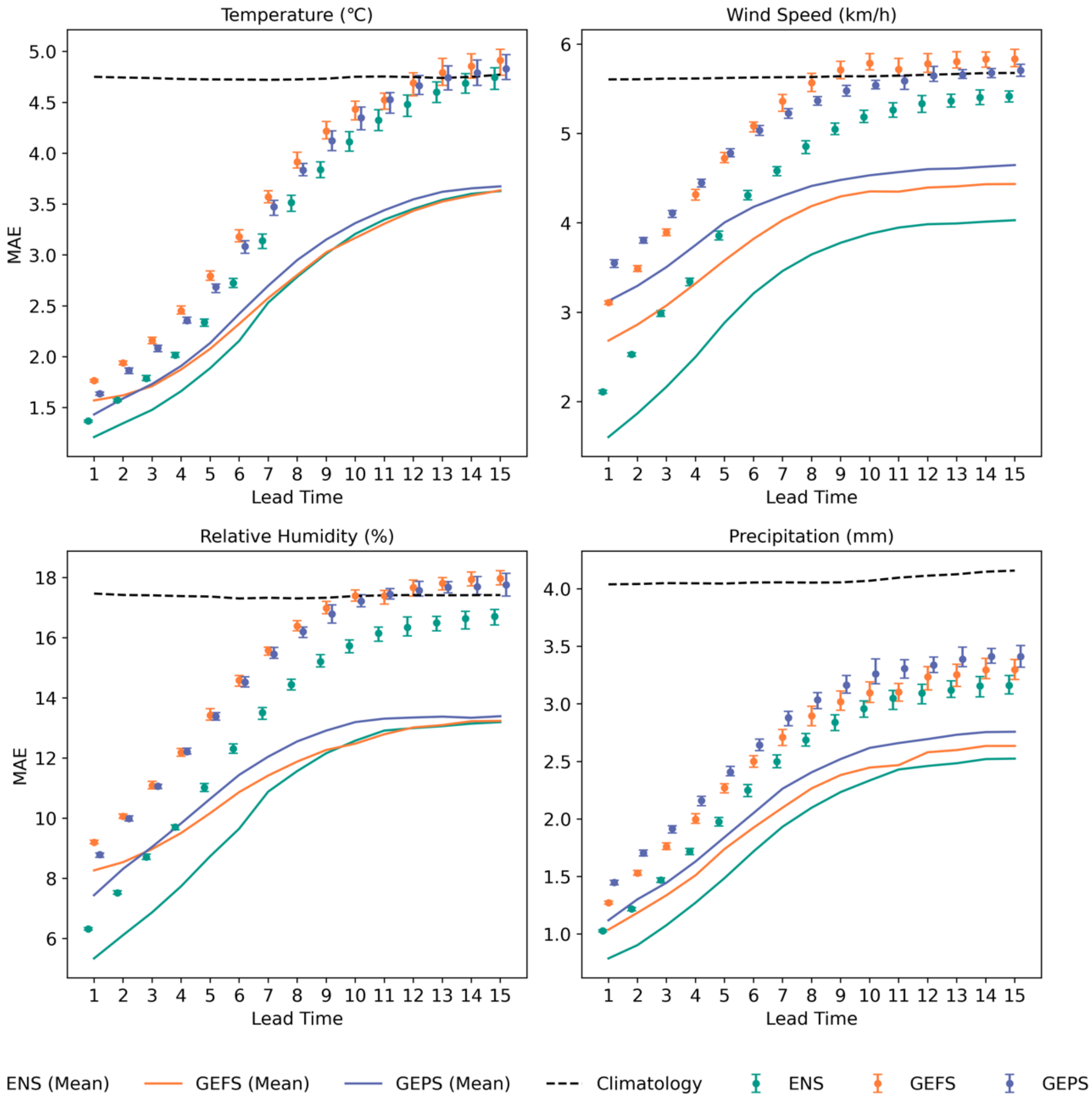


Fig. 3. Mean Absolute Error (MAE) of weather inputs for ENS, GEFS, and GEPS, verified against ERA5, across forecast lead times (in day). MAE is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). Dots indicate the average MAE across individual ensemble members, error bars represent the 5th-95th percentile confidence intervals, solid lines denote MAE for ensemble-mean-derived FWI components, and dashed lines show the climatological MAE, calculated by comparing ERA5 daily climatological mean values (1991-2020) with the ERA5 verification data during 2021-2023.

2) CONTINUOUS RANKED PROBABILITY SCORE (CRPS)

Fig. 4 and Fig. 5 illustrate CRPS values of FWI System components and the corresponding weather inputs for the three ensemble systems (ENS, GEFS, and GEPS), along with additional comparisons using reduced ensemble sizes and a “Super Ensemble” that combines all ensemble members from ENS, GEFS, and GEPS, for a total of 100 members. The inclusion of these comparisons allows us to evaluate both the overall probabilistic predictive skill of each ensemble and the impact of ensemble size on forecast reliability.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

As lead time increases, CRPS values rise, indicating both greater spread among ensemble members and increased deviations between forecast distributions and the observed values. Across all FWI System components, ENS consistently achieves the lowest CRPS values, demonstrating higher probabilistic skill than GEFS and GEPS. The relative performance between GEFS and GEPS varies by component and lead time: at shorter lead times (within 6 days), GEFS performs worse than GEPS for FFMC, ISI, and FWI, while their performance is comparable for DMC, DC, and BUI. However, at longer lead times, GEFS surpasses GEPS across all FWI System components.

ENS also maintains a clear advantages in all underlying weather input variables (Fig. 5). GEFS's poorer performance in temperature and relative humidity at short lead times likely contributes to its higher FFMC errors, which in turn affect ISI forecasts. Although GEFS outperforms GEPS in wind speed, this benefit does not translate into improved ISI performance due to the ISI's nonlinear and multiplicative sensitivity to both wind speed and FFMC. These results underscore the compounded influence of multiple weather inputs and error propagation within the FWI System.

To assess the impact of ensemble size, we compare subsets of ENS and GEFS using the first 20 ensemble members (i.e., members 0-19 by index), matching the size of GEPS. As expected, reducing ensemble size leads to an increase in CRPS, confirming that larger ensembles provide improved forecast reliability. However, the relative ranking of the models remains unchanged, suggesting that ensemble size affects absolute performance but does not alter their comparative ranking. The Super Ensemble (100 members) further demonstrates the benefits of larger ensembles, particularly at longer lead times (beyond 7 days). While its CRPS is comparable to ENS, GEFS, and GEPS in the short term, the advantages of the Super Ensemble become more pronounced at extended lead times (10-15 days), where forecast uncertainty grows. While the Super Ensemble generally improves forecast reliability, its performance for DC within the first 7 days is slightly worse than ENS. This is likely due to DC's long memory effect, where short-term variability in additional ensemble members introduces noise rather than enhancing predictive skill. In addition, while increasing ensemble size enhances forecast reliability, its practical implementation depends on computational resources and operational constraints. Further evaluation is required to determine the optimal trade-off between forecast accuracy and computational efficiency, particularly for operational fire weather prediction.

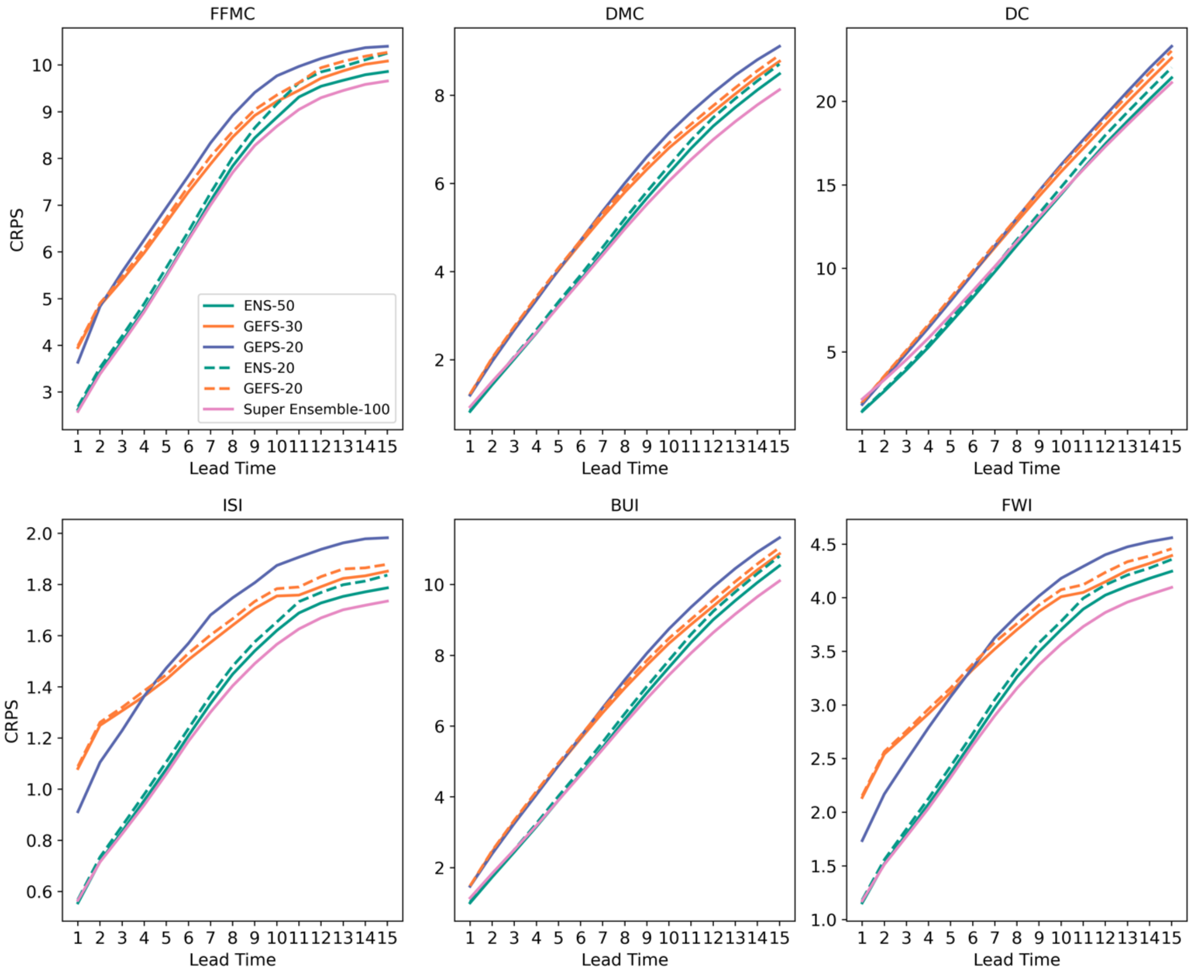


Fig. 4. Continuous Ranked Probability Score (CRPS) of FWI System components for ENS, GEFS, and GEPS across forecast lead times (in day), verified against ERA5. CRPS is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). Dashed lines indicate reduced ensemble subsets of ENS and GEFS (each with 20 members, matching GEPS), while the Super Ensemble (100 members) combines all ensemble members from ENS, GEFS, and GEPS.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

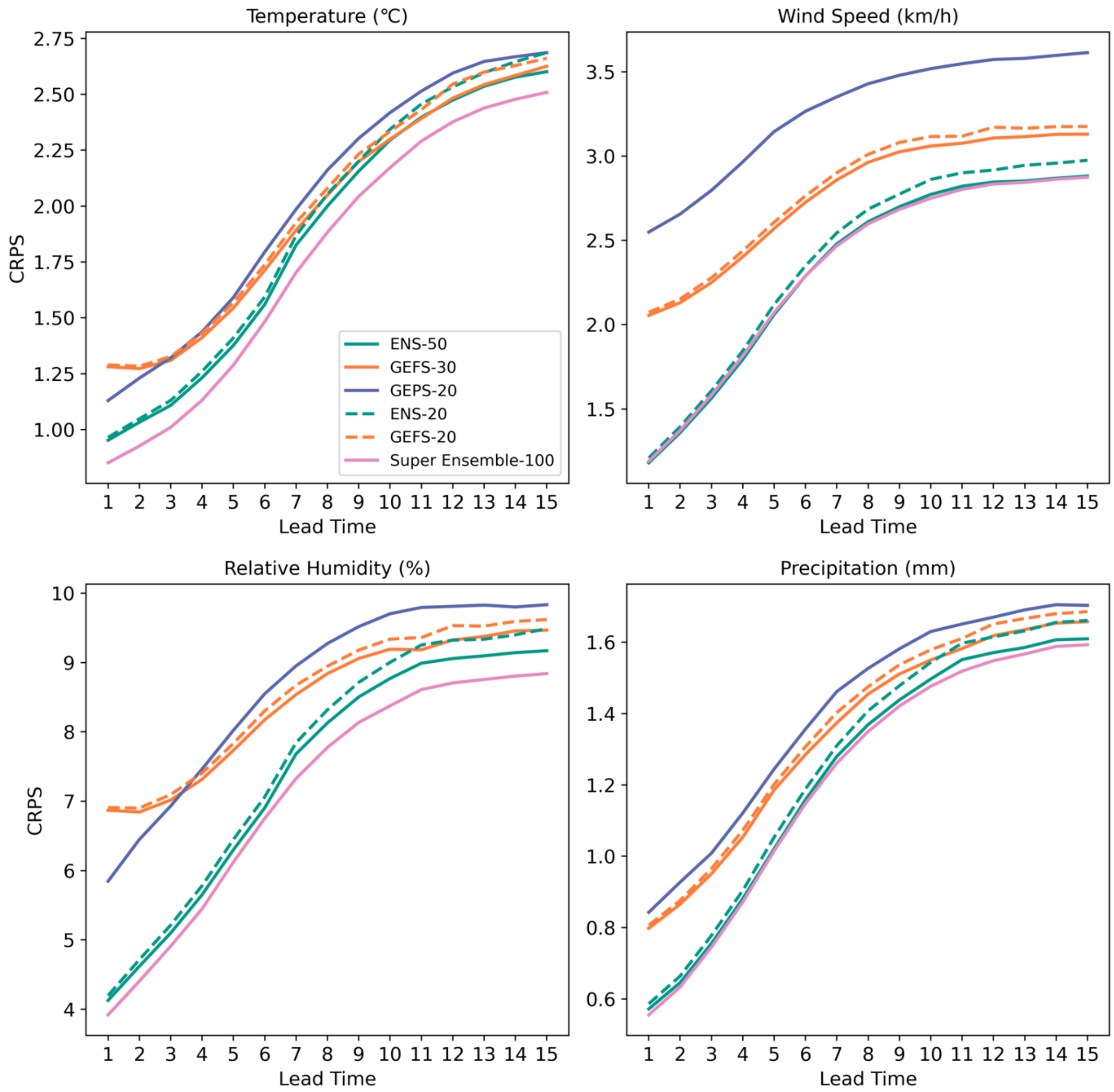


Fig. 5. Continuous Ranked Probability Score (CRPS) of weather inputs for ENS, GEFS, and GEPS across forecast lead times (in day), verified against ERA5. CRPS is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). Dashed lines indicate reduced ensemble subsets of ENS and GEFS (each with 20 members, matching GEPS), while the Super Ensemble (100 members) combines all ensemble members from ENS, GEFS, and GEPS.

To complement the numerical evaluation of probabilistic forecast skill, we also analyzed the spatial distribution of CRPS for the FWI across BC and AB at selected lead times (days 1, 3, 7, 10 and 15). The inclusion of spatial CRPS maps provides additional context for understanding the performance of ensemble models beyond aggregated statistical metrics. Fig. 6 presents the spatial maps of CRPS for FWI forecasts from ENS, GEFS, and GEPS, illustrating regional variations in forecast skills over time. FWI was selected because it is the most widely used fire danger indicator in the FWI System, providing an overall measure of fire potential based on weather conditions.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

The spatial distribution of CRPS highlights several key patterns. At short lead times (day 1 and day 3), CRPS values are generally lower across most regions, reflecting higher forecast skill. As the forecast lead time increases (day 7 to day 15), CRPS values increase across all regions, reflecting the expected decline in probabilistic forecast performance over time. The trend is particularly pronounced in areas of central to southern BC and southeastern AB, where forecast performance is lower. This trend can be attributed to the area's dry conditions, strong wind variability, and sparse observational data for model initialization. The sensitivity of FWI to these factors leads to larger ensemble spread and reduced alignment with observations, particularly at extended lead times.

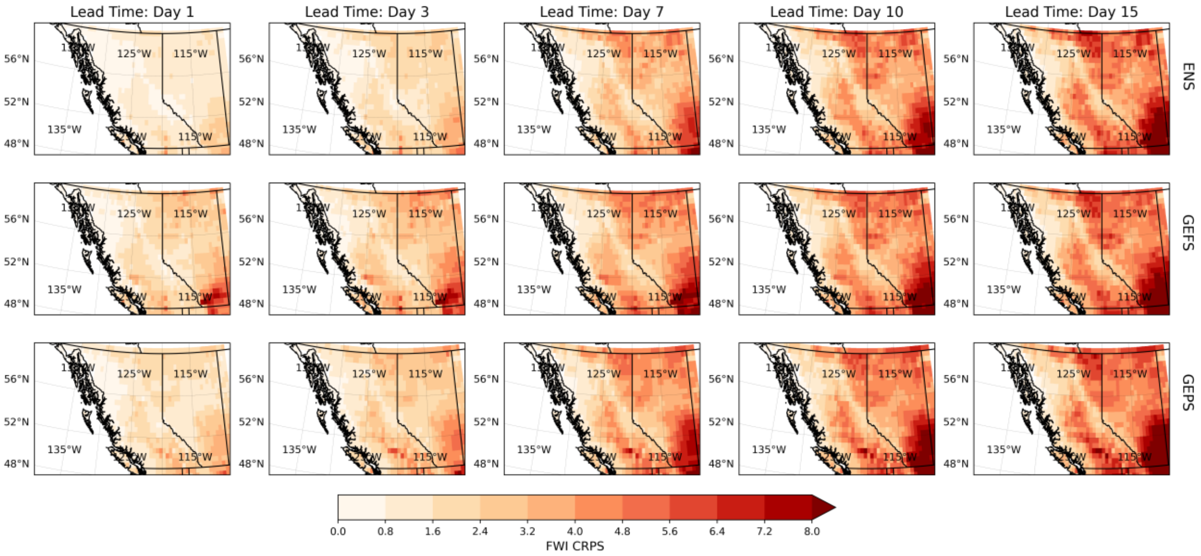


Fig. 6. Spatial distribution of Continuous Ranked Probability Score (CRPS) for FWI across BC and AB, evaluated for ENS, GEFS, and GEPS at selected lead times (Days 1, 3, 7, 10, and 15). CRPS values are averaged over the wildfire seasons (April-September 2021-2023), with darker shades indicating lower predictive skill.

b. Comparison of Ensemble and Deterministic Models

1) MEAN ABSOLUTE ERROR (MAE)

After we analyze the performance of the ensemble models verified against ERA5 reanalysis, we conclude that ENS demonstrates the best overall predictive performance across all FWI System components. Here we also assess whether a higher resolution deterministic model provides better accuracy than ensemble forecasts, at least at short lead times. To evaluate the differences between ensemble and deterministic models, we compare ECMWF ensemble forecasts (ENS) and deterministic forecasts (HRES). The first metric we consider is MAE. The analysis focused on all six FWI System components across 1-15-day lead times (Fig. 7). MAE was calculated for: (1) each ensemble member individually; (2) deterministic

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

forecasts (HRES); (3) ensemble-mean derived FWI System components (calculated after averaging each FWI System component across ensemble members); and (4) ensemble-median derived FWI System components (calculated after determining the median of each FWI System component across ensemble members).

In general, deterministic forecasts outperform individual ensemble members across all FWI System components. However, ensemble-mean and ensemble-median-derived forecasts perform better than HRES for most components. For short-term sensitive components (e.g., FFMC, ISI), ensemble-mean/median-derived values consistently outperform HRES across all lead times. For DMC, HRES initially performs better, but the ensemble mean/median surpass HRES beyond day 5. A similar trend is observed for DC, where HRES performs better for lead times less than 9-10 days. FWI, as an overall fire danger indicator, is better predicted by the ensemble mean/median across all lead times, despite ISI and BUI showing different relative performance patterns.

To better understand these results, we include the corresponding weather input variables (Fig. 8), which follows the same legend conventions as Fig. 7. Although HRES shows no clear advantage over the ensemble mean/median for any of the weather input variables, it still performs better for certain FWI components, specifically, DMC within the first 5 days and DC within the first 9-10 days. This may be related to how we initialize the models using ERA5. Since all models start from ERA5-derived values, the deterministic model may exhibit lower initial variability because it does not have ensemble spread. As a result, it may initially align more closely with the ERA5 verification data at short lead times, especially for components like DMC and DC that are less sensitive to daily fluctuations. However, at longer lead times, errors in HRES accumulate, leading to increased forecast deviations.

These findings suggest that while deterministic forecasts provide high-resolution predictions and outperform individual ensemble members, their usefulness in operational fire weather forecasting may be limited by their susceptibility to error propagation from single initial states and inherent variability in weather inputs. In contrast, ensemble-mean and -median-derived forecasts reduce such issues by averaging across multiple ensemble members, effectively smoothing out individual deviations and reducing the impact of input variability. As a result, ensemble-derived predictions offer more reliable and accurate guidance, especially at longer lead times, and may be more suitable for operational fire weather forecasting even when single deterministic values are required.

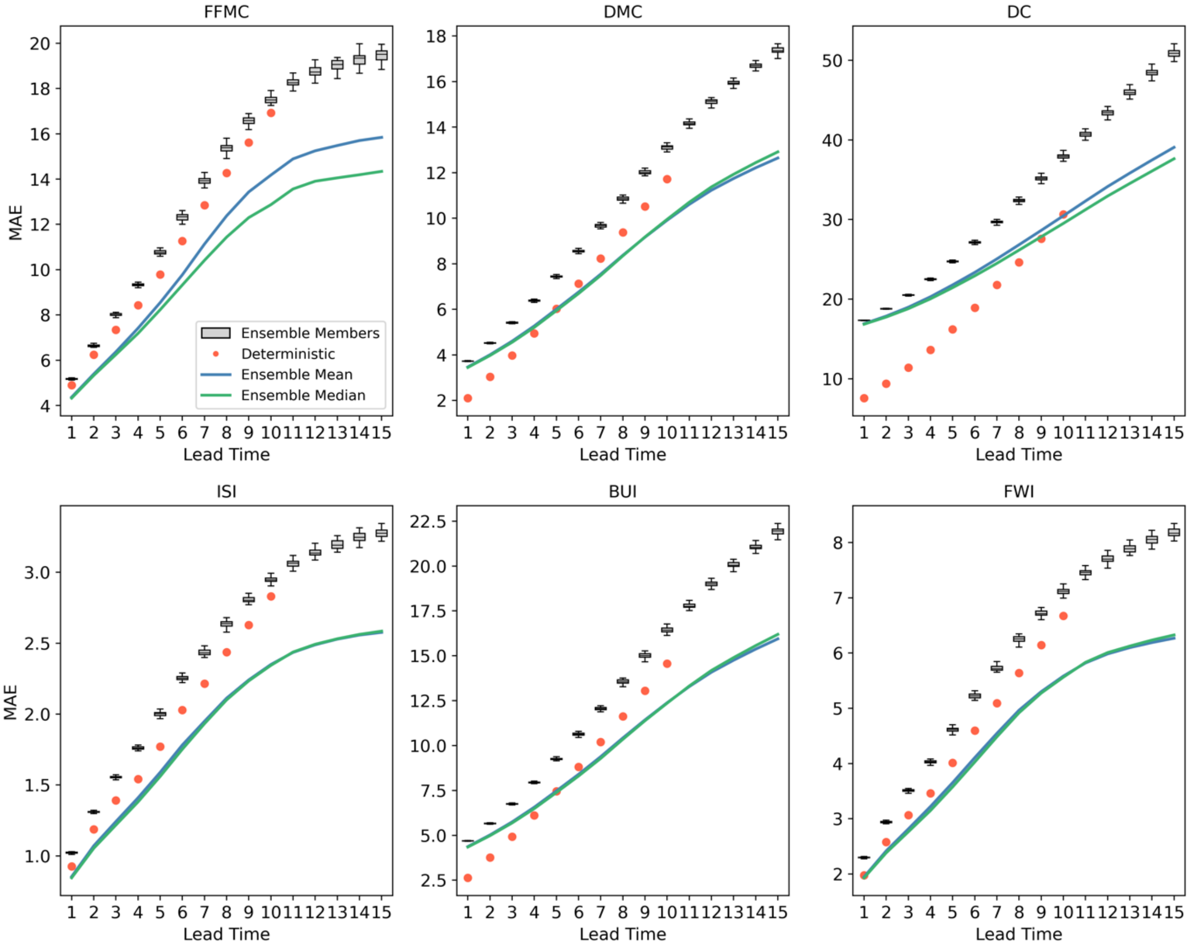


Fig. 7. Comparison of Mean Absolute Error (MAE) between ensemble and deterministic forecasts for FWI System components across lead times (in day). MAE is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). The boxplots represent the MAE values The boxplots represent the MAE values (max., 75th-, 50th-, 25th-percentile, min.) across all individual ensemble members for each forecast lead time. The dots indicate the deterministic (HRES) MAE. The solid lines correspond to the MAE of ensemble-mean-derived (blue) and ensemble-median-derived (green) FWI System components.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

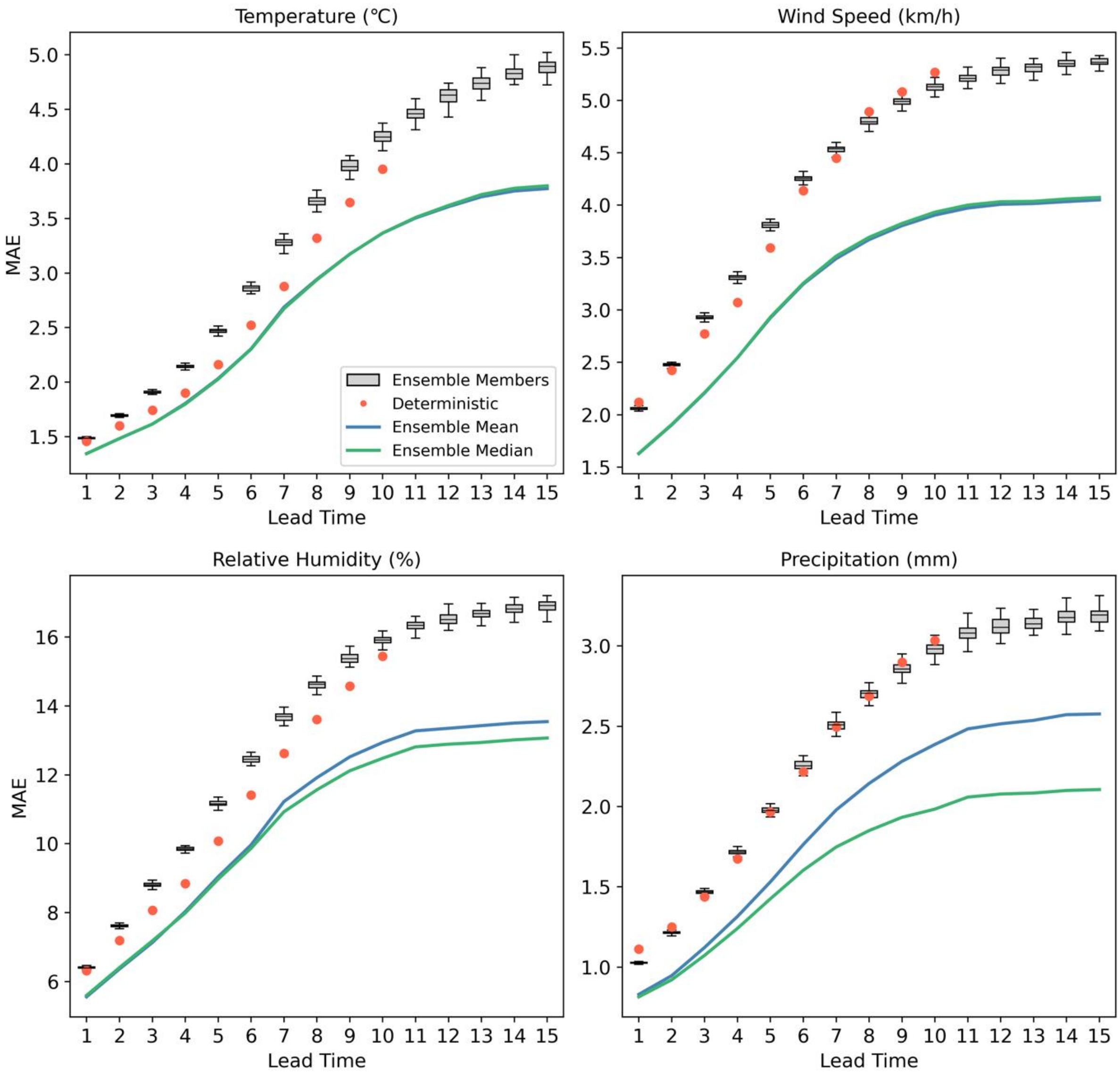


Fig. 8. Comparison of Mean Absolute Error (MAE) between ensemble and deterministic forecasts for weather inputs across lead times (in day). MAE is averaged over all grid points and forecast initialization times during the wildfire seasons (April-September 2021-2023). The boxplots (as described in Fig. 7) represent the MAE values across all individual ensemble members for each forecast lead time. The dots indicate the deterministic (HRES) MAE. The solid lines correspond to the MAE of ensemble-mean-derived (blue) and ensemble-median-derived (green) weather inputs.

2) PROBABILISTIC SKILL FOR HIGH-RISK FWI DAYS

From the previous analysis, we found that ensemble-mean-derived and ensemble-median-derived forecasts outperform single deterministic forecasts for the overall fire risk component FWI in terms of accuracy, with lower errors in the MAE metric evaluation, especially at longer lead time. In this section, we explore another key advantage of ensemble models: the ability to provide probabilistic forecasts, offering insights into forecast uncertainty that deterministic models lack.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

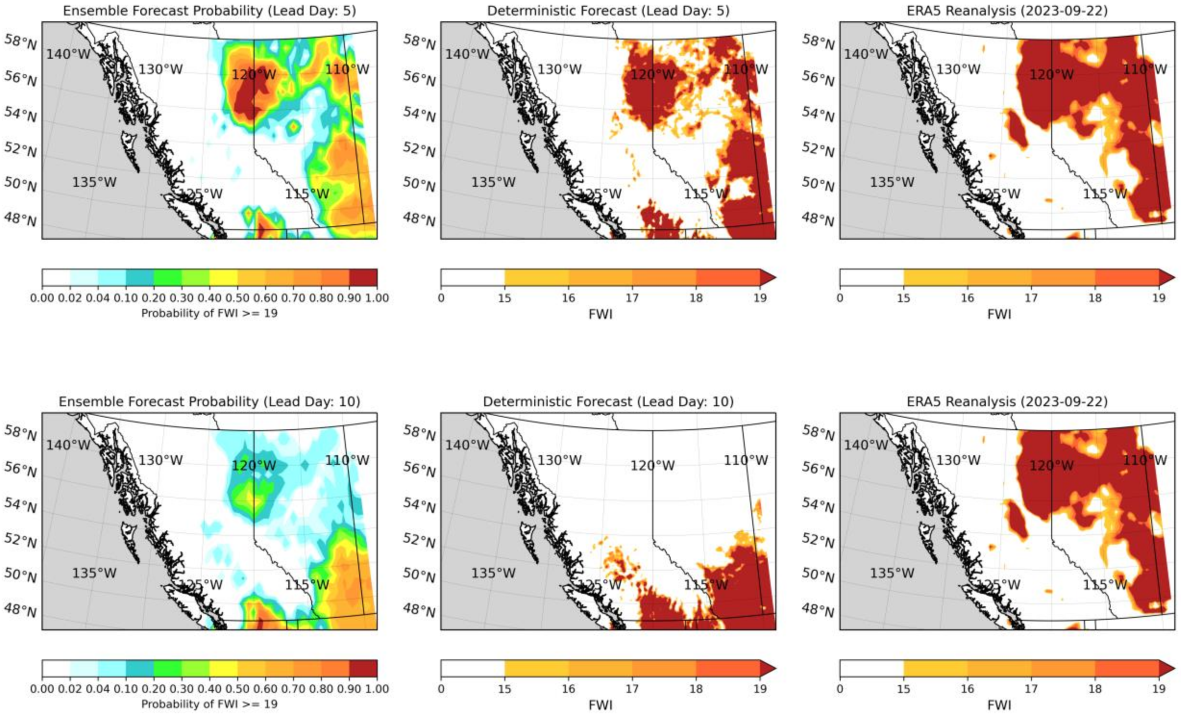
This comparison highlights the differences between ensemble probabilistic forecasts (ENS) and deterministic forecasts (HRES) for predicting extreme fire weather conditions (FWI $\geqslant$ 19) at different forecast lead times, using ERA5 reanalysis-derived FWI as the observational reference. Fig. 9 illustrates the spatial distribution of the ensemble forecast probability (left), deterministic forecast (middle), and ERA5 reanalysis (right) for the specified lead times (5 and 10 days). To illustrate the difference, we select September 22, 2023, a particularly active fire day during which over 400,000 hectares burned in a single day (Jain et al. 2024). Ensemble forecasts provide probabilities of extreme FWI, while the deterministic forecast and the ERA5 reanalysis represent specific FWI values.

At a lead time of day 5 (the upper subplots), both ensemble probabilistic forecast and deterministic forecast capture much of the spatial extent of high-risk regions. However, at a lead time of day 10 (the lower subplots), the differences between ensemble and deterministic forecasts become more pronounced. Ensemble probabilistic forecasts still capture high-risk areas with non-zero probabilities, but with reduced certainty. This decrease is expected due to the increasing uncertainty at longer lead times. In contrast, the deterministic forecasts fail to predict several regions of high risk entirely, especially around the BC and AB border in the vicinity of the Donnie Creek wildfire that burned through much of the 2023 fire season (Daniels et al. 2025). This underrepresentation shows the deterministic forecasts' limitations in capturing the spatial variability and uncertainty of extreme fire weather conditions at longer lead times.

To further quantify and evaluate the overall performance of ensemble and deterministic forecasts in predicting extreme fire weather conditions, we assess their ability to distinguish between high-risk and non-high-risk fire weather using the PR-AUC metric. We also include the other two ensemble models (GEFS and GEPS) into the comparison. Fig. 10 presents the PR-AUC values for ENS, GEFS, GEPS, and HRES across 10-15-day forecast lead times. At shorter lead times, deterministic forecasts (HRES) perform competitively with probabilistic forecasts, achieving high PR-AUC values (over 0.79 during the first 3 days). After lead time day 4, the gap between deterministic forecasts (HRES) and ensemble forecasts gradually becomes larger. The ensemble models maintain a consistent advantage, exhibiting higher PR-AUC values across all lead times, demonstrating their superior ability to identify extreme fire weather days (FWI $\geqslant$ 19) while minimizing false alarms. The PR-AUC analysis aligns with the spatial probability maps in Fig. 9, which demonstrates the advantages of ensemble

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

forecasts in capturing high-FWI areas with probabilistic certainty. While deterministic forecasts provide precise but single-outcome estimates, they fail to convey forecast uncertainty, leading to missed high-risk regions, particularly at longer lead times. Ensemble-based forecasts, however, provide decision-makers with probability distributions of extreme fire weather, supporting more effective wildfire preparedness and resource allocation strategies.



Ensemble Forecast Probability (Lead Day: 5)
Deterministic Forecast (Lead Day: 5)
ERA5 Reanalysis (2023-09-22)
Ensemble Forecast Probability (Lead Day: 10)
Deterministic Forecast (Lead Day: 10)
ERA5 Reanalysis (2023-09-22)
58°N
56°N
54°N
52°N
50°N
48°N
140°W
135°W
130°W
125°W
120°W
115°W
110°W
0.00 0.02 0.04 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Probability of FWI >= 19
0 15 16 17 18 19
FWI

Fig. 9. Spatial comparison of ensemble and deterministic forecasts of FWI for September 22, 2023. The left panel shows the probability of ensemble forecasts predicting FWI ≥ 19. Probabilities are calculated as the number of ensemble members forecasting the event divided by the total number of members (e.g., if 1 out of 50 members predicts FWI ≥ 19, the probability is 1/50 = 0.02). The middle panel presents the deterministic forecast, and the right panel displays ERA5 reanalysis. The top row corresponds to a 5-day lead time, while the bottom row corresponds to a 10-day lead time.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

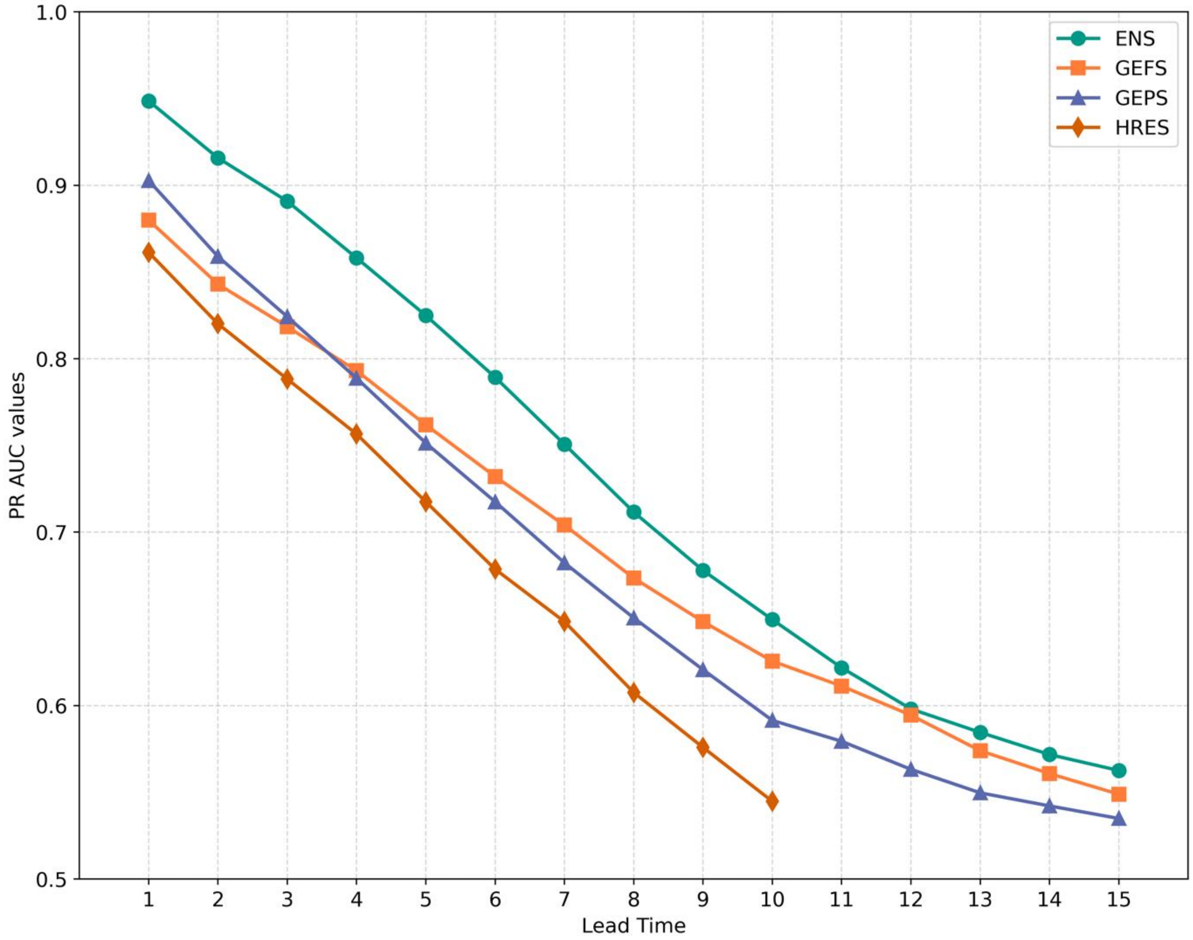


Fig. 10. Precision Recall (PR)-Area Under Curve (AUC) values for ENS, GEFS, GEPS, and HRES across forecast lead times (in day). PR-AUC evaluates the probabilistic skill in predicting high-risk fire weather conditions (FWI $\geq$ 19) compared to ERA5-derived observations.

5. Discussion and Conclusion

In this study, we evaluated the predictive performance of three ensemble forecasting systems (ENS, GEFS, and GEPS) and a deterministic model (HRES) for forecasting fire weather components over British Columbia and Alberta, Canada. Using ERA5 reanalysis data as the verification dataset, we assessed forecast quality through both deterministic and probabilistic metrics, including Mean Absolute Error (MAE), Continuous Ranked Probability Score (CRPS) and Precision Recall-Area Under Curve (PR-AUC).

For most FWI System components, ensemble forecasts retain predictive skill up to 15 days, particularly for long time-lag components like DC. In contrast, ISI remains sensitive to short-term variability in FFMC and wind speed, which in turn affects FWI and limits both indices from converging to climatology beyond lead time day 7.

Among the ensemble forecasting systems of ENS (50-member), GEFS (30-member) and GEPS (20-member), ENS consistently demonstrated superior performance across all FWI System components and weather inputs, as indicated by lower MAE and CRPS values. The ensemble size has influence on the forecast performance, with larger ensemble size generally showing better performance. However, even when we reduce the ensemble size to the same as GEPS's 20 members, the relative ranking among models remains unchanged. The Super Ensemble (100 members), which combines all ensemble members from ENS, GEFS, and GEPS, provided further incremental improvements, particularly at lead times beyond 7 days. This improvement may stem not only from increased ensemble size but also from enhanced member diversity across different modeling systems, consistent with the advantages of multi-core ensembles as discussed in Roberts et al. (2020).

While deterministic forecasts (HRES) provided higher resolution and better than individual ensemble member, they were outperformed by ensemble-mean and ensemble-median-derived forecasts for short-term sensitive FWI System components (e.g., FFMC, ISI and FWI) even at short lead times. A caveat is that HRES was regridded to match ERA5's resolution, which may introduce biases. The advantage of HRES for DMC ($\leq$ 5 days) and DC ($\leq$ 9-10 days) may be related to the ERA5 initialization.

Ensemble forecasting enhances muti-day predictability windows (i.e., short-, medium-, and long-term) by providing probabilistic forecasts rather than a single deterministic estimate. This allows fire managers to quantify uncertainty, improving decision-making at different timescales. Particularly over medium to long ranges, they provide confidence intervals around potential fire risk trends, reducing the bias from a single forecast outcome.

However, even though ensemble forecasts provide probabilistic information, operational decisions often require a single representative value, such as a fire danger index. In this context, ensemble-mean or -median-derived FWI forecasts offer a practical and more skillful alternative to high-resolution deterministic forecasts. Our results show that ensemble mean/median forecasts (ENS) outperform the deterministic forecasts (HRES) across most FWI System components, even at short lead times. This highlights the dual advantage of ensemble forecasts: they not only improve uncertainty quantification but also provide more accurate single-value guidance when required for operational decision-making.

One consideration for the evaluation of this study is the inherent connection between ENS and ERA5, as both originate from ECMWF modeling systems. Therefore, verification against

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

ERA5 could introduce biases, potentially making their results appear more closely aligned. As mentioned earlier, we also conducted a complementary validation using ground-based station observations (see Appendix A) to support the robustness of the evaluation. Although comparing point-based fire weather estimates with areal predictions remains challenging due to the uneven distribution of observation stations, the results show similar performance to the evaluation using ERA5 reanalysis.

Further research can enhance ensemble models by applying advanced post-processing methods, including bias correction and machine learning approaches, thereby improving predictive skill. Additionally, improving model resolution while maintaining ensemble forecasting capabilities could improve predictions in complex terrain and localized fire-prone regions. Notably, since June 2023, ECMWF has updated the resolution of ENS to 0.1°, which may further improve the performance of the ensemble models. This advancement also narrows the spatial resolution gap between ensemble and deterministic forecasts, reducing one of the primary advantages traditionally associated with deterministic models, though at the expense of significantly higher computational cost.

Acknowledgments.

This research was supported by the BC Wildfire Service. The authors are grateful to Dante Castellanos Acuna for assistance with the calculation of ERA5-derived FWI System components. We also thank Kate Bezooyen and Dave Schroeder for providing station observational data for British Columbia and Alberta, respectively. The authors declare no conflicts of interest.

Data Availability Statement.

The datasets used in this study are publicly available from their respective sources. NWP model forecasts, including ENS, GEFS, GEPS, and HRES, were accessed from ECMWF's Meteorological Archival and Retrieval System (MARS), the NOAA AWS Open Data Registry, and the Canadian Surface Prediction Archive (CaSPAr). ERA5 reanalysis data were obtained from the Copernicus Climate Data Store.

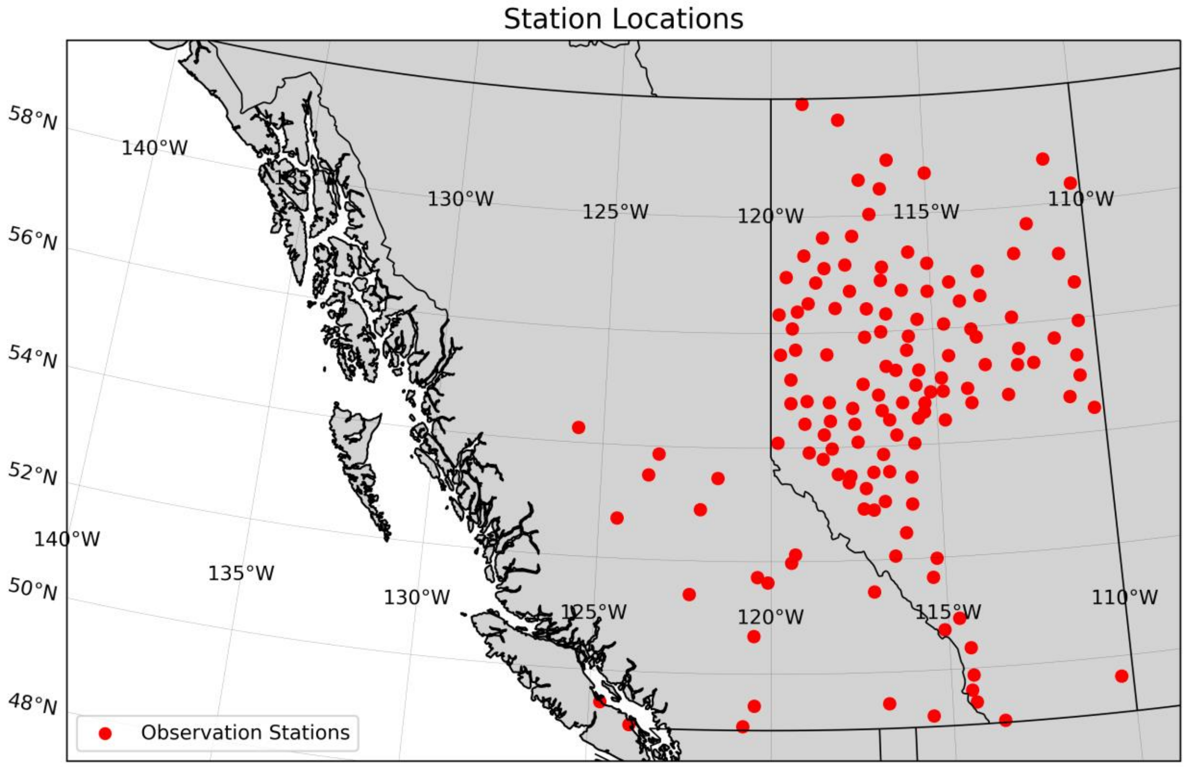
APPENDIX A

Validation of Station Observational Data

This supplementary analysis evaluates model performance using meteorological observations from 139 stations across BC and AB that maintained continuous records during April-September of 2021-2023. Only stations with complete data throughout the study period were included. The spatial distribution of these stations is shown in Fig. A1.

To ensure consistency with the primary analysis, we calculated mean absolute error (MAE) for the FWI System components and corresponding weather variables using station observations as the reference. Model forecasts were bilinearly interpolated to match station locations. All FWI System calculations were initialized using ERA5-derived moisture codes from the previous day, consistent with the initialization methodology applied in the gridded verification.

Figures A2-A3 present the inter-model comparisons across ensemble forecasts, while Figures A4-A5 compare ensemble and deterministic forecasts. Overall, the station-based evaluation broadly supports the findings from the gridded analysis, with similar model ranking and forecast performance patterns.



Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

Fig. A1. Locations of the 139 ground-based weather stations used for station-based verification across BC and AB. Only stations with continuous daily observations during the study period (April-September 2021-2023) were included.

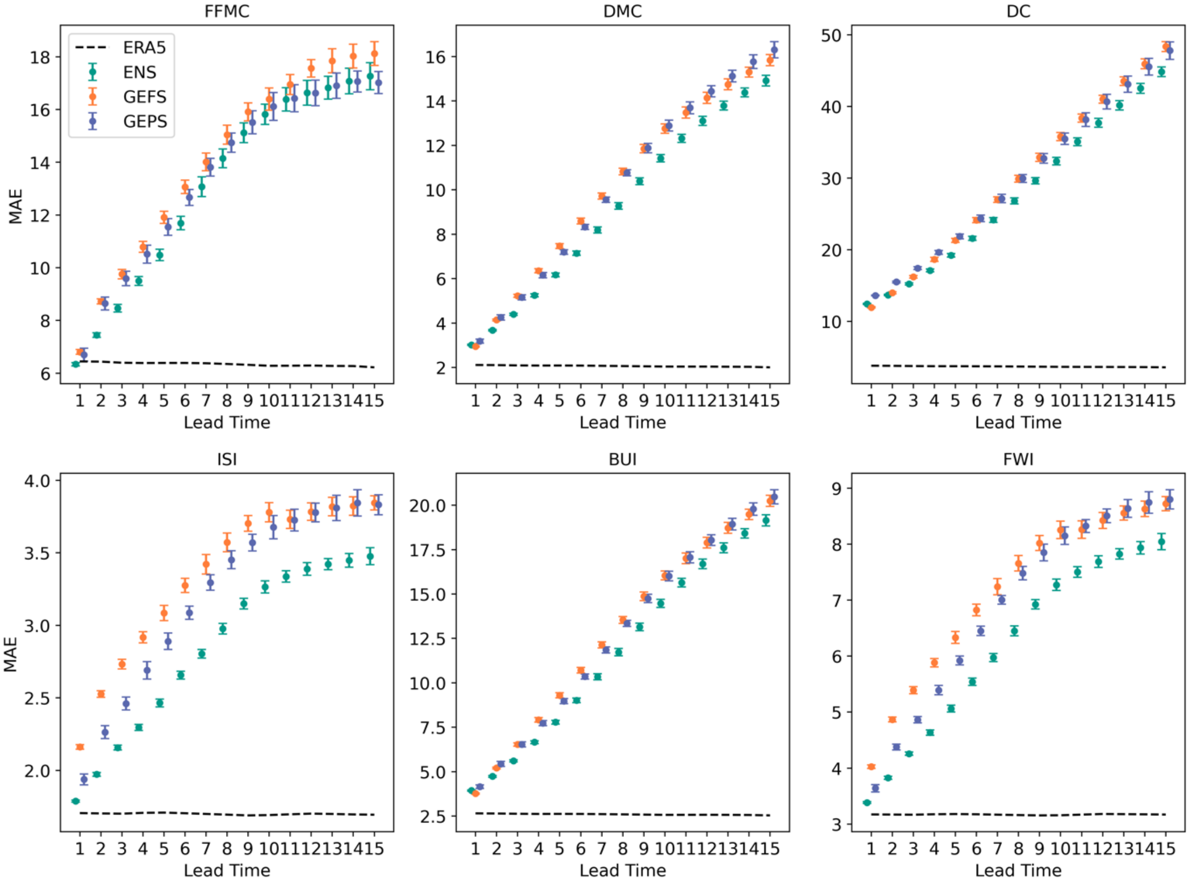


Fig. A2. Mean Absolute Error (MAE) of FWI System components for ENS, GEFS, and GEPS, verified against ground-based station observations across BC and AB during April-September 2021-2023. ERA5 reanalysis is included as a reference. Dots indicate the average MAE across ensemble members, error bars represent the 5th-95th percentile confidence intervals and dashed lines show the ERA5 MAE.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

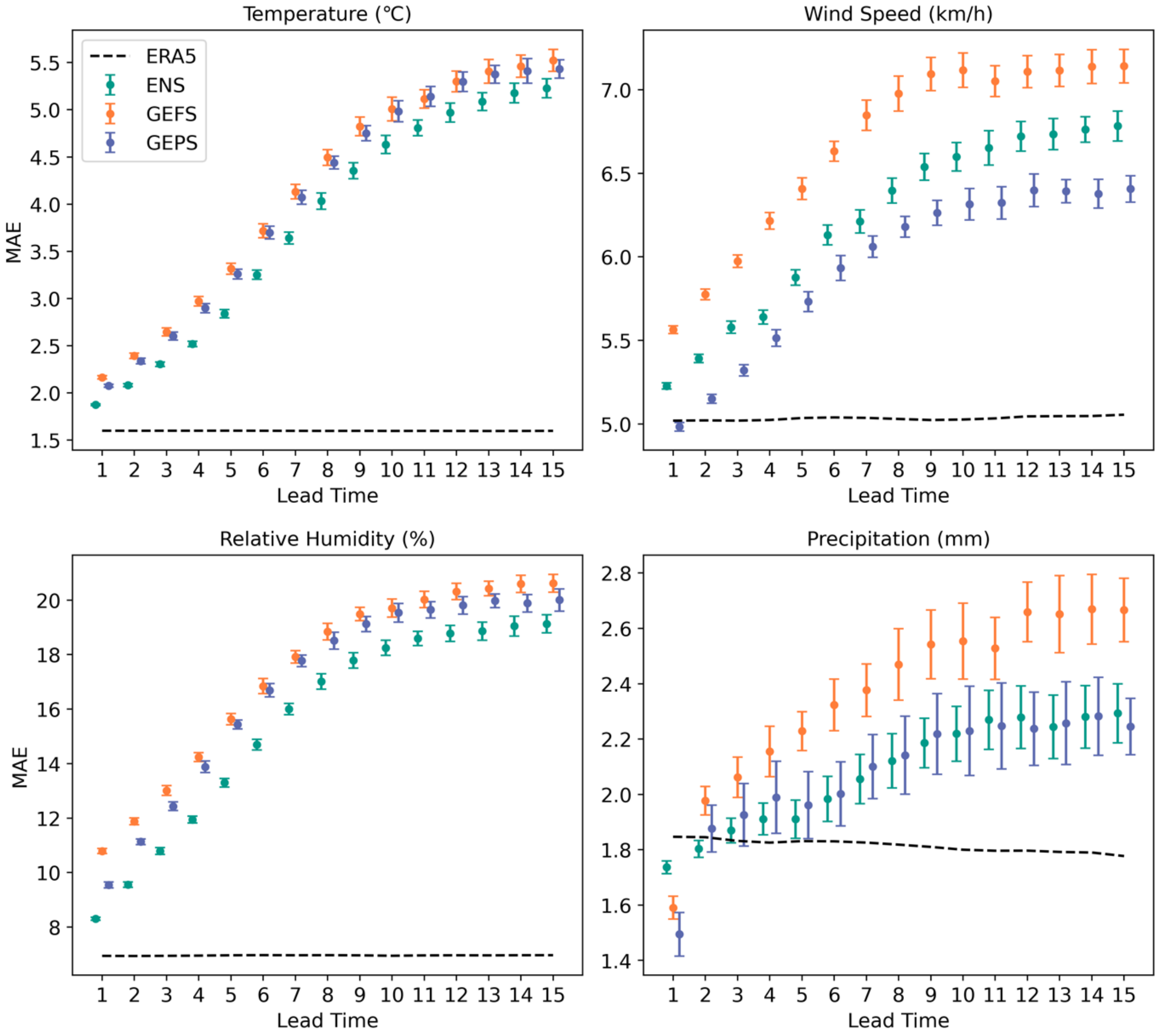


Fig. A3. Mean Absolute Error (MAE) of corresponding weather input variables for ENS, GEFS, and GEPS, verified against ground-based station observations across BC and AB during April-September 2021-2023. ERA5 reanalysis is included as a reference. Dots indicate the average MAE across ensemble members, error bars represent the 5th-95th percentile confidence intervals and dashed lines show the ERA5 MAE.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

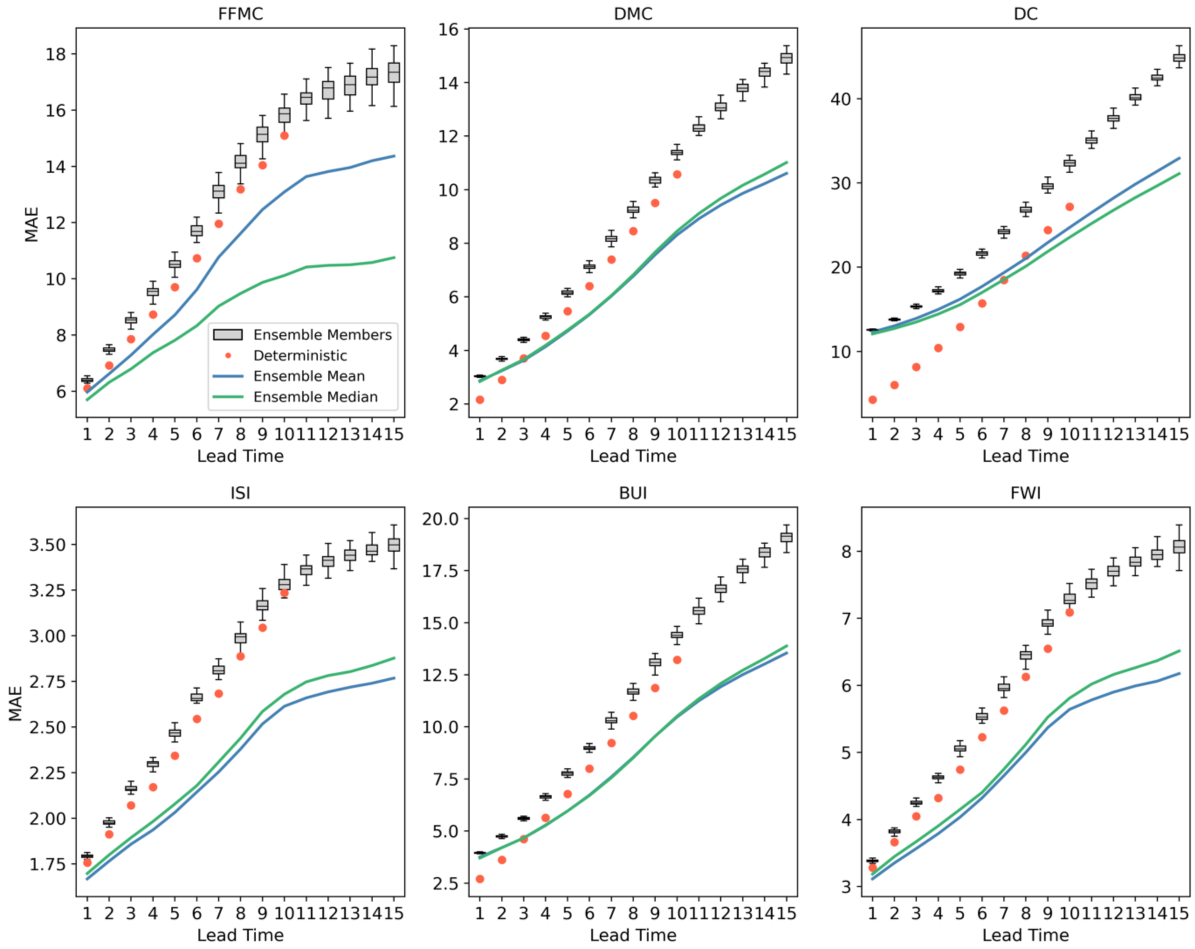


Fig. A4. Comparison of Mean Absolute Error (MAE) between ensemble (ENS) and deterministic (HRES) forecasts for FWI System components, verified against ground-based station observations across BC and AB during April-September, 2021-2023. The boxplots represent the distribution of MAE values across individual ensemble members, dots denote deterministic forecasts (HRES), and solid lines indicate ensemble-mean and ensemble-median-derived FWI components.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

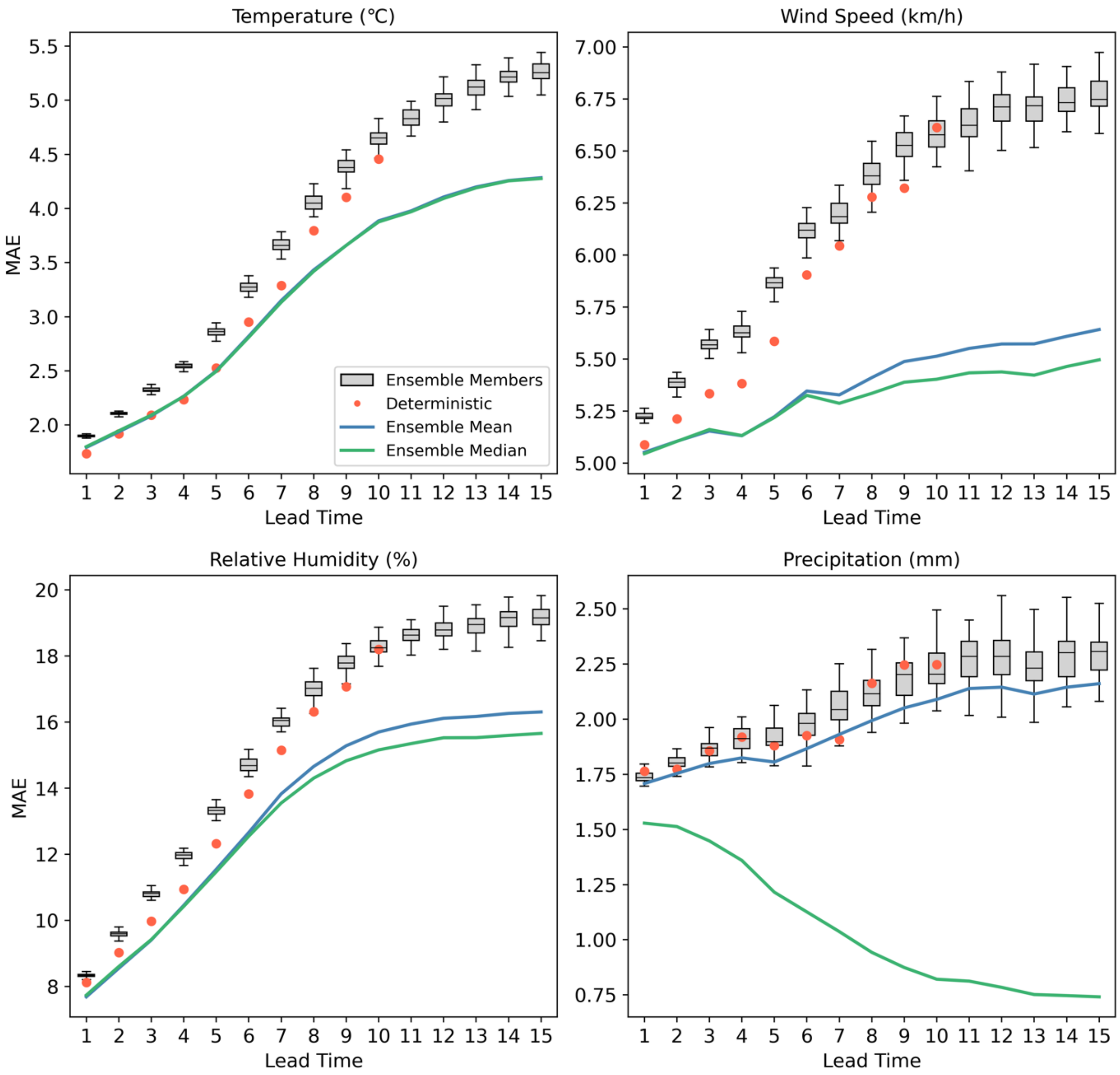


Fig. A5. Comparison of Mean Absolute Error (MAE) between ensemble (ENS) and deterministic (HRES) forecasts for corresponding weather input variables, verified against ground-based station observations across BC and AB during the wildfire seasons (April-September, 2021-2023). The boxplots represent the distribution of MAE values across individual ensemble members, red dots denote deterministic forecasts (HRES), and solid lines indicate ensemble-mean and ensemble-median-derived weather inputs.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

REFERENCES

Alduchov, O. A., and R. E. Eskridge, 1996: Improved Magnus Form Approximation of Saturation Vapor Pressure. *J. Appl. Meteorol.*, **35**, 601–609, https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2.

Balshi, M. S., A. D. McGUIRE, P. Duffy, M. Flannigan, J. Walsh, and J. Melillo, 2009: Assessing the response of area burned to changing climate in western boreal North America using a Multivariate Adaptive Regression Splines (MARS) approach. *Glob. Change Biol.*, **15**, 578–600, https://doi.org/10.1111/j.1365-2486.2008.01679.x.

Barbero, R., J. T. Abatzoglou, N. K. Larkin, C. A. Kolden, and B. Stocks, 2015: Climate change presents increased potential for very large fires in the contiguous United States. *Int. J. Wildland Fire*, **24**, 892, https://doi.org/10.1071/WF15083.

Beverly, J. L., and D. Schroeder, 2025: Alberta's 2023 wildfires: context, factors, and futures. *Can. J. For. Res.*, **55**, 1–19, https://doi.org/10.1139/cjfr-2024-0099.

Boychuk, D., and Coauthors, 2020: Assembling and Customizing Multiple Fire Weather Forecasts for Burn Probability and Other Fire Management Applications in Ontario, Canada. *Fire*, **3**, 16, https://doi.org/10.3390/fire3020016.

Buizza, R., 2006: The ECMWF ensemble prediction system. *Predictability of Weather and Climate*, Vol. 459 of, Cambridge University Press, p. 488.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Mon. Weather Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Carvalho, A., M. D. Flannigan, K. Logan, A. I. Miranda, and C. Borrego, 2008: Fire activity in Portugal and its relationship to weather and the Canadian Fire Weather Index System. *Int. J. Wildland Fire*, **17**, 328, https://doi.org/10.1071/WF07014.

Copernicus Climate Change Service, 2023: ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 18 February 2025, https://doi.org/10.24381/CDS.BD0915C6.

Daniels, L. D., and Coauthors, 2025: The 2023 wildfires in British Columbia, Canada: impacts, drivers, and transformations to coexist with wildfire. *Can. J. For. Res.*, **55**, 1–18, https://doi.org/10.1139/cjfr-2024-0092.

De Groot, W. J., 1998: Interpreting the Canadian forest fire weather index (FWI) system. *Proc. of the Fourth Central Region Fire Weather Committee Scientific and Technical Seminar*, 8.

Di Giuseppe, F., C. Vitolo, B. Krzeminski, C. Barnard, P. Maciel, and J. San-Miguel, 2020: Fire Weather Index: the skill provided by the European Centre for Medium-Range Weather Forecasts ensemble prediction system. *Nat. Hazards Earth Syst. Sci.*, **20**, 2365–2378, https://doi.org/10.5194/nhess-20-2365-2020.

Durão, R., C. Alonso, and C. Gouveia, 2022: The Performance of ECMWF Ensemble Prediction System for European Extreme Fires: Portugal/Monchique in 2018. *Atmosphere*, **13**, 1239, https://doi.org/10.3390/atmos13081239.

ECMWF, 2022: Medium-range forecasts. European Centre for Medium-Range Weather Forecasts, accessed 19 February 2025, https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts.

ECMWF, 2024: Land-Sea Mask. European Centre for Medium-Range Weather Forecasts, accessed 29 July 2025, https://confluence.ecmwf.int/display/FUG/Section+2.1.3.1+Land-Sea+Mask.

ECMWF, 2024: Set I - Atmospheric Model Ensemble 10-day Forecast (HRES), accessed 19 February 2025, https://www.ecmwf.int/en/forecasts/datasets/set-i.

ECWMF, 2024: Set III - Atmospheric Model Ensemble 15-day Forecast (ENS), accessed 18 February 2025, https://www.ecmwf.int/en/forecasts/datasets/set-i.

Flannigan, M. D., and J. B. Harrington, 1988: A Study of the Relation of Meteorological Variables to Monthly Provincial Area Burned by Wildfire in Canada (1953–80). *J. Appl. Meteorol.*, **27**, 441–452, https://doi.org/10.1175/1520-0450(1988)027<0441:ASOTRO>2.0.CO;2.

Flannigan, M. D., and C. E. Van Wagner, 1991: Climate change and wildfire in Canada. *Can. J. For. Res.*, **21**, 66–72, https://doi.org/10.1139/x91-010.

Flannigan, M. D., K. A. Logan, B. D. Amiro, W. R. Skinner, and B. J. Stocks, 2005: Future Area Burned in Canada. *Clim. Change,* **72**, 1–16, https://doi.org/10.1007/s10584-005-5935-y.

Flannigan, M. D., B. M. Wotton, G. A. Marshall, W. J. De Groot, J. Johnston, N. Jurko, and A. S. Cantin, 2016: Fuel moisture sensitivity to temperature and precipitation: climate change implications. *Clim. Change*, **134**, 59–71, https://doi.org/10.1007/s10584-015-1521-0.

Gillett, N. P., A. J. Weaver, F. W. Zwiers, and M. D. Flannigan, 2004: Detecting the effect of climate change on Canadian forest fires. *Geophys. Res. Lett.*, **31**, 2004GL020876, https://doi.org/10.1029/2004GL020876.

Gneiting, T., and A. E. Raftery, 2005: Weather Forecasting with Ensemble Methods. *Science*, **310**, 248–249, https://doi.org/10.1126/science.1115255.

Gneiting, T., and A. E. Raftery, 2007: Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Gralewicz, N. J., T. A. Nelson, and M. A. Wulder, 2012: Factors influencing national scale wildfire susceptibility in Canada. *For. Ecol. Manag.*, **265**, 20–29, https://doi.org/10.1016/j.foreco.2011.10.031.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

Hanes, C., X. Wang, P. Jain, M.-A. Parisien, J. M. Little, and M. D. Flannigan, 2019: Fire-regime changes in Canada over the last half century. *Can. J. For. Res.*, **49**, 256–269, https://doi.org/10.1139/cjfr-2018-0293.

Hanley, J. A., and B. J. McNeil, 1982: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36, https://doi.org/10.1148/radiology.143.1.7063747.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather Forecast.*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Holden, Z. A., and Coauthors, 2018: Decreasing fire season precipitation increased recent western US forest wildfire activity. *Proc. Natl. Acad. Sci.*, **115**, https://doi.org/10.1073/pnas.1802316115.

Holsinger, L. M., S. A. Parks, L. B. Saperstein, R. A. Loehman, E. Whitman, J. Barnes, and M. Parisien, 2022: Improved fire severity mapping in the North American boreal forest using a hybrid composite method. *Remote Sens. Ecol. Conserv.*, **8**, 222–235, https://doi.org/10.1002/rse2.238.

Hoyer, S., and J. Hamman, 2017: xarray: N-D labeled Arrays and Datasets in Python. *J. Open Res. Softw.*, **5**, 10, https://doi.org/10.5334/jors.148.

Jain, P., and Coauthors, 2024: Canada Under Fire – Drivers and Impacts of the Record-Breaking 2023 Wildfire Season, https://doi.org/10.22541/essoar.170914412.27504349/v1.

Kirchmeier-Young, M. C., and Coauthors, 2024: Human driven climate change increased the likelihood of the 2023 record area burned in Canada. *Npj Clim. Atmospheric Sci.*, **7**, 316, https://doi.org/10.1038/s41612-024-00841-9.

Lawson, B. D., and O. B. Armitage, 2008: *Weather guide for the Canadian Forest Fire Danger Rating System*. Northern Forestry Centre.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-Based Monthly Prediction at the Canadian Meteorological Centre. *Mon. Weather Rev.*, **144**, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

Mai, J., and Coauthors, 2020: The Canadian Surface Prediction Archive (CaSPAr): A Platform to Enhance Environmental Modeling in Canada and Globally. *Bull. Am. Meteorol. Soc.*, **101**, E341–E356, https://doi.org/10.1175/BAMS-D-19-0143.1.

Marlon, J. R., and Coauthors, 2009: Wildfire responses to abrupt climate change in North America. *Proc. Natl. Acad. Sci.*, **106**, 2519–2524, https://doi.org/10.1073/pnas.0808212106.

McElhinny, M., J. F. Beckers, C. Hanes, M. Flannigan, and P. Jain, 2020: A high-resolution reanalysis of global fire weather from 1979 to 2018 – overwintering the Drought Code. *Earth Syst. Sci. Data*, **12**, 1823–1833, https://doi.org/10.5194/essd-12-1823-2020.

McKenzie, D., Z. Gedalof, D. L. Peterson, and P. Mote, 2004: Climatic Change, Wildfire, and Conservation. *Conserv. Biol.*, **18**, 890–902, https://doi.org/10.1111/j.1523-1739.2004.00492.x.

Mermoz, M., T. Kitzberger, and T. T. Veblen, 2005: LANDSCAPE INFLUENCES ON OCCURRENCE AND SPREAD OF WILDFIRES IN PATAGONIAN FORESTS AND SHRUBLANDS. *Ecology*, **86**, 2705–2715, https://doi.org/10.1890/04-1850.

Mölders, N., 2010: Comparison of Canadian Forest Fire Danger Rating System and National Fire Danger Rating System fire indices derived from Weather Research and Forecasting (WRF) model data for the June 2005 Interior Alaska wildfires. *Atmospheric Res.*, **95**, 290–306, https://doi.org/10.1016/j.atmosres.2009.03.010.

MSC, 2024: Canadian Global Ensemble Prediction System (GEPS). Accessed 10 February 2025, https://eccc-msc.github.io/open-data/msc-data/nwp_geps/readme_geps_en/.

Natural Earth, 2024: Timezones - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales. Accessed 24 February 2025, https://www.naturalearthdata.com/downloads/10m-cultural-vectors/timezones/.

NCEP, 2024: Global Ensemble Forecast System (GEFS). Accessed 10 February 2025, https://www.nco.ncep.noaa.gov/pmb/products/gens/.

NOAA, 2024: NOAA Global Ensemble Forecast System (GEFS) - Registry of Open Data on AWS. Accessed 19 February 2025, https://registry.opendata.aws/noaa-gefs/.

Parker, W. S., 2010: Predicting weather and climate: Uncertainty, ensembles and probability. *Stud. Hist. Philos. Sci. Part B Stud. Hist. Philos. Mod. Phys.*, **41**, 263–272, https://doi.org/10.1016/j.shpsb.2010.07.006.

Pausas, J. G., and J. E. Keeley, 2021: Wildfires and global change. *Front. Ecol. Environ.*, **19**, 387–395, https://doi.org/10.1002/fee.2359.

Podur, J., and B. M. Wotton, 2011: Defining fire spread event days for fire-growth modelling. *Int. J. Wildland Fire*, **20**, 497, https://doi.org/10.1071/WF09001.

Richardson, D. S., H. L. Cloke, and F. Pappenberger, 2020: Evaluation of the Consistency of ECMWF Ensemble Forecasts. *Geophys. Res. Lett.*, **47**, e2020GL087934, https://doi.org/10.1029/2020GL087934.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What Does a Convection-Allowing Ensemble of Opportunity Buy Us in Forecasting

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.

Thunderstorms? *Weather Forecast.*, **35**, 2293–2316, https://doi.org/10.1175/waf-d-20-0069.1.

Rocklin, M., 2015: Dask: Parallel computation with blocked algorithms and task scheduling. SciPy, 126–132.

Saito, T., and M. Rehmsmeier, 2015: The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, **10**, e0118432, https://doi.org/10.1371/journal.pone.0118432.

Schoennagel, T., and Coauthors, 2017: Adapt to more wildfire in western North American forests as climate changes. *Proc. Natl. Acad. Sci.*, **114**, 4582–4590, https://doi.org/10.1073/pnas.1617464114.

Stocks, B. J., T. J. Lynham, B. D. Lawson, M. E. Alexander, C. E. Van Wagner, R. S. McAlpine, and D. E. Dubé, 1989: The Canadian Forest Fire Danger Rating System: An Overview. *For. Chron.*, **65**, 450–457, https://doi.org/10.5558/tfc65450-6.

Thompson, J. R., and T. A. Spies, 2009: Vegetation and weather explain variation in crown damage within a large mixed-severity wildfire. *For. Ecol. Manag.*, **258**, 1684–1694, https://doi.org/10.1016/j.foreco.2009.07.031.

Van Wagner, C. E., 1987: *Development and structure of the Canadian forest fire weather index system.* Canadian Forestry Service, Headquarters, Ottawa, https://ostrnrcan-dostrncan.canada.ca/handle/1845/228434.

Wang, W., X. Wang, M. D. Flannigan, L. Guindon, T. Swystun, D. Castellanos-Acuna, W. Wu, and G. Wang, 2025: Canadian forests are more conducive to high-severity fires in recent decades. *Science*, **387**, 91–97, https://doi.org/10.1126/science.ado1006.

Wotton, B. M., 2009: Interpreting and using outputs from the Canadian Forest Fire Danger Rating System in research applications. *Environ. Ecol. Stat.*, **16**, 107–131, https://doi.org/10.1007/s10651-007-0084-2.

Wotton, B. M., and M. D. Flannigan, 1993: Length of the fire season in a changing climate. *For. Chron.*, **69**, 187–192, https://doi.org/10.5558/tfc69187-2.

xarray-contrib/xskillscore, 2024. https://xskillscore.readthedocs.io

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the New NCEP Global Ensemble Forecast System in a Parallel Experiment. *Weather Forecast.*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhou, X., and Coauthors, 2022: The Development of the NCEP Global Ensemble Forecast System Version 12. *Weather Forecast.*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhuang, J., and Coauthors, 2022: pangeo-data/xESMF: v0.7.0, https://doi.org/10.5281/ZENODO.7447707.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-25-0069.1.