

RESEARCH ARTICLE

Subseasonal predictability of cold spells in the central United States

Oliver T. Millin  | Jason C. Furtado

School of Meteorology, University of
Oklahoma, Norman, Oklahoma

Correspondence

Oliver T. Millin, School of Meteorology,
University of Oklahoma, Norman,
Oklahoma, 73072, USA.
Email: omillin@ou.edu

Funding information

NSF Oklahoma EPSCoR, Grant/Award
Number: OIA-1946093

Abstract

Wintertime cold spells in the central United States pose significant socioeconomic and environmental risks, yet actionable subseasonal forecasts beyond 2 weeks remain a challenge. Here, we evaluate central United States cold spell predictability at weeks 3–4 using reanalysis and a suite of subseasonal model hindcasts, focusing on the physical mechanisms underlying forecast success and failure. We find that prediction skill declines rapidly beyond week 2, but forecasts retain sufficient resolution and sharpness to identify precursor signals linked to performance. Skillful and confident week 3 forecasts are preceded by a weakening stratospheric polar vortex (SPV), persistent high-latitude blocking, and robust stratosphere-troposphere coupling. In contrast, confident false alarms are often associated with unrealistic North American winter weather regime transitions and weak or absent stratospheric signals, particularly under strong SPV conditions. By stratifying confident forecasts by SPV state at initialization, we demonstrate that weak SPV conditions enhance forecast reliability, whereas strong SPV states can degrade it. These findings reveal that the vertical structure and timing of SPV variability play a critical role in modulating extended-range predictability of central United States cold spells, and capturing both SPV variability and weather regime persistence is essential for improving weeks 3–4 cold spell forecasts. Overall, the study offers a mechanistic standpoint and guidance for interpreting and developing more reliable extended-range predictions of extreme winter cold in the central United States.

KEYWORDS

cold spells, stratospheric polar vortex, subseasonal prediction, weather regimes

1 | INTRODUCTION

Persistent anomalous wintertime cold across the central United States (US) can have substantial socioeconomic, environmental, and infrastructural impacts (e.g., Cellitti *et al.*, 2006; Smith & Sheridan, 2018, 2020; Wheeler *et al.*, 2011). These impacts include stress on the electrical grid due to excessive demand, loss of life, and billions of dollars in damages (Doss-Gollin *et al.*, 2021;

Millin *et al.*, 2024; NCEI, 2025). In particular, spatially and temporally extended extreme cold events, termed cold air outbreaks (CAOs), have some of the largest impacts. One recent example is the February 2021 CAO, which was the second most severe CAO in the Great Plains between 1950 and 2021—based on a severity measure incorporating both magnitude and longevity in Millin *et al.* (2022)—and caused a wide range of impacts, including numerous power outages, traffic incidents, widespread

winter storm warnings, \$27.5bn in CPI-adjusted cost, and, unfortunately, various cold-related fatalities (Doss-Gollin *et al.*, 2021; Millin & Furtado, 2022; NCEI, 2025). As such, it is important to obtain better, longer-lead forecasts of anomalous cold, such as on the subseasonal-to-seasonal (S2S) timescale (i.e., 2 weeks to 2 months lead; Vitart *et al.*, 2017).

The demand for better S2S forecasts of such extreme weather events is growing, as decision-makers, end users, and stakeholders seek enhanced prediction skill for planning purposes and mitigation efforts. Indeed, the S2S timescale is notoriously difficult to forecast on as it lies between the shorter memory of atmospheric features on weather timescales and the longer memory that is conserved in boundary conditions of the ocean and land surface for seasonal forecasts (Millin *et al.*, 2024; White *et al.*, 2017). Enhancing prediction skill of such cold events can be through making use of windows of forecast opportunity, which are periods where the skill of an S2S forecast is temporarily higher than normal due to favorable conditions within the climate system (Mariotti *et al.*, 2020).

One such mode of variability that can provide enhanced prediction skill of surface cold extremes is the stratospheric polar vortex (SPV). Variability in the SPV can have substantial impacts on surface weather, particularly in regard to extreme midlatitude cold following major sudden stratospheric warmings (SSWs; e.g., Baldwin & Dunkerton, 2001; Butler *et al.*, 2017; Lee *et al.*, 2019; Afarghan-Gerstman & Domeisen, 2020; Millin *et al.*, 2022). In these events, the climatological westerly winds in the polar stratosphere reverse to easterly due to strong upward and poleward wave flux from the troposphere to stratosphere, ultimately depositing heat and momentum, thus weakening the vortex (Baldwin & Dunkerton, 2001; Charlton & Polvani, 2007; Nakagawa & Yamazaki, 2006). Subsequently, the anomalously weak winds propagate downward into the troposphere, enabling the negative phase of the Northern Annular Mode, anomalous ridging across northern Canada/Greenland, and an equatorward shift of the jet stream, which facilitates midlatitude CAOs (e.g., Baldwin & Dunkerton, 2001; Hitchcock & Simpson, 2014; Millin *et al.*, 2022; Thompson & Wallace, 2000; White *et al.*, 2019).

Prior work has shown that this process is especially robust in developing anomalous cold across Eurasia on subseasonal timescales (e.g., Butler *et al.*, 2017; Huang *et al.*, 2022; Kautz *et al.*, 2019), with the probability of cold spells across northern Eurasia being higher when there is a weak SPV at initialization (Erner *et al.*, 2022). Alternatively, a major SSW is not always required for this SPV–cold connection, with minor warmings also having an impact (Liu *et al.*, 2022; Wang & Chen, 2010). Interestingly, strong SPV conditions, developing from a lack of

upward and poleward wave activity, have typically been associated with anomalous warmth across the midlatitudes in Eurasia (Tripathi *et al.*, 2015). However, more recent analysis of CAOs across the US Great Plains shows that a strong mid-stratosphere circulation may actually be conducive for cold development through stratospheric wave reflection, a process whereby planetary-scale wave activity originating in Siberia propagates vertically into the stratosphere and subsequently is reflected back down into the troposphere over northern Canada. The result of this stratospheric wave reflection is changes in the polar jet stream, often with anomalous ridging over/near Alaska (Messori *et al.*, 2022; Millin *et al.*, 2022; Schutte *et al.*, 2025).

Indeed, North American winter weather regimes—that is, large-scale, quasi-persistent, recurrent circulation patterns that can last for days or weeks (Lee *et al.*, 2019; Millin *et al.*, 2022)—have been linked with potential long-lead prediction of extreme cold in the central US and have unique ties to SPV variability (Lee *et al.*, 2019; Millin *et al.*, 2022). The Arctic High (ArH) regime, characterized by anomalous ridging over Greenland/northern Canada, commonly occurs post-weak-SPV event and is important for the development of ArH-type CAOs in the Great Plains (Lee *et al.*, 2019; Millin *et al.*, 2022). Additionally, the Alaskan Ridge (ArR) regime has been connected to central US CAOs and a strong mid-stratospheric polar vortex (Messori *et al.*, 2022; Millin *et al.*, 2022). Furthermore, Lee *et al.* (2022) has demonstrated that specific weather regime transitions may be more or less likely during different vortex strengths. For example, under weak SPV conditions, the dominant transition is nearly always toward the ArH regime. However, under strong SPV conditions, the regime transition is dominantly toward the Arctic Low (ArL) regime, characterized by a pan-Arctic trough, or the Pacific Trough (PT) regime, characterized by anomalous troughing in the North Pacific. Thus, regimes and the state of the SPV both have individual and collective potential impacts on subseasonal predictability of central US cold.

In this study, we aim to quantify the subseasonal predictability of cold spells in the central US using various skill metrics in S2S models and assess the tropospheric circulation, regime, and stratospheric characteristics associated with confident hit and false alarm forecasts. We predominantly follow the methods of Erner and Karpechko (2024), who showed that confident cold spell predictions for northern Eurasia have cold already in the initial conditions in conjunction with a negative North Atlantic Oscillation (NAO) or Scandinavian block, as well as finding that S2S models tend to overpredict cold spells, with false alarms occurring more often during periods of a weak SPV. We will replicate some of the same analyses but for central US cold spells to explore the SPV–regime

relationships in S2S forecasts and quantify the prediction skill of the subseasonal forecast systems.

2 | DATA AND METHODS

2.1 | Verification data and identifying cold spells

This work uses daily-averaged October through April (ONDJFMA) atmospheric fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) 5th Reanalysis (ERA5) over the period October 1, 1993, to April 30, 2021, for our verification dataset (Hersbach *et al.*, 2020). This reanalysis product is available on a 0.5° horizontal resolution with 37 vertical levels from 1000 hPa to 1 hPa. We obtain both 2-m temperature (T2M) and geopotential height (GPH) for our analyses. Anomalies are calculated with respect to the 1993–2020 climatology, and the seasonal cycle is removed from each day. This climatological window is chosen because it matches the period over which model reforecasts are obtained.

To define “cold spells”, we follow the method of Erner and Karpechko (2024) but for the central US, a region that experiences some of the most extreme cold waves in North America (Millin *et al.*, 2022). For this study, the central US is defined by the region bounded by 91.5–103.5°W and 30–48°N. We choose this domain to ensure continuity with prior work linking extreme central US cold with key modes of variability on S2S timescales (Millin *et al.*, 2022; Millin & Furtado, 2022). We begin with area-averaging the November through April ERA5 T2M anomalies across this domain and compute weekly averages by employing a forward-looking overlapping rolling mean starting on November 1 in each season. For example, the weekly average for week commencing November 1, 1993, will be the average of November 1–7, 1993, and the weekly average

for week commencing November 2, 1993, will be the average of November 2–8, 1993, and so forth. Cold spells are defined when the weekly-averaged T2M anomalies across the central US fall below the lower tercile (33rd percentile). Finally, for verification scoring, we assign binary values to the weekly-averaged T2M anomalies (1 for a cold spell; 0 otherwise).

2.2 | S2S prediction systems and model cold weeks

In this study, we use reforecasts from six S2S prediction systems in the S2S Project Database (Table 1; Vitart *et al.*, 2017). We have chosen these specific models for several reasons:

1. They provide a minimum of 19 seasons of reforecasts to allow for a robust statistical analysis.
2. They have seven or more ensemble members, which is important for probabilistic forecasts of the cold spells.
3. Each model is initialized on the same calendar day each year, which allows for a robust climatology calculation for each model.

The models chosen for this study are listed in Table 1. We select the reforecasts that are initialized in November through March only between 1993 and 2020. Each model's reforecast span may differ within this time range (Table 1). Collectively, these models yield a total of 3,454 reforecasts for analysis. From these S2S models, we retrieve T2M and GPH fields. Note that, in the S2S models, the T2M data are daily-averaged and the GPH data are instantaneous (taken at 0000 UTC each day). Furthermore, the S2S model data are retrieved at a 1.5° horizontal resolution with 10 vertical levels from 10 hPa to 1000 hPa. Since the Bureau of Meteorology (BOM) model output is provided in its native grid, we bilinearly interpolate the

TABLE 1 The S2S Project Database models used in the study.

Model	Model version	Time range (days)	Time period used	Reforecast size	Runs per NDJFM	Total runs used
ECMWF	CY47R3	0–46	2001–2021	11	43	880
UKMO	GloSea6	0–60	1993–2016	7	20	460
KMA	GloSea6–GC3.2	0–60	1993–2016	7	20	460
BOM	POAMA P24	0–62	1993–2013	33	30	600
CNR-ISAC	GLOBO	0–35	2001–2020	8	30	570
HMCR	RUMS	0–46	1993–2015	11	22	484

Abbreviations: NDJFM: November–March; ECMWF: European Centre for Medium-Range Weather Forecasts; UKMO: United Kingdom Meteorological Office; KMA: Korean Meteorological Agency; BOM: Bureau of Meteorology; CNR-ISAC: Italian National Research Council–Institute of Atmospheric Sciences and Climate; HMCR: Hydro-Meteorological Centre of Russia.

data to the same 1.5° horizontal resolution as the other models. Anomalies are calculated with respect to a lead- and initialization-dependent model climatology with a leave-one-out method. That is, the climatology for each reforecast initialization date is the mean of all reforecasts across year and ensemble member space in each model's reforecast period that is initialized on the same calendar day except for the reforecast on that particular date.

To calculate model forecast cold spells, we first calculate the area-averaged T2M anomalies over the central US before weekly averaging the forecast data. Forecast anomalies are averaged between leads 0–6 days (week 1), 7–13 days (week 2), 14–20 days (week 3), and 21–27 days (week 4). For each initialization date and week, we compute the lower tercile across year and ensemble member space in each model's reforecast period on the same calendar day, except for the reforecast on that particular date. Each week is then classified as a non-cold (0) or cold (1) week. Forecast probabilities of a cold week are then computed by taking the mean of the binary cold week forecast across ensemble members and comparing that with ERA5. Though forecast skill is assessed for weeks 1–4 initially, we focus most of the analysis on weeks 3 and 4 as we want to gauge the subseasonal prediction capability and drivers in these models.

2.3 | Skill assessment and metrics

To assess the skill of the subseasonal prediction systems in forecasting cold weeks, we use several skill scores (CAWCR, 2015), computed using Python's *climpred* package (Brady & Spring, 2021).

- The “Heidke Skill Score” (HSS), a generalized skill score that measures the proportion of correct forecasts after removing those that would be correct due to random chance and is given by the following equation:

$$\text{HSS} = \frac{(a + d) - (n_{\text{exp}})_{\text{random}}}{N - (n_{\text{exp}})_{\text{random}}}, \quad (1)$$

where the number of expected correct forecasts by random chance $(n_{\text{exp}})_{\text{random}}$ is

$$(n_{\text{exp}})_{\text{random}} = \frac{(a + c)(a + b) + (d + c)(d + b)}{N}, \quad (2)$$

where a , b , c , and d represent the forecast hits, false alarms, misses, and correct rejections (i.e., contingency table components), respectively, and N is the total number of forecasts.

- The “Probability of Detection” (POD), which describes the proportion of correct forecasts, and the Miss Rate

(MR), which describes the proportion of events not detected by the models:

$$\text{POD} = \frac{a}{a + c}, \quad (3)$$

$$\text{MR} = 1 - \text{POD}. \quad (4)$$

- The “False Alarm Ratio” (FAR), which describes the fraction of predicted events that did not occur:

$$\text{FAR} = \frac{b}{a + b}. \quad (5)$$

Table 2 shows the range of possible scores and the scores equating to the perfect and “no skill” outcomes. In the case of HSS, the lower bound of the skill range is -1 , which occurs in the extreme case where a forecast model consistently predicts an event when it does not occur and consistently fails to predict an event when it does occur.

In addition to skill scores, we also examine reliability diagrams for weeks 3 and 4 cold week forecasts to assess how model forecast probabilities compare with the observed frequencies (Met Office, 2025). To calculate the reliability diagrams, we use the *xskillscore* Python package, which bins the forecast probabilities and then compares them with the relative frequencies in ERA5. Several lines comprise a reliability diagram: (a) the diagonal “perfect reliability” 1:1 line representing a perfect forecast, where forecast probabilities are equal to observed relative frequencies; (b) the climatological event frequency lines, which are vertical and horizontal lines placed at 0.33 (i.e., 33rd percentile) in our case; and (c) the “no skill” line, which lies halfway between the “perfect reliability” line and the climatological event frequency. If the reliability curve has a slope that is smaller than the “no skill” line, then the prediction system is said to be overconfident (i.e., cold weeks are overpredicted), whereas above this line and the forecast is overall adding value.

We complement the reliability diagrams with sharpness histograms, which show the relative proportion of forecasts falling into each forecast probability bin (Met Office, 2025). As noted in Erner and Karpechko (2024), the expectation in this analysis is that the frequency

TABLE 2 The range of possible scores, the value of a perfect score, and the value of a “no skill” score for the Heidke skill score, probability of detection, miss rate, and false alarm ratio.

Skill score	Range	Perfect score	No skill score
Heidke skill score	−1 to 1	1	0
Probability of detection	0 to 1	1	0
Miss rate	0 to 1	0	1
False alarm ratio	0 to 1	0	1

peak occurs at around 0.33 forecast probability, since we used the lower tercile threshold in weekly averaged T2M anomaly to define cold weeks. Finally, we use probability bins at 0.1 spacing between 0 and 1 for models that have 10 or more ensemble members (i.e., ECMWF, BOM, and Hydro-Meteorological Centre of Russia (HMCR)), but the reliability is calculated using bins with $1/M$ spacing for those with less than 10 ensemble members (i.e., United Kingdom Meteorological Office (UKMO), Korean Meteorological Agency (KMA), and Italian National Research Council–Institute of Atmospheric Sciences and Climate (CNR-ISAC)) for appearance purposes, where M is the number of ensemble members. Thus, sharpness histograms are occasionally missing counts for UKMO, KMA, and CNR-ISAC as they do not contain forecast probabilities that fall into every bin spaced at even 0.1 intervals.

2.4 | Windows of forecast opportunity

Whereas skill scores and reliability metrics provide a broad assessment of model performance, more targeted insights are gained by focusing on forecasts with higher ensemble agreement (i.e., windows of forecast opportunity). Bloomfield *et al.* (2021) found that there is an increase in extended-range forecast skill over Europe when at least 50% of ensemble members agree on the flow pattern, and Erner and Karpechko (2024) successfully used this definition in their study. As such, we also focus on central US cold spells that are predicted by at least 50% of ensemble members, herein referred to as “confident forecasts”, to assess the key signatures of enhanced predictability for central US cold spells on subseasonal timescales. Specifically, we define a “hit confident forecast” as a forecast that correctly predicts a cold spell with a probability of at least 50%. Subsequently, “a false alarm confident forecast” is defined as a forecast that incorrectly forecasts a cold spell with a probability of at least 50%. Finally, a “miss forecast” is when a forecast does not predict an observed week 3 cold spell with at least 50% probability. Most importantly, tracking the hit and false alarm confident forecasts allows us to conclude what conditions may yield enhanced predictability subseasonally versus overconfidence. For each category of forecast, we pool the initializations in that group together to represent the multimodel mean (MMM). For model field composites and other statistics, all initializations are kept, including initialization dates that match between models, as each forecast is considered an independent sample. However, for composites of the ERA5 fields for these different forecast categories, we remove repeating initialization dates to ensure we are not averaging the same map from the verification data each time. Statistical significance on all analyses is determined via a

bootstrapping approach with 5,000 iterations with replacement for significance testing on regime frequencies and model composites, but without replacement for ERA5 composites.

2.5 | Weather regimes

North American winter weather regimes have a substantial influence on the development of central US cold spells due to their quasi-persistent nature (Lee *et al.*, 2019; Millin *et al.*, 2022). In this work, we define North American winter weather regimes using the methods of Millin *et al.* (2022), an adapted methodology from that of Lee *et al.* (2019). Instead of November–March (NDJFM), we use ONDJFMA linearly-detrended 500 hPa GPH anomalies between 1993 and 2020, with respect to the 1993–2020 climatology, in ERA5 to define the regimes across the North American region (180° – 330° E, 19.5° – 79.5° N), to account for end of March forecasts that go into April and also lagged composites that extend into October. Empirical orthogonal function analysis is performed on the 500 hPa GPH anomalies, where the data are weighted by the square root of the cosine of latitude prior to eigenanalysis. We obtain the principal component time series of the 12 leading modes of variability, which represent $\sim 80\%$ of the variance explained, before employing the k -means clustering algorithm from Python’s sci-kit learn package (Pedregosa & Coauthors, 2011). Based on Millin *et al.* (2022), $k = 5$ is chosen for the number of clusters for this analysis. Each day in the record is then assigned a regime through calculating the minimum Euclidean distance between a day’s 500 hPa GPH anomaly field and each cluster centroid.

Figure 1 shows the composite maps of the 500 hPa GPH anomalies (shading) for each regime in ERA5 between 1993 and 2020, along with their occupation frequency. The least frequent regime is the AkR regime, featuring anomalous ridging across Alaska and occupying 14.6% of days (Figure 1a). The second least frequent regime is the ArH regime, characterized by anomalous ridging across northern Canada/Greenland and occurring on 17.7% of days (Figure 1b). Next, occupying 20.7% of ONDJFMA days between 1993 and 2020, is the West Coast ridge (WCR) regime with anomalous ridging across the US West Coast as part of a Rossby wave train (Figure 1c). The PT regime is the fourth most frequent regime, occupying 23.0% of days and featuring an enhanced Aleutian low (Figure 1d). The regime with the highest frequency is the ArL regime, occupying 24.0% of days and defined by pan-Arctic troughing (Figure 1e). To investigate the evolution of North American winter weather regimes in confident forecasts of subseasonal cold spells, we also employ these ERA5 empirical

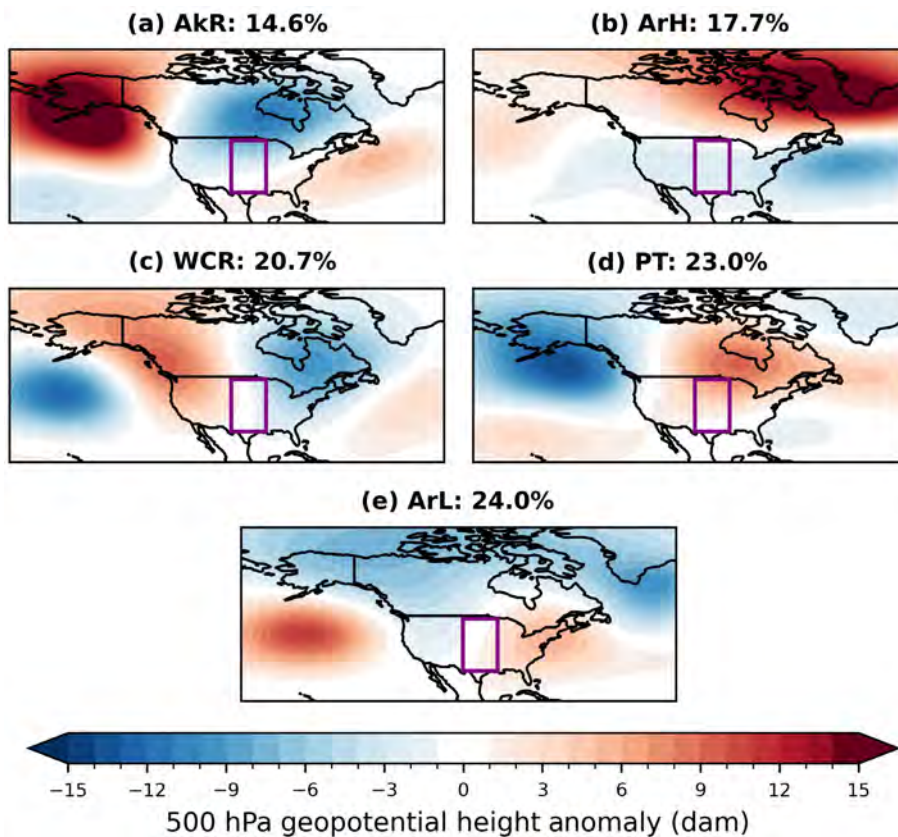


FIGURE 1 Composite mean maps of 500 hPa GPH anomalies (shading, dam) for the five North American winter weather regimes in ERA5: (a) the Alaskan Ridge (AkR) regime, (b) the Arctic High (ArH) regime, (c) the West Coast Ridge (WCR) regime, (d) the Pacific Trough (PT) regime, and (e) the Arctic Low (ArL) regime. Percentages indicate the occupation frequency of each regime during October to April over 1993–2020. Purple boxes indicate the central US region used to define the cold spells (30–48°N, 91.5–103.5°W). [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

orthogonal function and k -means objects to project onto the ensemble mean 500 hPa GPH anomalies in the S2S models to obtain the regime days in the models.

2.6 | SPV forcing

To assess the strength of the SPV and its potential relationship with the central US cold spells, we use the 50 hPa GPH anomalies from ERA5 at the time of initialization to determine the SPV strength, as in Erner and Karpechko (2024). Specifically, weak and strong SPV conditions are defined by the polar cap averaged (60–90°N) 50 hPa GPH anomalies being above or below the 80th and 20th percentiles of the NDJFM 1993–2020 period, respectively.

3 | RESULTS

3.1 | Forecast skill and reliability

Figure 2 shows the skill scores for all six S2S models in forecasting cold weeks, stratified by forecast week. In week 1, five models (KMA, UKMO, ECMWF, HMCR, and CNR-ISAC) exhibit strong performance, with HSS values between 0.7 and 0.9 (Figure 2a) and POD above

0.8 (Figure 2b). The FAR and MR are generally below 0.2 (Figure 2c,d). The BOM model lags behind, with an HSS near 0.5 and higher FAR and MR (both around 0.3). In week 2, skill declines across all metrics. ECMWF maintains the greatest skill (e.g., HSS around 0.4), followed by UKMO, KMA, CNR-ISAC, and HMCR. BOM again performs noticeably worse, primarily due to elevated FAR and MR (Figure 2c,d).

By week 3, model performance declines notably, with HSS values falling below 0.2 across all models and BOM becoming more comparable to the other prediction systems (Figure 2a). POD values hover around 0.4, whereas FAR and MR increase to approximately 0.6 (Figure 2b–d). At week 4, HSS drops further to below 0.1 for all models (Figure 2a), although POD and FAR remain similar to those in week 3 (Figure 2b,c). This suggests that rising MR and FAR are primarily responsible for the continued decline in HSS at longer lead times. Nevertheless, all six models maintain a positive HSS and a POD of around 0.4 at weeks 3 and 4, indicating that they retain meaningful skill in forecasting cold weeks at those lead times, albeit with some overconfidence. We also tested whether skill holds for predicting CAO days instead of cold weeks, using the NDJFM 20th percentile threshold over three or more consecutive days—as per methodology in Millin *et al.* (2022)—to ensure sample stability. The weeks 1 and 2 skill remains high and comparable to cold weeks for most

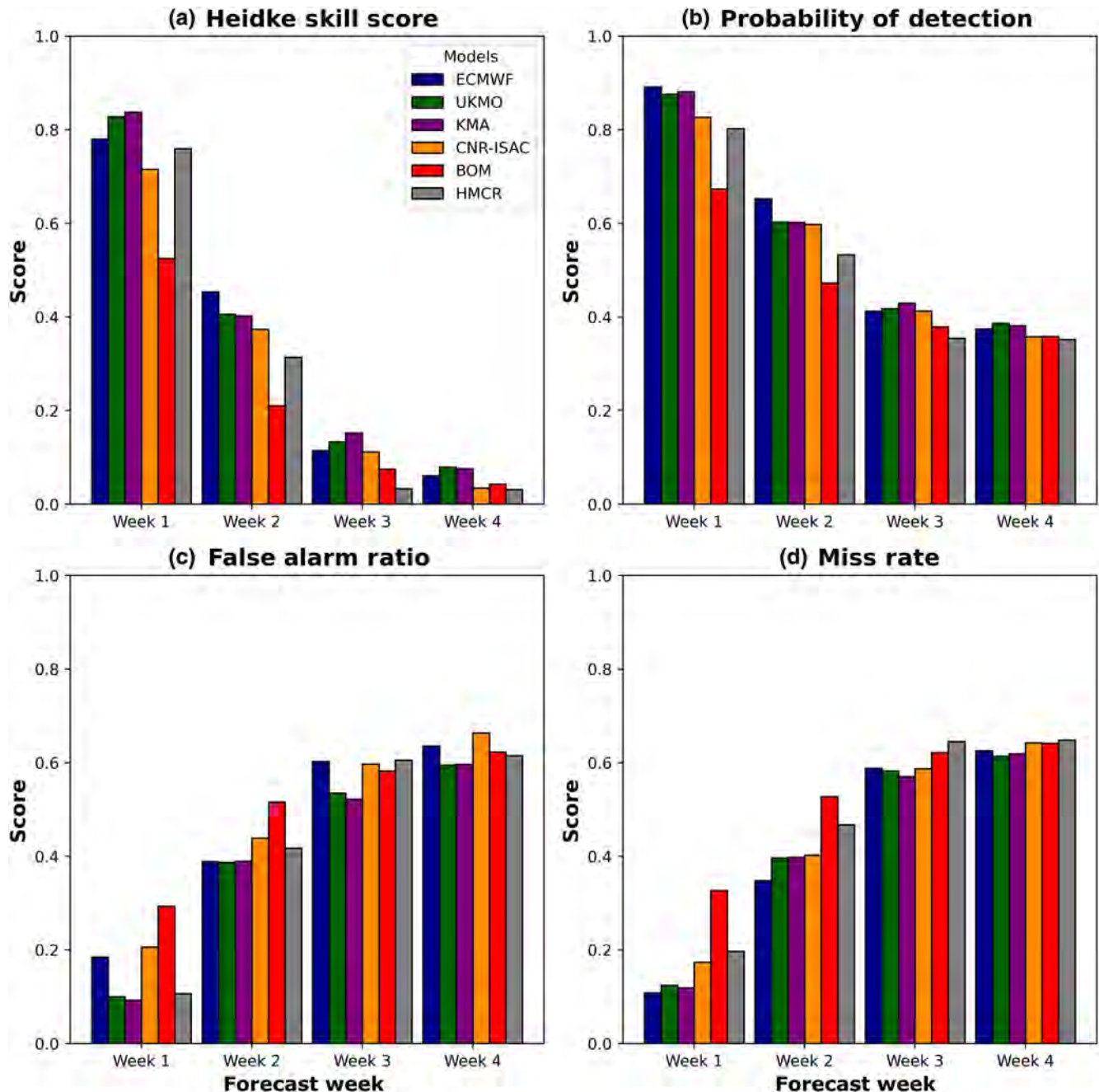


FIGURE 2 The skill of European Centre for Medium-Range Weather Forecasts (ECMWF; blue bars), United Kingdom Meteorological Office (UKMO; green bars), Korean Meteorological Agency (KMA; purple bars), Italian National Research Council–Institute of Atmospheric Sciences and Climate (CNR-ISAC; orange bars), Bureau of Meteorology (BOM; red bars), and Hydro-Meteorological Centre of Russia (HMCR; gray bars) in forecasting cold weeks as a function of forecast week in terms of (a) Heidke skill score, (b) Probability of detection, (c) False alarm ratio, and (d) Miss rate. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

models (except BOM), but drops sharply after week 2 (not shown).

We further evaluate the S2S models through a reliability assessment. Figure 3 presents reliability diagrams and sharpness histograms for week 3 and week 4 cold spells. For week 3, most models exhibit underconfidence at low forecast probabilities (0–0.3), with reliability curves lying

above the perfect reliability and no-skill lines, indicating too many false negatives (Figure 3a). As forecast probabilities increase, the curves typically fall below the no-skill line, reflecting overconfidence and an excess of false positives. However, ECMWF (blue line), KMA (purple line), and BOM (red line) show some limited skill at higher probabilities (> 0.4), with curves that lie between the no-skill

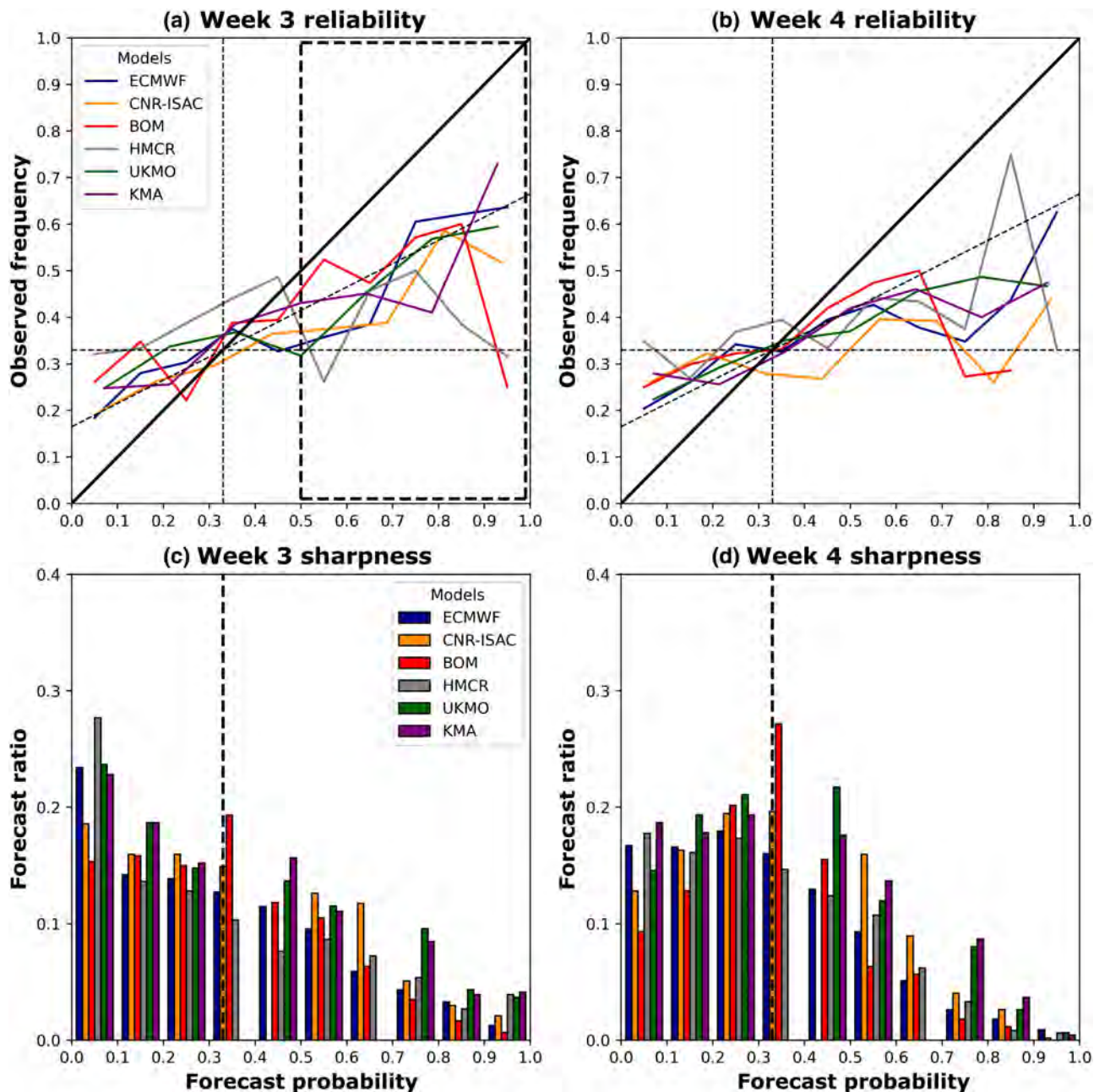


FIGURE 3 Reliability diagrams and sharpness histograms for the subseasonal-to-seasonal models in forecasting (a, c) week 3 and (b, d) week 4 cold spells. For the reliability diagrams, the forecast probability is shown on the x-axis and the observed relative frequency is shown on the y-axis, the horizontal and vertical dashed gray lines represent the climatological frequency, the diagonal black line represents the perfect reliability 1:1 line, the diagonal dashed gray line represents the “no-skill” line, and the dashed black box is the region of confident forecasts (i.e., $\geq 50\%$ forecast probability). The sharpness histograms have forecast probability on the x-axis and forecast ratio on the y-axis, where the vertical dashed black line represents the climatological frequency (i.e., the climatological expected rate of occurrence). BOM: Bureau of Meteorology; ECMWF: European Centre for Medium-Range Weather Forecasts; CNR-ISAC: Italian National Research Council–Institute of Atmospheric Sciences and Climate; HMCR: Hydro-Meteorological Centre of Russia; KMA: Korean Meteorological Agency; UKMO: United Kingdom Meteorological Office. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

and perfect reliability lines. In particular, ECMWF and KMA display higher observed frequencies at higher forecast probabilities. Despite biases at both low and high forecast probabilities, the overall reliability curves show a rising pattern from bottom left to top right, suggesting

that the models retain some resolution in week 3 forecasts. The accompanying sharpness histograms reveal that most models issue the majority of forecasts in the lower probability bins; the exception is the BOM model, which peaks in the 0.3–0.4 range, likely due to the scarcity of confident

forecasts at long leads (Figure 3c). Across models, the number of forecasts generally declines with increasing forecast probability, indicating that while the models can produce high probability forecasts for cold spells at week 3, such forecasts are relatively rare. Overall, the S2S models (except the BOM model) demonstrate sharpness by issuing forecasts with probabilities distinct from the climatological frequency (0.33).

For week 4 cold spells, the reliability curves indicate overconfident forecasts by the models, particularly at forecast probabilities above 0.4, (Figure 3b). Notable exceptions include BOM (red line) for forecast probabilities between 0.4 and 0.7 and HMCR (gray line) between 0.8 and 0.9, both of which show some skillful confident forecasts. The sharpness histogram further reveals that most forecasts cluster around the climatological frequency (dashed black line), especially in the 0.2–0.5 probability range (Figure 3d). This peak near the climatological value suggests that most forecasts offer limited deviation from the base rate and thus exhibit little sharpness at week 4. Given that week 3 forecasts demonstrate greater reliability, sharpness, and overall skill compared with week 4 (Figures 2 and 3), the remainder of our analysis focuses primarily on week 3.

3.2 | Tropospheric circulation and weather regimes in confident forecasts

To better understand the skill and limitations of week 3 cold spell forecasts, we next analyze confident forecasts across the six S2S models. Specifically, we examine composites of 500 hPa GPH anomalies and T2M anomalies for different outcomes (Figure 4). In ERA5, the initialization day composite for confident hits shows pronounced anomalous ridging (up to +3 dam) over Greenland and northeastern Canada, accompanied by troughing to the south over the North Atlantic, resembling the negative phase of the NAO and the ArH regime (Figure 4a). This supports the idea that confident hit forecasts may be linked to ArH regimes at initialization, consistent with previous studies and potential associations with weak SPV variability on subseasonal timescales (e.g., Lee *et al.*, 2019; Millin *et al.*, 2022, 2024). In contrast, confident false alarms show much weaker circulation anomalies at initialization, with only faint ridging over the North Atlantic and Arctic Ocean (Figure 4d), suggesting a lack of strong pre-existing signals. Interestingly, the miss forecasts exhibit a different signal entirely, with evidence of anomalous ridging over Siberia at initialization (Figure 4g), hinting at a circulation pattern not conducive to accurate cold spell prediction over the central US.

For week 3 confident hit forecasts in ERA5, the verifying week 3 composite shows strong anomalous ridging (up to +6 dam) over Alaska and the US West Coast, flanked by deep troughing over the North Pacific, the central/eastern US, and the North Atlantic. This pattern drives widespread negative T2M anomalies across the central and eastern US (Figure 4b). Additional ridging over Greenland and northern Canada further suggests that a persistent ArH regime may play a key role in enabling skillful subseasonal forecasts of central US cold spells. The overall circulation pattern exhibits features of AkR-, ArH-, and WCR-like regimes (Figure 1), including anomalous ridging over Alaska, the US West Coast, and northern Canada/Greenland, alongside anomalous troughing across the central/eastern US, all of which have been associated with CAOs in Millin *et al.* (2022). In contrast, ERA5 composites for confident false alarms at week 3 reveal a circulation pattern that is weaker and of opposite sign to that of the confident week 3 hit forecasts in Figure 4b. Specifically, there is a wave train pattern with anomalous troughing over Alaska and anomalous ridging over the central and northern US, leading to positive T2M anomalies in those regions (Figure 4e). These forecasts are not only confidently wrong, but they also misrepresent the sign of temperature anomalies, highlighting a fundamental circulation mismatch.

ERA5 patterns for the miss group closely resemble those of the confident hits, with anomalous ridging over Alaska and a downstream anomalous trough. However, they lack anomalous ridging over Greenland and northern Canada (Figure 4h). This absence suggests that the ArH regime may serve as an important distinguishing feature between successful and missed forecasts of central US cold, potentially in conjunction with AkR and WCR patterns. The MMM forecast composites support these interpretations. For confident hits, the MMM forecasts a circulation very similar to ERA5, with anomalous ridging over Alaska, the western US, and Greenland/northern Canada, along with downstream troughing (Figure 4c). In contrast, confident false alarm MMM forecasts feature anomalous ridging over Alaska but lack a Greenland high feature (Figure 4f). This may indicate that confident false alarms are more likely when the forecast features a pattern resembling AkR or WCR regimes with a lack of an anomalous Greenland block. The lack of ArH also hints at a potential connection to stratospheric preconditioning (e.g., SPV variability) for skillful week 3 cold forecasts. Finally, the MMM composites for week 3 misses show weak and incoherent circulation patterns, consistent with the poor performance, and accompanied by weak positive T2M anomalies across the US (Figure 4i).

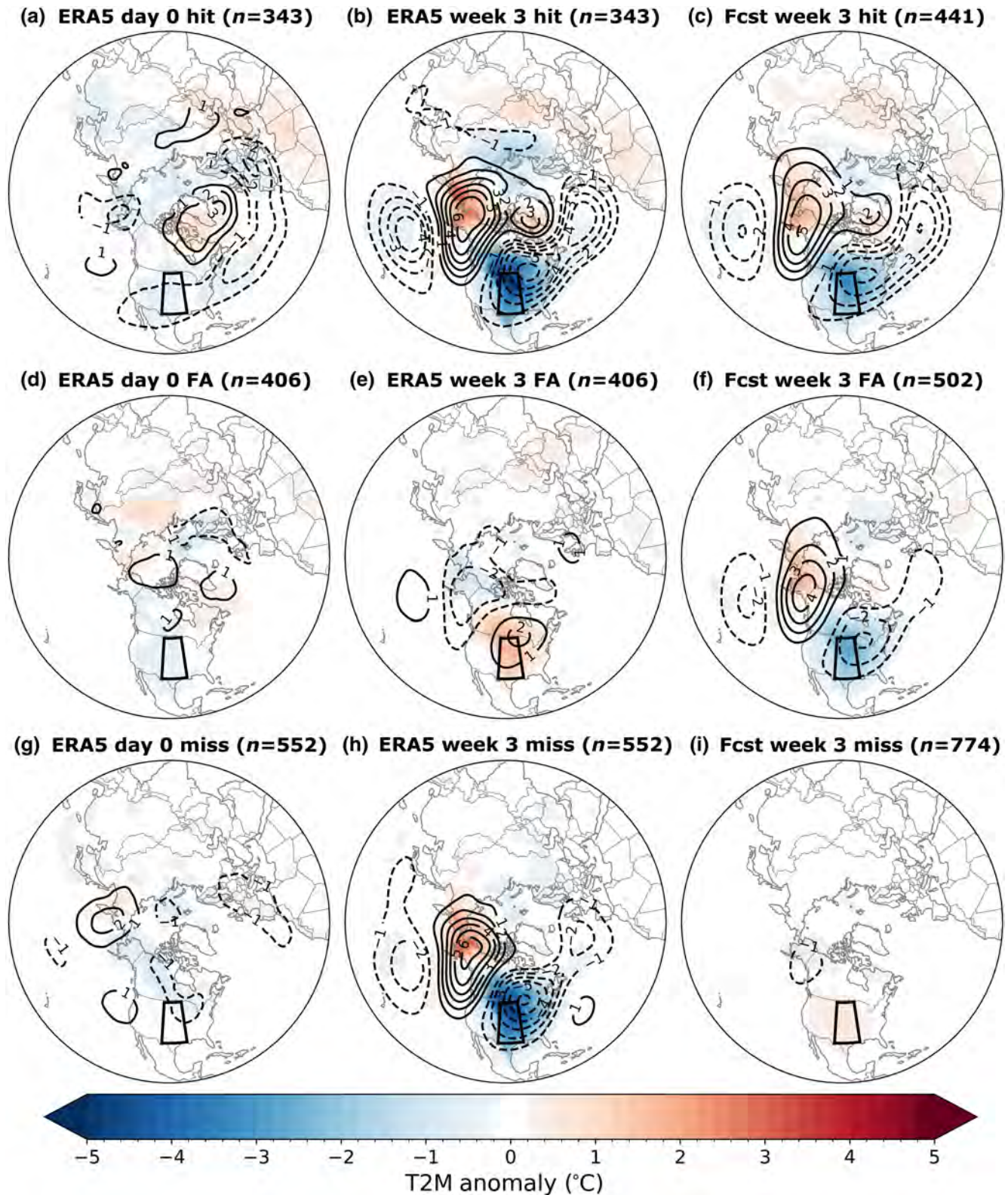


FIGURE 4 Composite maps of 2-m temperature (T2M) anomalies (shading, °C) and 500 hPa geopotential height (GPH) anomalies (contours, dam) for confident hit, false alarm (FA), and miss forecasts for (a, d, g) European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) initialization day, (b, e, h) ERA5 week 3, and (c, f, i) multimodel mean week 3. The black box shows the region used to define central US cold spells, whereas solid contours show positive GPH anomalies and dashed contours show negative GPH anomalies. The sample sizes n for each composite are given in the panel titles. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

Since the composite maps for confident hit and false alarm forecasts of week 3 cold spells reveal features that are also associated with North American winter weather regime patterns, we now examine the frequency evolution of the five regimes in Figure 5, relative to forecast initialization and lead time in ERA5 and the models. Frequencies are computed at each lag day with respect to day 0 (initialization), normalized by the number of samples in each forecast group. For confident hit forecasts, the PT regime (green line) dominates the 2–4 weeks preceding initialization, with the frequency reaching around 30% (Figure 5a). Two weeks before initialization, the ArH regime (red line) becomes dominant. This change in regime dominance aligns well with the observed anomalous ridging over the Arctic seen in Figure 4a, reinforcing the importance of the ArH as a precursor signal. The ArH regime maintains a dominant presence into week 3, reaching over 30% frequency in ERA5. As the event nears, however, the WCR (Figure 5a, blue line) regime emerges as the dominant regime during week 3, accompanied by a sharp uptick in AkR (Figure 5a, orange line). Simultaneously, the PT

and ArL (Figure 5a, purple line) regimes decline to less than 10% frequency, consistent with their unfavorable conditions for central US cold spells (Millin *et al.*, 2022, 2024). The strong WCR regime signal in week 3 is somewhat unexpected given the spatial prominence of the anomalous Alaskan ridging in Figure 4b. However, this likely reflects overlapping features in WCR-like events, including deep North Pacific troughing and possible strong AkR contributions to certain CAOs (compare Figures 1c and 4b). Importantly, the model forecast frequencies largely mirror those observed in ERA5 for all five regimes in confident hit forecasts (Figure 5b), highlighting the models' ability to reproduce the key regime evolutions that lead to successful cold spell forecasts. In particular, the significantly elevated ArH frequency prior to and at initialization suggests a critical preconditioning role for this regime in enabling skillful week 3 cold spell forecasts (Millin *et al.*, 2022).

The regime frequency evolution for the confident false alarm forecasts provides insight into how the models may become overconfident in predicting week 3 cold spells.

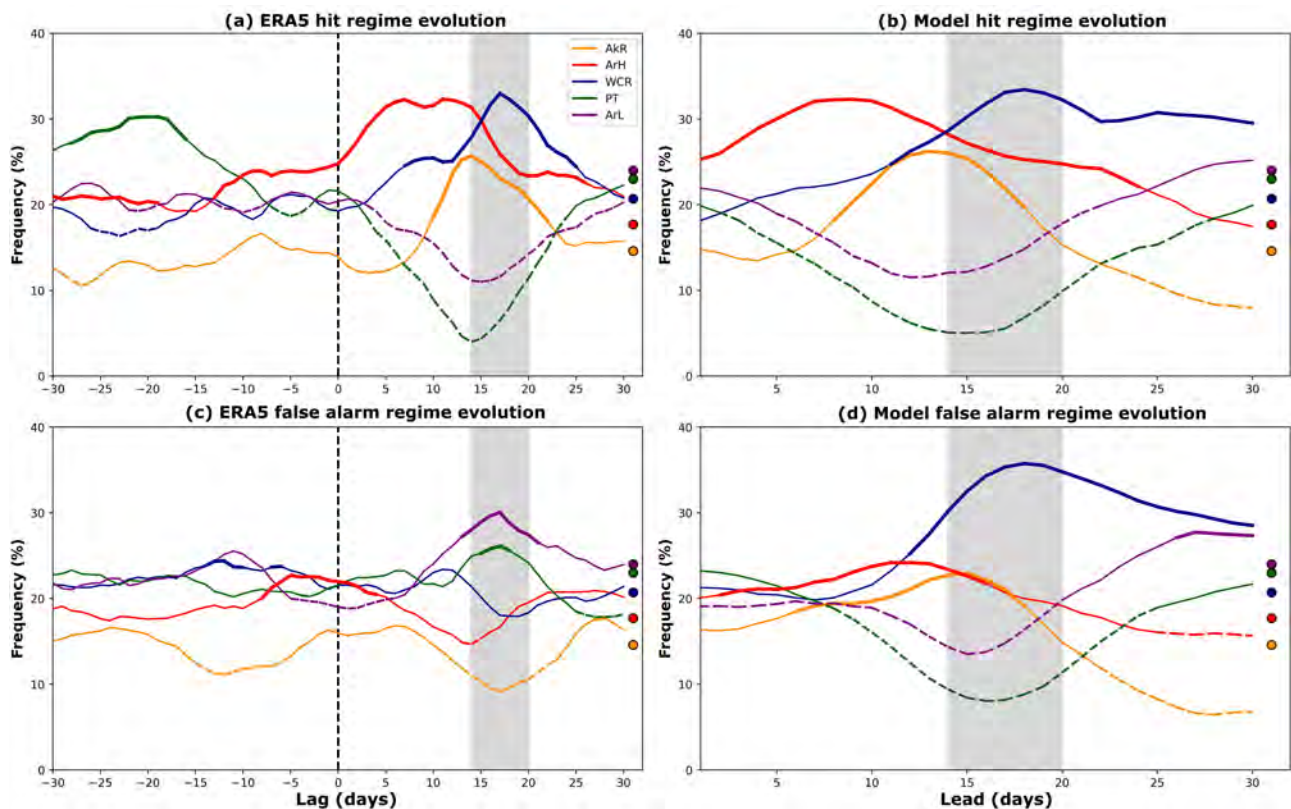


FIGURE 5 Lagged North American winter weather regime frequencies with respect to forecast initialization time (lag 0) for (a) the confident week 3 hit forecast initializations and (c) the confident week 3 false alarm forecast initializations in European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5). (b, d) As (a) and (c) but for the model forecasts as a function of lead time. Regime frequencies are calculated relative to the total number of samples in each group and are then smoothed with a 5-day centered filter. Bold and dashed lines respectively represent statistically significant enhanced or suppressed frequencies relative to climatology at the 95th percentile via 5,000-iteration bootstrapping from ERA5 data. Week 3 is shown in gray shading, and the colored dots indicate the climatological frequency of the five winter weather regimes. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.40081)]

In ERA5, no regime stands out 2–4 weeks before initialization; frequencies for all five regimes hover near 20% (Figure 5c). In the 2 weeks leading up to initialization, the WCR regime becomes significantly enhanced, followed by a brief increase in ArH frequency starting during the week before initialization. However, as week 3 approaches, the frequencies of the ArH and AkR regimes decline. Strikingly, during week 3, the PT and ArL regimes become significantly enhanced in ERA5, reaching frequencies of 25–30%. Since these regimes are generally associated with anomalous warmth over the central US, this suggests that false alarm forecasts are not only wrong about cold occurrence but also fail to systematically predict circulation patterns that drive warm anomalies instead (see Figure 4e).

Model forecasts for these false alarm events reflect some aspects of the observed evolution. They capture the initial rise in ArH frequency during the first lead week, peaking near 20% (Figure 5d). However, unlike reanalysis, the models continue to increase the ArH frequency toward week 3, maintaining a significantly enhanced signal. Similarly, the AkR regime is significantly overestimated throughout the forecast period, despite its decline in the verifying state. Perhaps most tellingly, the WCR regime dominates the false alarm model forecasts for week 3, with peak frequencies reaching around 35% during the target week (Figure 5d, blue line). This evolution suggests that false alarm forecasts are often linked to a more progressive, wave-train-like circulation across North America, consistent with Figure 4f. Prior work has shown that WCR-associated CAOs tend to be shorter and less intense, often lacking strong stratospheric precursors (Millin *et al.*, 2022). This may help explain the relatively muted ArH and AkR regime signals the model forecasts for false alarms, in contrast to the stronger SPV-linked signals seen in confident hit forecasts (Figure 5b,d). These distinctions highlight the potential role of stratospheric variability in differentiating between skillful and overconfident cold forecasts.

3.3 | Relating the state of the SPV to confident cold spell predictions

Figure 6 presents the composite standardized GPH anomalies averaged over the polar cap (60–90°N) for confident hit and false alarm forecasts of week 3 cold spells, along with their differences, in both ERA5 and the S2S models. For confident hit forecasts, ERA5 shows significantly positive GPH anomalies in the mid-to-upper stratosphere beginning around 4 weeks prior to initialization (Figure 6a), indicating a weaker-than-normal SPV. These positive anomalies gradually descend through

the stratosphere, reaching the lower levels near the initialization date and projecting onto the troposphere as anomalous ridging, likely linked to the ArH signal seen in ERA5 500 hPa GPH anomalies on day 0 (Figures 4a, 5a, and 6a). The weakened SPV signal persists in the lower stratosphere through the week 3 forecast period, with the most pronounced and statistically significant downward propagation into the troposphere coinciding with the cold spell. A recovery of the SPV, characterized by negative GPH anomalies in the upper stratosphere, occurs roughly 20–30 days after initialization, suggesting that weak vortex events, including major SSWs, play a role in the subseasonal predictability of central US cold spells (Figure 6a). Importantly, the S2S models show a comparable evolution. The forecast composite for confident hit cases also features significantly positive GPH anomalies in the mid-to-lower stratosphere, persisting from initialization through week 3 (Figure 6d). These anomalies exhibit a “dripping” structure indicative of downward propagation into the troposphere, aligning well with the observed timing of the cold spell. This agreement between the operational forecast model and reanalysis signals reinforces the notion that weak SPV conditions are one important preconditioning mechanism for skillful week 3 forecasts of central US cold spells.

Analyzing the stratospheric conditions associated with confident false alarm forecasts of week 3 cold spells provides further insight into the absence, or misrepresentation, of key stratospheric signals that may contribute to erroneous predictions. In ERA5, these false alarm cases exhibit little to no significant stratospheric anomalies prior to initialization (Figure 6b). Following initialization, the observed evolution in ERA5 shows a strengthening of the SPV, indicated by negative GPH anomalies in the upper stratosphere that descend toward the troposphere through week 3. This downward propagation of a stronger SPV signal is associated with the development of warmer winter weather regimes (Figure 5c), consistent with anomalous warmth observed during these falsely predicted cold spells. In contrast, the S2S model composite shows a divergent evolution from the reanalysis. Although the models simulate a stronger SPV aloft, with persistent negative GPH anomalies in the upper stratosphere, they simultaneously forecast a rapidly developing and significant tropospheric ridging signal, confined below 100 hPa. This low-level ridging emerges a few days before week 3 and persists through the forecast cold spell period (Figure 6e). This disconnect between the model’s stratospheric and tropospheric evolution may explain the overconfidence in predicting cold spells despite the lack of conducive stratospheric preconditioning, highlighting a potential source of false alarm error in subseasonal forecasts.

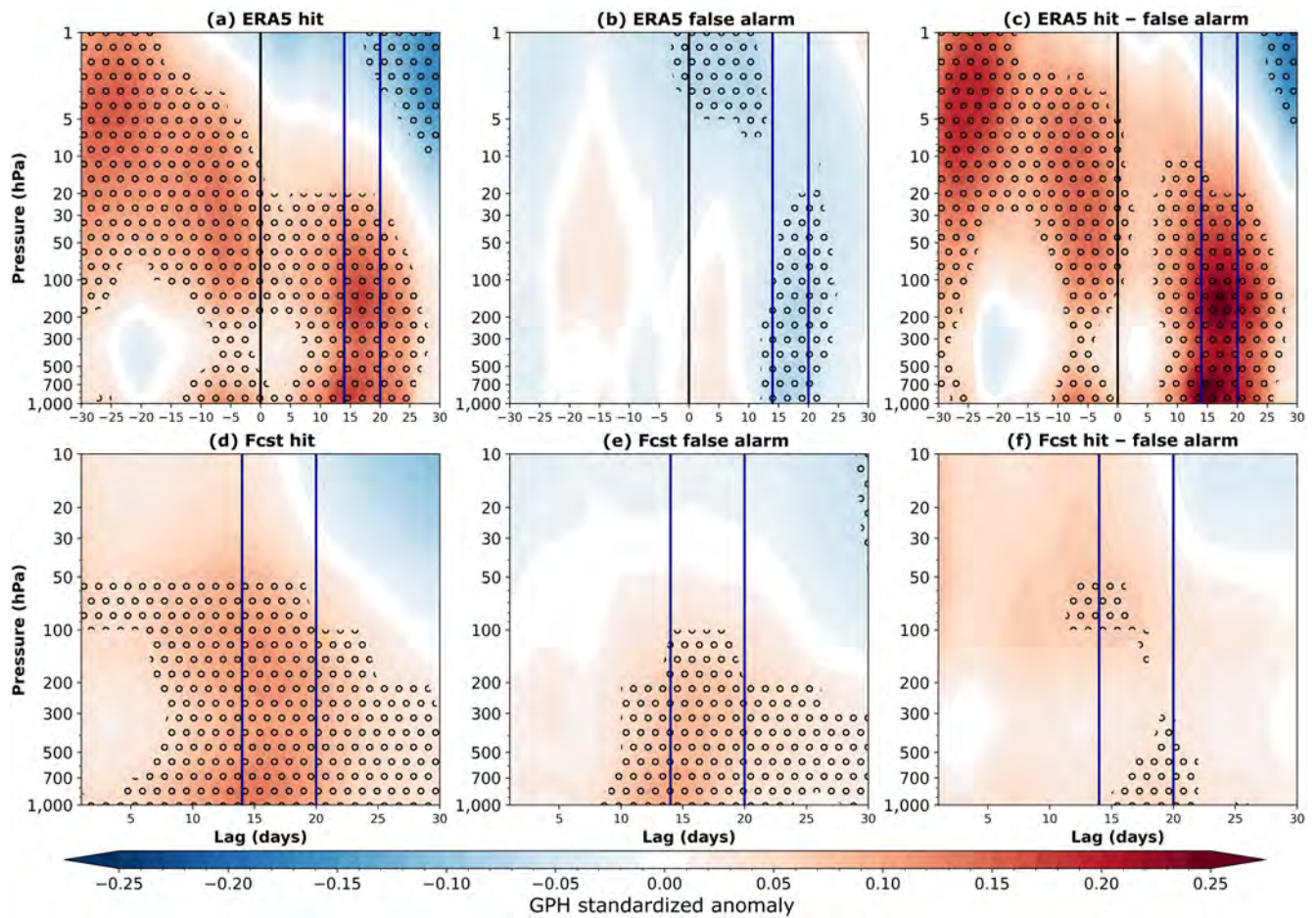


FIGURE 6 Composites of polar cap (60–90°N) averaged standardized geopotential height (GPH) anomalies (shading) as a function of lag from initialization and pressure level for week 3 cold spell confident hit forecasts, confident false alarm forecasts, and the difference between the confident hit and false alarm forecasts in (a)–(c) European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) and (d)–(f) the S2S models. Circles represent statistical significance at the 95th percentile via 5,000-iteration bootstrapping, the blue lines represent the week 3 cold, and the black solid line represents initialization date. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)] [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

Figure 6c,f presents the composite differences in polar cap GPH standardized anomalies between confident hit and false alarm forecasts in ERA5 (Figure 6c) and the subseasonal prediction systems (Figure 6f). These differences highlight the circulation features that help distinguish correct from incorrect forecasts when a week 3 cold spell is predicted. In ERA5, a notably weaker SPV, manifested as statistically significant positive GPH anomalies in the mid-to-upper stratosphere, separates confident hit forecasts from false alarms prior to initialization (Figure 6c). In the models, the same distinction holds: confident hit forecasts are characterized by persistent mid-to-upper stratospheric positive GPH anomalies, whereas false alarm forecasts lack this signal entirely (Figure 6f). These results suggest that confident subseasonal cold spell forecasts are more likely to be correct when preceded by a substantial weakening of the SPV. Conversely, confident false alarms appear to arise when forecasts are made in the

absence of strong stratospheric forcing, pointing to a more tropospheric origin of the predicted circulation pattern. Specifically, false alarms tend to be associated with the simulation of WCR- and AkR-like patterns, which may be driven by tropospheric dynamics or potentially shallower stratospheric processes, such as wave reflection (Messori *et al.*, 2022; Millin *et al.*, 2022; Schutte *et al.*, 2025).

3.4 | Confident forecasts under weak and strong SPV conditions

3.4.1 | Proportions of confident forecasts as a function of SPV strength

Before evaluating the reliability of subseasonal cold spell forecasts initialized in weak and strong SPV states, we

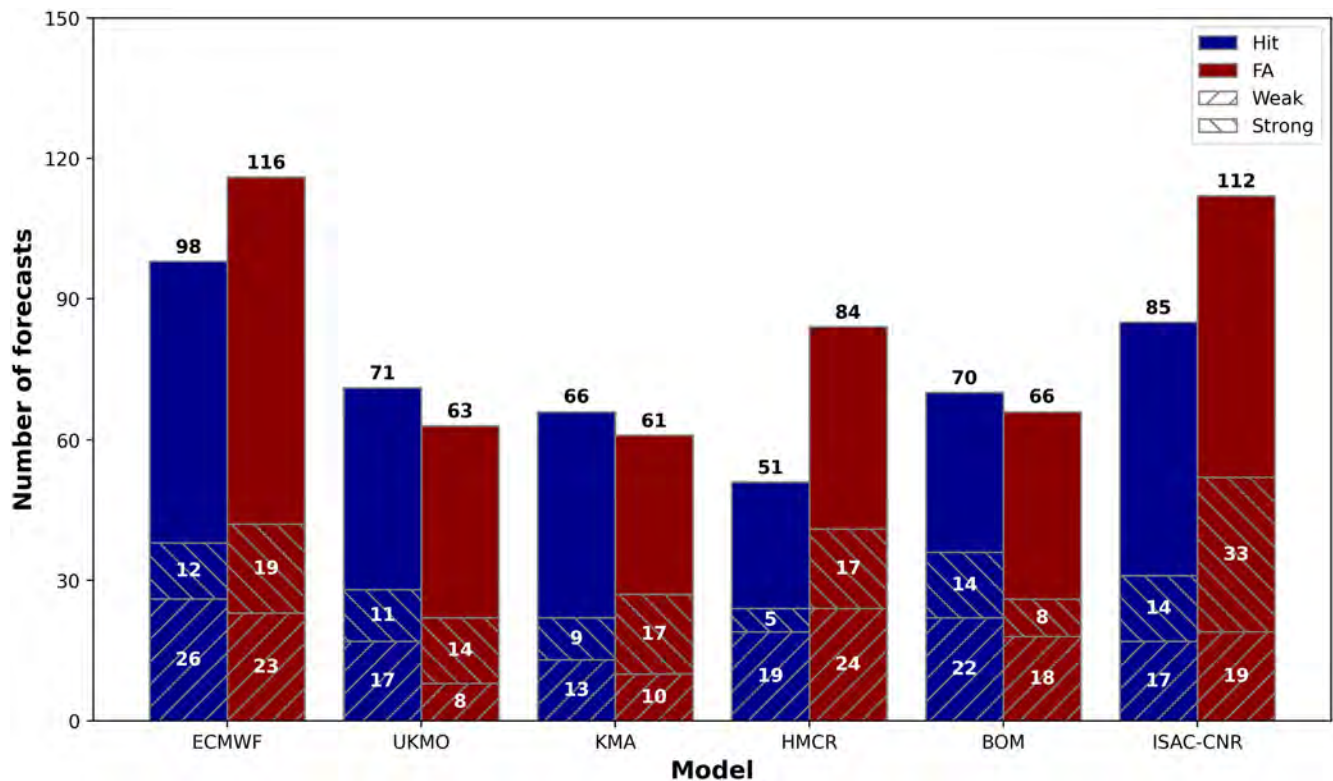


FIGURE 7 The total number of confident hit (blue solid bars) and confident false alarm (FA; red solid bars) forecasts as a function of subseasonal-to-seasonal model for week 3 cold spells. The left-to-right hatching indicates the proportion of confident hit and false alarm forecasts initialized in a weak stratospheric polar vortex (SPV) state whereas the right-to-left hatching indicates the proportion of confident hit and false alarm forecasts initialized in a strong SPV state. The bold numbers show the respective counts of these different proportions. BOM: Bureau of Meteorology; ECMWF: European Centre for Medium-Range Weather Forecasts; CNR-ISAC: Italian National Research Council–Institute of Atmospheric Sciences and Climate; HMCR: Hydro-Meteorological Centre of Russia; KMA: Korean Meteorological Agency; UKMO: United Kingdom Meteorological Office. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

examine how confident forecasts are distributed as a function of SPV strength. Figure 7 presents the number of confident hit forecasts (blue bars), confident false alarm forecasts (red bars), and the proportion of each that were initialized in either strong or weak SPV states (indicated by hatching). Among the models, ECMWF contains the highest number of both confident hits (98) and false alarms (116), likely due to its larger sample size (Table 1). Though the CNR-ISAC model has fewer forecasts overall than both ECMWF and BOM, its relatively high number of false alarms is consistent with its tendency toward overconfidence (Figure 3a). The BOM model, despite having 600 forecasts, produces substantially fewer confident hits and false alarms compared with ECMWF and CNR-ISAC. UKMO, KMA, and HMCR exhibit a similar number of confident hits and false alarms, although HMCR tends to produce more false alarms overall.

Importantly, Figure 7 illustrates that all models show more confident hit forecasts when initialized during weak SPV conditions than during strong SPV conditions. For instance, ECMWF generates 26 confident hits under weak

vortex conditions compared with just 12 under strong SPV conditions. In contrast, the number of confident false alarms under weak and strong SPV states is more balanced across the models. In fact, for UKMO, KMA, and ISAC-CNR, there are more false alarms initiated under strong versus weak SPV cases. These results suggest that weak SPV conditions may enhance forecast skill for week 3 central US cold spells (see also Figure 6). Although weak SPV states can still lead to false alarms, strong SPV conditions appear more likely to be associated with overconfidence, and ultimately, incorrect forecasts in these subseasonal prediction systems.

3.4.2 | Reliability

Figure 8 presents reliability diagrams and sharpness histograms for week 3 cold spell forecasts from the six chosen S2S models, stratified by whether forecasts were initialized under strong or weak SPV conditions. The week 3 reliability curves for forecasts initialized in a weak vortex

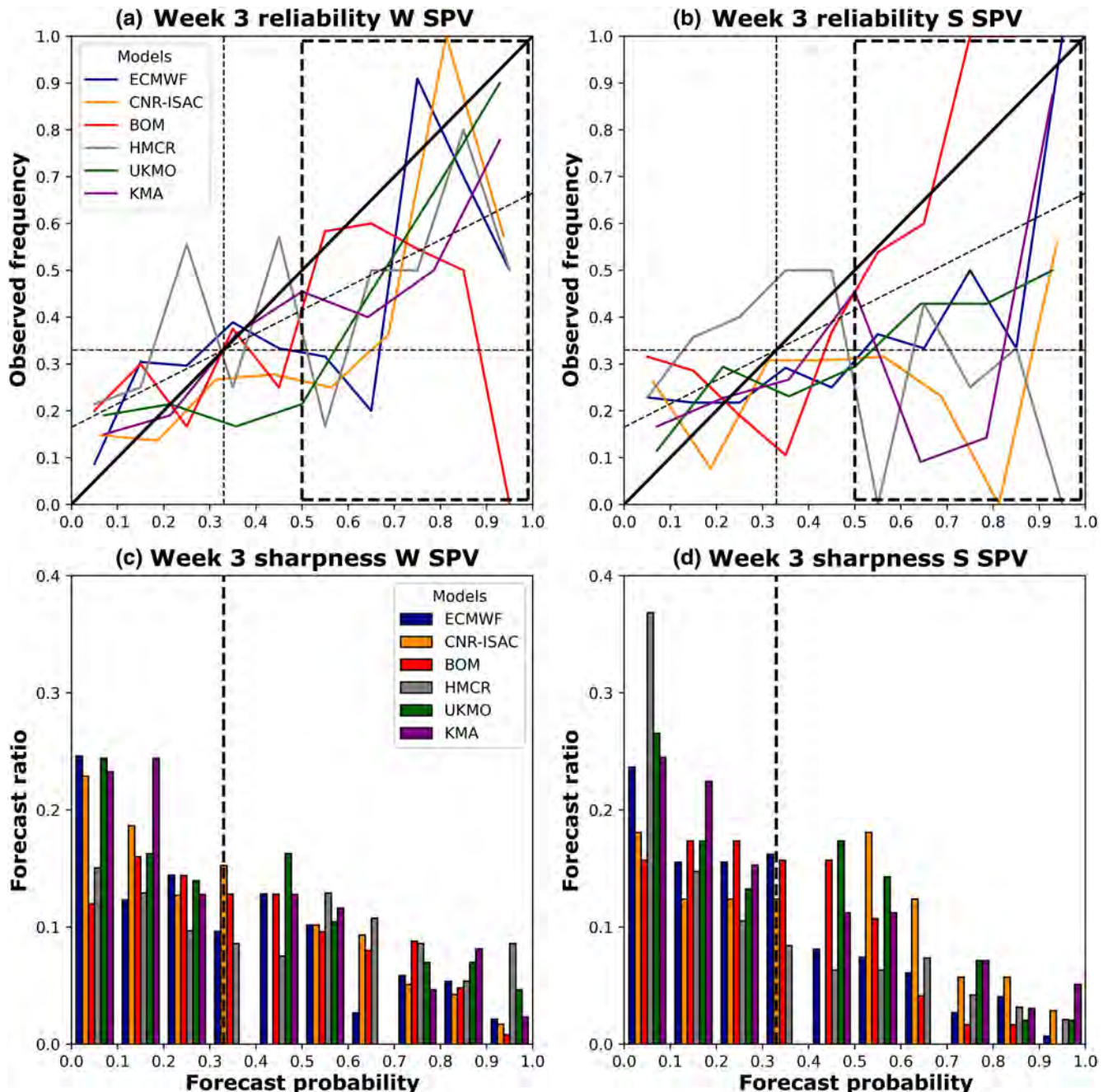


FIGURE 8 As in Figure 3 but the forecast reliability diagrams and sharpness histograms of week 3 cold spells initialized in (a, c) weak stratospheric polar vortex (SPV) conditions and (b, d) strong SPV conditions. Note that here the cumulative number of forecasts used for this analysis is 695 for forecasts initialized in a weak SPV and 665 for forecasts initialized in a strong SPV, compared with the total number of 3454 used in Figure 3. BOM: Bureau of Meteorology; ECMWF: European Centre for Medium-Range Weather Forecasts; CNR-ISAC: Italian National Research Council–Institute of Atmospheric Sciences and Climate; HMCR: Hydro-Meteorological Centre of Russia; KMA: Korean Meteorological Agency; UKMO: United Kingdom Meteorological Office. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

(Figure 8a) show a noticeable shift toward the perfect reliability line compared with the reliability for all week 3 forecasts (Figure 3a). These curves are also more consistently sloped from bottom left to upper right, particularly for forecast probabilities above 0.5, indicating that a weak SPV at initialization improves both the confidence

and accuracy of cold spell predictions. The corresponding sharpness histogram for week 3 cold spells initialized in weak SPV states (Figure 8c) shows the highest concentration of forecasts in the lowest probability bins (< 0.3), but also displays a more even distribution across the full range of forecast probabilities compared with

all forecasts (Figure 8c). This broader spread, including a higher frequency of high-probability forecasts, reflects improved sharpness when forecasts are initialized in weak SPV conditions.

We also analyzed the reliability and sharpness of week 4 cold spell forecasts initialized with a weak SPV (not shown). The week 4 reliability curves similarly show improvement relative to all forecasts, particularly at higher forecast probabilities. However, these curves are generally flatter than their week 3 counterparts, indicating more limited skill at longer lead times. The week 4 sharpness histograms mirror those of week 3 in shape. However, the distributions are flatter, with more forecasts clustered near the climatological frequency (0.33) and some increase in high-probability forecasts, indicating modest gains in sharpness relative to Figure 3d. Overall, though initializing in a weak SPV enhances both skill and sharpness for week 4 cold spell forecasts, the improvements are smaller than those seen in week 3.

The reliability of week 3 cold spell forecasts initialized with strong SPV conditions shows an opposite pattern to that seen for forecasts initialized with weak SPV conditions (Figure 8b). Forecasts initialized with strong SPV conditions are generally overconfident, with reliability curves shifted toward the lower right of the diagram; some models, such as ECMWF, KMA, and BOM, display more reliable behavior at higher forecast probabilities than the others. The sharpness histograms for week 3 forecasts initialized with a strong SPV (Figure 8d) indicate some degree of sharpness, with a broad spectrum of forecast probabilities. However, the number of forecasts exceeding a probability of 0.7 is notably lower than for forecasts initialized with a weak SPV (Figure 8c), suggesting that strong SPV conditions reduce the issuance of high probability week 3 cold spell forecasts. Additionally, week 3 cold spell forecasts initialized with a strong SPV show a higher frequency of low-probability predictions, further reflecting reduced probabilistic confidence and skill. Week 4 cold spell forecasts initialized with strong SPV conditions exhibit similar behavior to week 3 forecast counterparts (not shown), with more clustering around the climatological frequency in the sharpness histograms. This behavior implies that week 3 cold spell forecasts initialized under a strong SPV are overall less reliable.

3.4.3 | Tropospheric circulation and weather regimes

We now repeat the analysis of Figure 4, focusing on confident hit and confident false alarm forecasts that were initialized in either weak or strong SPV conditions. Figure 9 shows the composite 500 hPa GPH anomalies

(contours) and T2M anomalies (shading) for confident forecasts initialized during weak SPV states. For both confident hits and false alarms, the ERA5 mean pattern at initialization is similar, with a characteristic weak SPV signature; anomalous ridging over the Arctic (Figure 9a,d). In both cases, the central US features negative T2M anomalies and anomalous troughing, along with negative T2M anomalies across northern Eurasia, highlighting the canonical weak SPV-related cold response in that region (Butler *et al.*, 2017). A key difference between the confident hit and false alarm forecasts in weak SPV states is the magnitude of Arctic GPH anomalies at initialization, which is stronger in false alarms ($\sim +8$ dam) than in the hits ($\sim +6$ dam). This suggests that more pronounced Arctic ridging at initialization may bias the model toward forecasting central US cold spells later (Figure 9a,d). Week 3 verification fields for hits show negative T2M anomalies and anomalous troughing in the central US, alongside strong anomalous ridging over northern Canada and Greenland (~ 8 dam), connecting with anomalous ridging over Alaska. This pattern is indicative of the ArH regime and a negative NAO-like structure (Figure 9b).

In contrast, the verifying mean state for confident false alarm forecasts reveals anomalous warmth across North America, and the anomalous ridging signal over northern Canada and Greenland is weaker and less coherent. The lack of an AkR-like signal in this case suggests that both AkR and ArH regimes may be important features following a weak SPV, consistent with prior findings (Figure 9e; Millin *et al.*, 2022). The corresponding week 3 patterns for model confident hit forecasts closely resemble the observed ERA5 pattern (Figure 9c). By contrast, the forecast pattern for false alarms features weaker and more high-latitude-confined ridging, implying that simulating a stronger and more spatially extensive Arctic ridging enhances forecast accuracy for week 3 cold spells (Figure 9f). Finally, a notable feature across all panels in Figure 9 is the consistent presence of negative T2M anomalies over northern Eurasia, suggesting the potential for compound cold events affecting both Eurasia and the central US following weak SPV conditions.

Figure 10 is the same as Figure 9 but for strong SPV initial conditions. The first notable feature is that the number of confident hit forecasts under strong SPV conditions is about half of that of weak SPV cases, suggesting less favorable conditions for skillful long-lead forecasts of central US cold spells. At initialization time, in ERA5, both confident hit and false alarm forecasts exhibit strong anomalous troughing over the Arctic (< -6 dam), a hallmark of strong SPV states (Figure 10a,d). The confident hit forecasts show the large negative GPH anomalies across the Arctic at

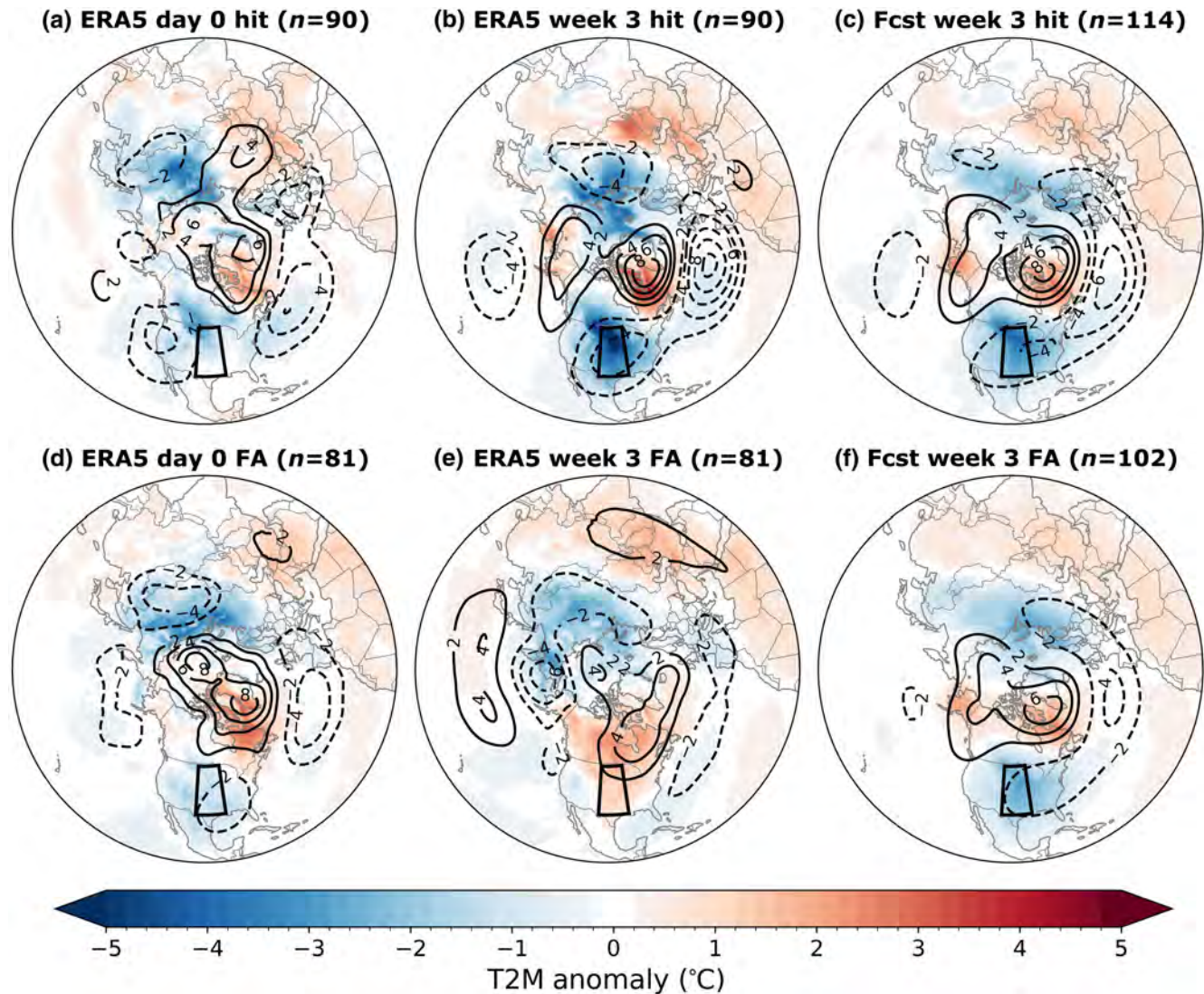


FIGURE 9 As in Figure 4 but for forecasts initialized in weak polar vortex states. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

this stage. In the verifying week 3 ERA5 fields for confident hit forecasts, strong negative T2M anomalies develop over the central US, co-occurring with a well-defined wave train pattern: anomalous ridging over Alaska and Eurasia flanked by deep anomalous troughing in the North Pacific and central/eastern US (Figure 10b). The model forecast captures much of this structure, though ridging is more isolated in a pure AkR-like configuration, lacking the strong North Pacific troughing associated with the WCR pattern (Figure 10c). This configuration resembles that seen during AkR CAOs over the Great Plains (Millin *et al.*, 2022).

In contrast, confident false alarm forecasts initialized under strong SPV conditions generally feature a weaker wave pattern during week 3, appearing as a broad dipole over North America, with anomalous ridging over Alaska and anomalous troughing over Canada (Figure 10f). The

verifying ERA5 state for these false alarms in week 3 maintains a pattern similar to that at initialization, with a tightly confined Arctic trough and positive T2M anomalies across the central US (Figure 10e). These findings suggest that the strength and structure of the simulated wave pattern, and potentially the configuration of the SPV, play important roles in whether S2S models can confidently and correctly predict week 3 cold spells under strong SPV conditions.

The weather regime analyses for confident hit and false alarm forecasts of week 3 cold spells, stratified by SPV strength, reinforce the circulation-based results discussed previously (Figures 11 and 12). In ERA5, confident hit forecasts initialized in weak SPV conditions exhibit a markedly enhanced frequency of the ArH regime, which becomes apparent as early as 2 weeks prior to initialization and increases to nearly 50% in the days leading

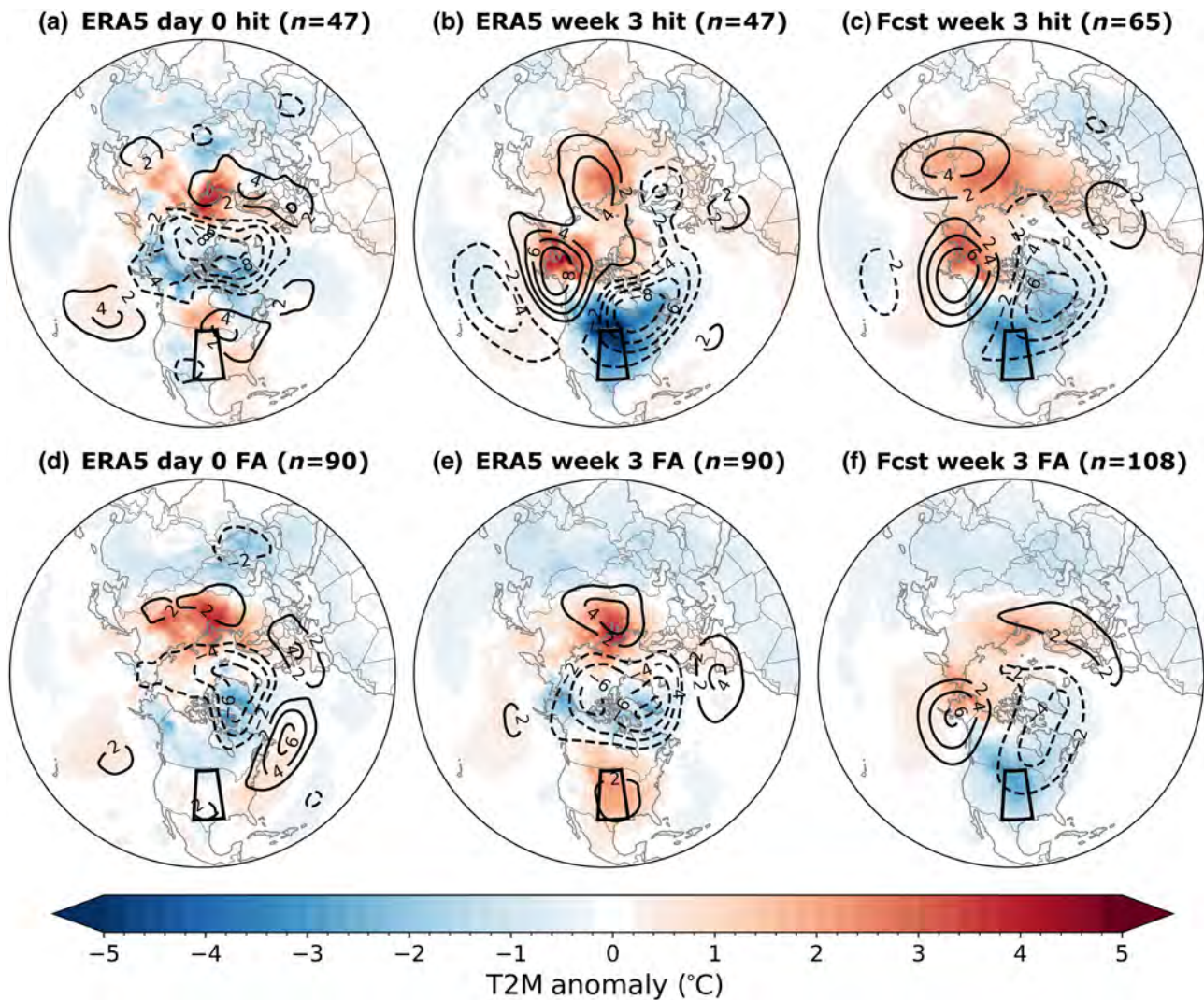


FIGURE 10 As in Figure 4, but for forecasts initialized during strong stratospheric polar vortex states. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

up to the cold spell (Figure 11a). During this period, the other weather regimes show little or no change in occurrence, underscoring the dominant role of the ArH regime in the lead up to week 3 confident hit cold spells following weak SPV conditions. Consistently, the model forecasts for this same group also simulate significantly elevated ArH regime frequencies (Figure 11b). This alignment between the observed and simulated regime evolution highlights the importance of capturing the emergence and persistence of the ArH regime on subseasonal timescales. Indeed, the tropospheric circulation pattern is very unlikely to move away from the ArH regime under weak SPV conditions (Lee *et al.*, 2022).

For the false alarm confident forecasts initialized with a weak SPV, the ArH regime is consistently observed at frequencies of around 30% for approximately 4 weeks before initialization (Figure 11c). However, after initialization, the ArH regime diminishes in frequency, allowing the PT

regime to dominate the cold week period in ERA5. The PT regime is not conducive to extreme cold in the central US and explains why anomalous warmth across the central US tends to occur in the verification (Figure 9e). Interestingly, in the models, the confident false alarm forecasts start with a relatively lower ArH frequency than the confident hit forecasts. However, this frequency increases as the cold week approaches, eventually exceeding the ArH frequency seen in confident hit forecasts (Figure 11d). This suggests that the persistence of the ArH regime prior to week 3 cold spells is a key differentiator between confident hit and confident false alarm forecasts initialized with a weak SPV.

The evolution of regime frequencies in ERA5 for confident hit forecasts initialized with strong SPV conditions reveals that, prior to initialization, the ArL regime is a favorable weather regime (Figure 12a). In the days leading up to the week 3 confident hit cold spells in ERA5,

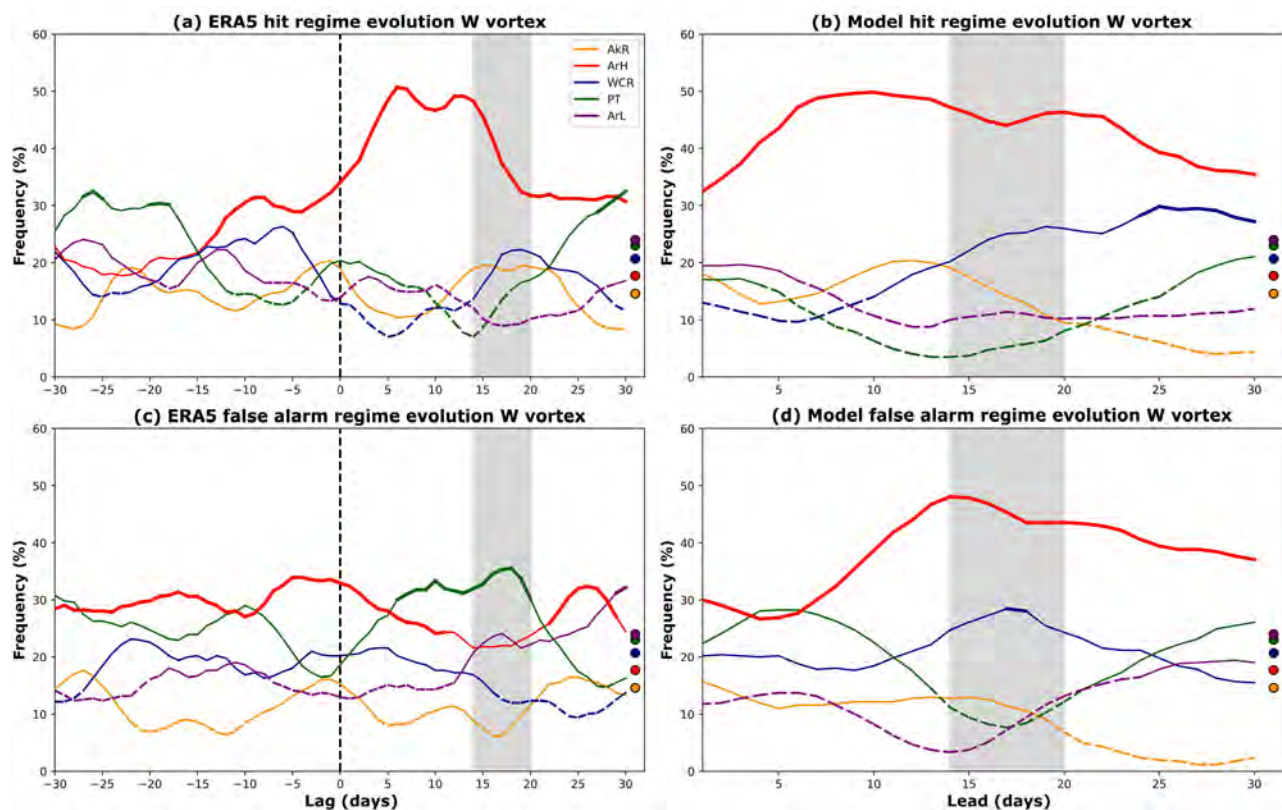


FIGURE 11 As in Figure 5, but for forecasts initialized in weak polar vortex states. [Colour figure can be viewed at wileyonlinelibrary.com]

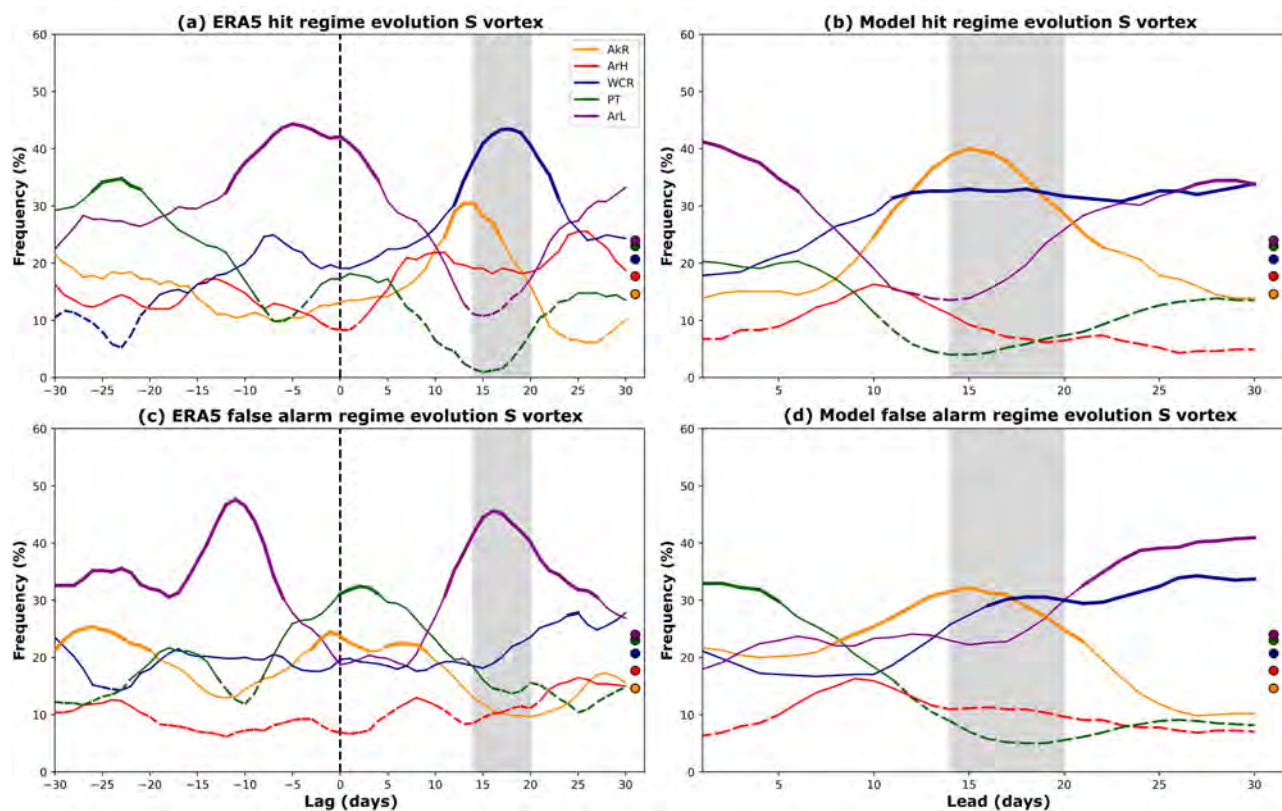


FIGURE 12 As in Figure 5, but for forecasts initialized in strong stratospheric polar vortex states. [Colour figure can be viewed at wileyonlinelibrary.com]

the regime frequencies shift, with the WCR and AkR regimes becoming dominant and significantly enhanced, reaching frequencies between 30% and 40%. Similar to the observed evolution, the confident hit model forecasts show an increase in the frequency of the AkR and WCR regimes but also simulate AkR conditions more frequently than in ERA5 (Figure 12b). This difference in the dominant regime frequency is likely attributable to the positioning of the GPH anomalies shown in Figures 10b,c. Nevertheless, the rapid but significant transition to WCR and AkR regimes in the confident hit forecasts initialized under strong SPV conditions mirrors the synoptic-scale development of these features, potentially associated with stratospheric wave reflection (Millin *et al.*, 2022).

The confident false alarm forecasts reveal that the ArL regime is significantly enhanced up to approximately 5 days prior to initialization in ERA5, after which the PT regime frequency becomes enhanced around initialization (Figure 12c). In contrast to the confident hit forecasts, the ArL regime dominates in ERA5 during the cold week. Meanwhile, the model forecasts show significantly enhanced frequencies of the AkR and WCR regimes

during the cold week (Figure 12d). Interestingly, the model forecast regime frequencies show an evolution that results in a transition from the PT to the AkR and WCR regimes, which are very unlikely transitions under strong SPV conditions (Lee *et al.*, 2022).

3.4.4 | Stratospheric signals for forecasts initialized in weak and strong vortex

To further explore the differences between forecast groups, Figure 13 shows composites of polar-cap-averaged standardized GPH anomalies in both ERA5 and the model forecast for confident hit and false alarm week 3 cold spell forecasts initialized under weak SPV conditions. In ERA5, the weak SPV hit group exhibits strong and statistically significant positive GPH anomalies throughout the mid-to-upper stratosphere at initialization, particularly between 10 and 50 hPa (Figure 13a). Several “drips” of significant positive GPH standardized anomalies descend into the troposphere prior to initialization and become increasingly persistent approaching the cold

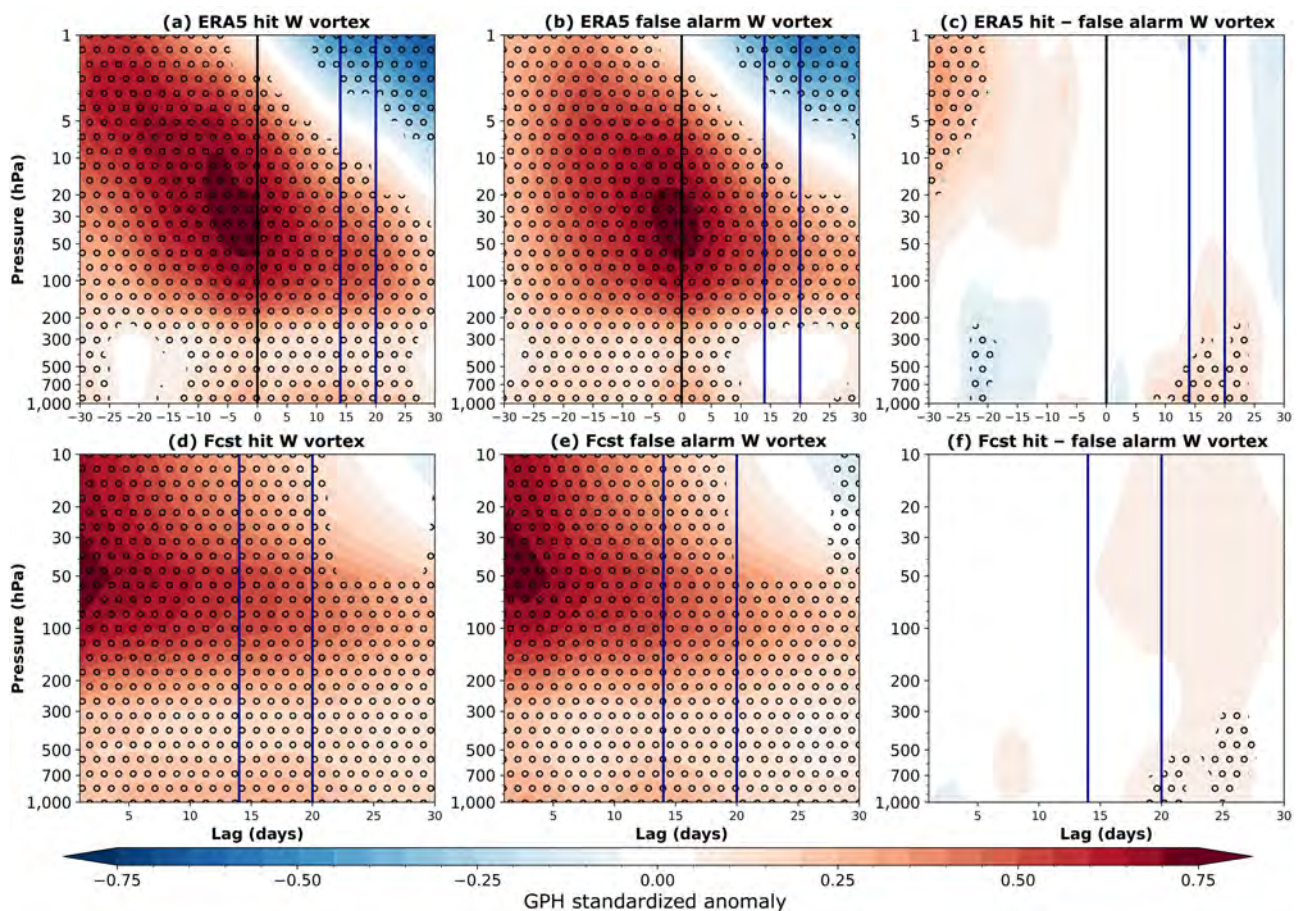


FIGURE 13 As in Figure 6, but for forecasts initialized during weak polar vortex states. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

spell, consistent with the sustained ArH regime signal seen in Figure 11a. The corresponding ERA5 composite for the weak SPV false alarm group shows a broadly similar stratospheric evolution (Figure 13b) but without positive GPH standardized anomalies in the troposphere during week 3. The most pronounced difference between the ERA5 composites emerges in the upper stratosphere approximately 3–4 weeks prior to initialization, during which the hit forecasts exhibit substantially stronger positive GPH standardized anomalies between 1 and 10 hPa (Figure 13c). This difference suggests that a key distinguishing factor between the two groups is the earlier onset and greater intensity of upper stratospheric weakening in the hit forecasts. In contrast, the difference in the model forecast evolution between the two groups is minimal (Figure 13d–f), apart from the hit group showing a slightly larger near-surface positive GPH anomaly beyond week 3. Hence, false alarm forecasts initialized under weak SPV conditions tend to overestimate the extent of downward coupling.

Figure 14 presents composites of polar-cap-averaged standardized GPH anomalies in both ERA5 and the

model forecast for confident hit and false alarm week 3 cold spell forecasts initialized under strong SPV conditions. In ERA5, the hit forecast group shows a more isolated negative GPH anomaly confined to the mid-to-lower stratosphere, emerging approximately 10 days prior to initialization and persisting until the week 3 cold period (Figure 14a). This signal resembles the wave-reflection-driven SPV signature associated with AkR-type central US CAOs (Millin *et al.*, 2022). The ERA5 composite for the false alarm group displays a stronger and more vertically extensive SPV signal throughout the stratosphere, which is also more persistent prior to initialization. However, this does not translate into a positive tropospheric GPH anomaly during the week 3 period in ERA5 (Figure 14b), resulting in a statistically significant positive anomaly difference between the two groups in the troposphere (Figure 14c). The model forecast fields for both groups are similar, exhibiting strong SPV conditions at initialization and a more positive GPH standardized anomaly in the troposphere during week 3 (Figure 14d–f). Taken together, these results suggest that the structure and timing of the SPV preconditioning may be critical.

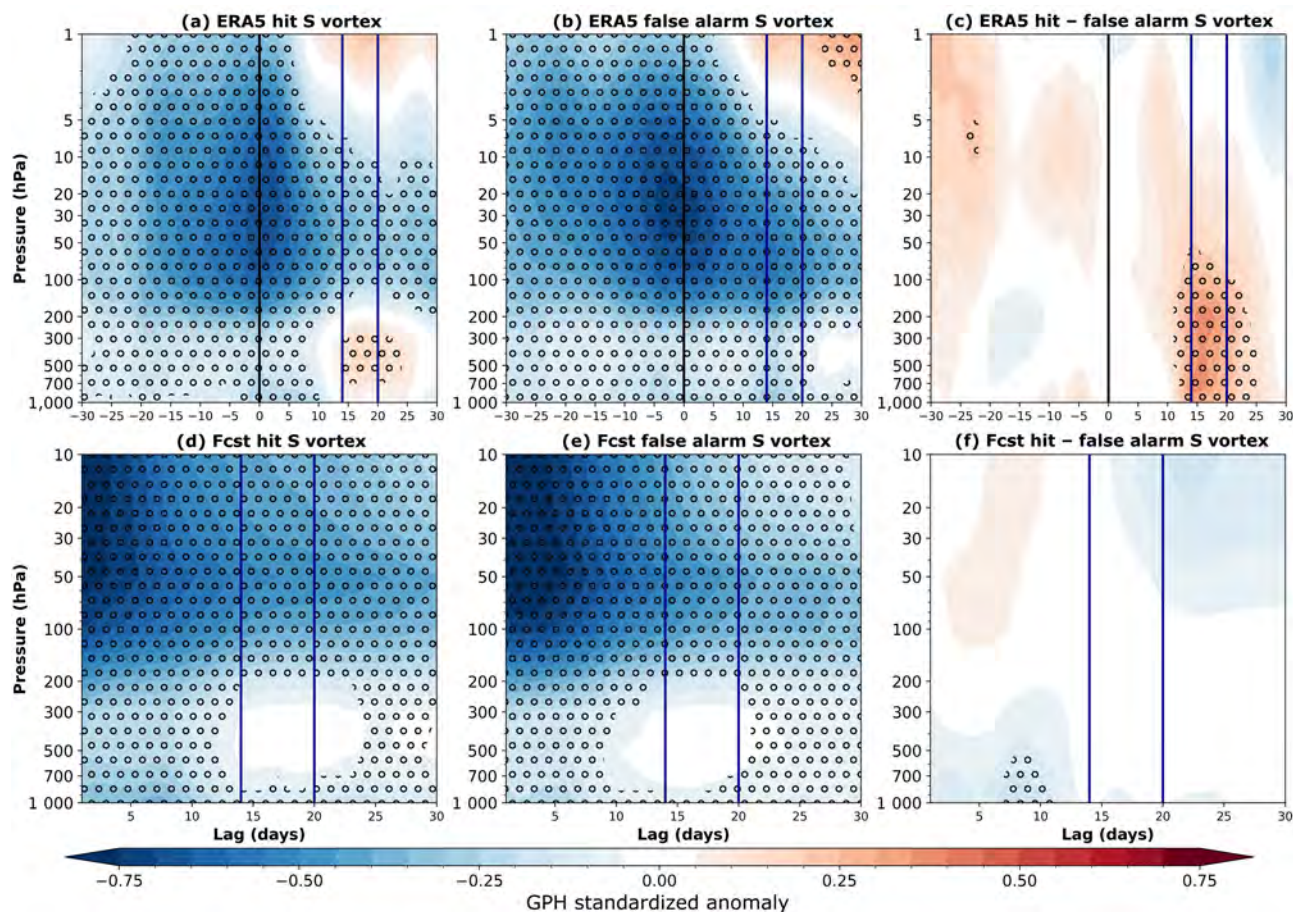


FIGURE 14 As in Figure 6 but for forecasts initialized in strong polar vortex states. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/qj.70081)]

A more shallow and isolated stratospheric anomaly, potentially linked to wave reflection near initialization, may distinguish hit forecasts. In such cases, this isolated strong SPV signature may increase the curvature of the zonal wind profile over Canada, creating a stratospheric reflecting surface (Matthias & Kretschmer, 2020; Millin *et al.*, 2022; Perlwitz & Harnik, 2003), which could allow upward-propagating planetary-scale waves to be reflected back into the troposphere, reinforcing existing circulation anomalies and promoting North Pacific/Alaskan anomalous ridging. Conversely, the false alarm cases may represent situations where the model attempts to predict a weather regime transition away from the PT regime toward the AkR regime in response to the strong SPV. In reality, such transitions during strong SPV conditions are less likely to occur (Lee *et al.*, 2022), especially when the strong vortex state is deeply established, potentially inhibiting downward coupling or suppressing the necessary tropospheric precursors for a regime change. Additionally, false alarms initialized in strong SPV conditions could also be a result of the model falsely simulating wave reflection conditions, which should be studied further. Overall, our interpretations are consistent with prior work highlighting that not all strong SPV states have the same tropospheric impact; rather, the vertical structure of the SPV anomaly and the presence or absence of reflection events may shape whether the stratosphere reinforces or fails to project onto the troposphere.

4 | CONCLUSIONS

This study provides a comprehensive assessment of the predictability of cold spells in the central US in six S2S models over the period 1993–2020. We find that forecasts of central US cold spells exhibit a rapid degradation in HSS, POD, FAR, and MR beyond week 2 (Figure 2). The BOM model generally exhibits the lowest skill; this model is also the one with the lowest model lid (10 hPa) and coarser native resolution than the others. Indeed, differences in forecast skill among models may also stem from systematic biases in the representation of the SPV and in the frequency and position of atmospheric blocking, the latter being a well-documented challenge in S2S models (Quinting & Vitart, 2019). Additionally, increased model vertical and horizontal resolution and improvements in the representation of quasi-stationary waves and diabatic heating could further enhance the skill of forecasts for central US cold spells and their associated weather regimes. Confident hit week 3 cold spell forecasts are characterized by a consistent dynamical evolution: (a) enhanced ArH regime frequency prior to onset (Figures 4a, 5a, and 5b), (b) persistence of ArH, AkR, and/or WCR regimes during

the cold spell (Figure 5a,b), and (c) a gradually weakening SPV signal propagating downward subseasonally (Figure 6a,d). These findings align with prior work, linking ArH-dominated regimes and stratospheric weakening to CAO development (Lee *et al.*, 2019; Millin *et al.*, 2022). In contrast, confident false alarms typically lack a coherent stratospheric precursor and are instead associated with shallow, tropospheric ridging (Figure 6b,e), underscoring the role of stratosphere–troposphere coupling in enabling skillful subseasonal cold spell forecasts.

Stratifying forecasts by the state of the SPV at initialization further reveals that weak SPV conditions are associated with improved reliability in weeks 3 and 4, whereas strong SPV conditions tend to degrade forecast quality (Figure 8). This finding echoes earlier findings over Europe (e.g., Erner & Karpechko, 2024). Specifically, confident hits initialized in weak SPV conditions are distinguished by earlier and more persistent SPV weakening and sustained ArH regime conditions, whereas false alarms initialized in weak SPV conditions lack this long-lived precursor signal (Figures 11 and 13). Meanwhile, confident hits initialized in strong SPV conditions resemble wave-reflection-driven CAOs, showing an isolated strong SPV signal near initialization and forecast transitions to the AkR or WCR regime. In contrast, false alarms initialized in strong SPV conditions show an unrealistic regime transition from the PT regime to the AkR regime, highlighting model deficiencies in representing regime evolution under strong vortex conditions (Figures 12 and 14). These findings emphasize the potential of the SPV state as a key predictor of subseasonal forecast confidence. In particular, a weakened SPV may offer a “window of opportunity” for skillful week 3–4 cold spell forecasts, especially when paired with a persistent ArH. Furthermore, the regimes emphasized here, especially the ArH, AkR, and WCR regimes, are connected to but not equivalent to teleconnections such as the NAO and the Pacific–North American (PNA) pattern. For instance, the ArH regime captures a more pan-Arctic circulation than the NAO does, and both the ArH and AkR regimes also exhibit unique linkages to SPV variability (Lee *et al.*, 2019; Millin *et al.*, 2022). These distinctions highlight that though our identified regimes share influences with teleconnections such as the NAO and the PNA, they provide complementary perspectives on the teleconnections shaping central US temperature variability. Future research should explore how wave reflection and regime transition probabilities influence forecast evolution. Operationally, incorporating SPV state and regime tendencies into probabilistic forecast interpretation could enhance the reliability of extended-range cold spell guidance in North America.

Although our results suggest a possible role for wave reflection in strong SPV cases, we did not explicitly

quantify this mechanism. Future work should compute composites of a measure of stratospheric wave reflection (e.g., the reflection index; Messori *et al.*, 2022) to assess its role more directly. Additionally, incorporating more models from the full S2S Project Database could improve robustness of the results, with the caveat of rectifying ensemble sizes and hindcast periods. Figure 9 also suggests the potential for compound cold spells across North America and Eurasia; the role of a weak SPV in such events merits further investigation. Finally, although we focused on regime and stratospheric influences, future studies should examine the role of other modes of variability, particularly tropical modes of variability, in modulating cold spell prediction skill. For example, Büeler *et al.* (2021) demonstrated that European weather regime forecast skill improves following phases 4 and 7 of the Madden–Julian Oscillation, but is reduced after phase 2. This shows the value of examining influences of tropical convection not only on central US cold spell prediction but also on the associated North American weather regimes.

ACKNOWLEDGEMENTS

We thank the National Science Foundation (NSF) Oklahoma Established Program to Stimulate Competitive Research (EPSCoR) project grant OIA-1946093 for the financial support to enable the completion of this work. We also acknowledge that this work has been partly completed using the Casper supercomputer at the National Center for Atmospheric Research (CISL, 2023).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

ERA5 data can be downloaded from <https://cds.climate.copernicus.eu/datasets> and S2S model reforecasts can be downloaded from <https://apps.ecmwf.int/datasets/data/s2s/levtype=sfc/type=cf/>. All code for analyses can be found at <https://doi.org/10.5281/zenodo.18100939>.

ORCID

Oliver T. Millin  <https://orcid.org/0000-0002-9033-763X>

REFERENCES

- Afarghan-Gerstman, H. & Domeisen, D.I.V. (2020) Pacific modulation of the North Atlantic storm track response to sudden stratospheric warming events. *Geophysical Research Letters*, 47(2), 1–10. Available from: <https://doi.org/10.1029/2019GL085007>
- Baldwin, M.P. & Dunkerton, T.J. (2001) Stratospheric harbingers of anomalous weather regimes. *Science*, 294(5542), 581–584. Available from: <https://doi.org/10.1126/science.1063315>
- Bloomfield, H.C., Brayshaw, D.J., Gonzalez, P.L.M. & Charlton-Perez, A. (2021) Pattern-based conditioning enhances sub-seasonal prediction skill of European national energy variables. *Meteorological Applications*, 28(4), 1–16. Available from: <https://doi.org/10.1002/met.2018>
- Brady, R.X. & Spring, A. (2021) Climpred: verification of weather and climate forecasts. *Journal of Open Source Software*, 6(59), 1–5.
- Büeler, D., Ferranti, L., Magnusson, L., Quinting, J.F. & Grams, C.M. (2021) Year-round sub-seasonal forecast skill for Atlantic–European weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 147(741), 4283–4309. Available from: <https://doi.org/10.1002/qj.4178>
- Butler, A.H., Sjöberg, J.P., Seidel, D.J. & Rosenlof, K.H. (2017) A sudden stratospheric warming compendium. *Earth System Science Data*, 9(1), 63–76. Available from: <https://doi.org/10.5194/essd-9-63-2017>
- CAWCR. (2015) *Forecast verification*. [Accessed: 27th May 2025] <https://www.cawcr.gov.au/projects/verification/>
- Cellitti, M.P., Walsh, J.E., Rauber, R.M. & Portis, D.H. (2006) Extreme cold air outbreaks over the United States, the polar vortex, and the large-scale circulation. *Journal of Geophysical Research – Atmospheres*, 111(D2), 1–14. Available from: <https://doi.org/10.1029/2005JD006273>
- Charlton, A.J. & Polvani, L.M. (2007) A new look at stratospheric sudden warmings. part I: climatology and modeling benchmarks. *Journal of Climate*, 20(3), 449–469. Available from: <https://doi.org/10.1175/JCLI3996.1>
- CISL. (2023) *Derecho: HPE cray EX system (university community computing)*. Boulder, CO: NSF National Center for Atmospheric Research. Available from: <https://doi.org/10.5065/qx9a-pg09>
- Doss-Gollin, J., Farnham, D.J., Lall, U. & Modi, V. (2021) How unprecedented was the February 2021 Texas cold snap? *Environmental Research Letters*, 16(6), 1–7. Available from: <https://doi.org/10.1088/1748-9326/ac0278>
- Erner, I. & Karpechko, A. (2024) Factors influencing subseasonal predictability of northern Eurasian cold spells. *Quarterly Journal of the Royal Meteorological Society*, 150(762), 2955–2975. Available from: <https://doi.org/10.1002/qj.4744>
- Erner, I., Karpechko, A., Kämäräinen, M. & Järvinen, H. (2022) Stratosphere–troposphere coupling enhances subseasonal predictability of Northern Eurasian cold spells. *Quarterly Journal of the Royal Meteorological Society*, 148, 2769–2783. Available from: <https://doi.org/10.1002/qj.4335>
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. Available from: <https://doi.org/10.1002/qj.3803>
- Hitchcock, P. & Simpson, I.R. (2014) The downward influence of stratospheric sudden warmings. *Journal of the Atmospheric Sciences*, 71(10), 3856–3876. Available from: <https://doi.org/10.1175/JAS-D-14-0012.1>
- Huang, J., Hitchcock, P., Tian, W. & Sillin, J. (2022) Stratospheric influence on the development of the 2018 late winter European cold air outbreak. *Journal of Geophysical Research – Atmospheres*, 127(5), 1–21. Available from: <https://doi.org/10.1029/2021JD035877>
- Kautz, L., Polichtchouk, I., Birner, T., Garny, H. & Pinto, J.G. (2019) Enhanced extended-range predictability of the 2018 late-winter Eurasian cold spell due to the stratosphere. *Quarterly Journal of the Royal Meteorological Society*, 146(727), 1040–1055. Available from: <https://doi.org/10.1002/qj.3724>

- Lee, S.H., Furtado, J.C. & Charlton-Perez, A.J. (2019) Wintertime North American weather regimes and the arctic stratospheric polar vortex. *Geophysical Research Letters*, 46(24), 14 892–14 900. Available from: <https://doi.org/10.1029/2019GL085592>
- Lee, S.H., Charlton-Perez, A.J., Woolnough, S.J. & Furtado, J.C. (2022) How do stratospheric perturbations influence North American weather regime predictions? *Journal of Climate*, 35(18), 5915–5932. Available from: <https://doi.org/10.1175/JCLI-D-21-0413.1>
- Liu, M., Hu, D. & Guan, Z. (2022) Modulation of a long-lasting extreme cold event in Siberia by a minor sudden stratospheric warming and the dynamical mechanism involved. *Climate Dynamics*, 60, 797–811. Available from: <https://doi.org/10.1007/s00382-022-06353-7>
- Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C. et al. (2020) Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, 101(5), E608–E625. Available from: <https://doi.org/10.1175/BAMS-D-18-0326.1>
- Matthias, V. & Kretschmer, M. (2020) The influence of stratospheric wave reflection on north american cold spells. *Monthly Weather Review*, 148(4), 1675–1690. Available from: <https://doi.org/10.1175/MWR-D-19-0339.1>
- Messori, G., Kretschmer, M., Lee, S.H. & Wendt, V. (2022) Stratospheric downward wave reflection events modulate north American weather regimes and cold spells. *Weather and Climate Dynamics*, 3(4), 1215–1236. Available from: <https://doi.org/10.5194/wcd-3-1215-2022>
- Met Office, 2025: *Reliability and sharpness diagrams*. [Accessed: 22nd May 2025], <https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/user-guide/interpret-reliability:text=Reliability%20diagrams%20provide%20additional%20information,chance%20of%20observing%20the%20event>
- Millin, O.T. & Furtado, J.C. (2022) The role of wave breaking in the development and subseasonal forecasts of the February 2021 Great Plains cold air outbreak. *Geophysical Research Letters*, 49(21), 1–10. Available from: <https://doi.org/10.1029/2022GL100835>
- Millin, O.T., Furtado, J.C. & Basara, J.B. (2022) Characteristics, evolution, and formation of cold air outbreaks in the Great Plains of the United States. *Journal of Climate*, 35(14), 4585–4602. Available from: <https://doi.org/10.1175/JCLI-D-21-0772.1>
- Millin, O.T., Furtado, J.C. & Malloy, C. (2024) The impact of North American winter weather regimes on electricity load in the central United States. *npj Climate and Atmospheric Science*, 7, 254. Available from: <https://doi.org/10.1038/s41612-024-00803-1>
- Nakagawa, K.I. & Yamazaki, K. (2006) What kind of stratospheric sudden warming propagates to the troposphere? *Geophysical Research Letters*, 33(4), 1–4. Available from: <https://doi.org/10.1029/2005GL024784>
- NCEI. (2025) *Billion-dollar weather and climate disasters: events*. [Accessed: 27th February 2025] <https://www.ncei.noaa.gov/access/billions/events>
- Pedragose, F. & Coauthors. (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Perlitz, J. & Harnik, N. (2003) Observational evidence of a stratospheric influence on the troposphere by planetary wave reflection. *Journal of Climate*, 16(18), 3011–3026. Available from: [https://doi.org/10.1175/1520-0442\(2003\)016<3011:OEOASI>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<3011:OEOASI>2.0.CO;2)
- Quinting, J. & Vitart, F. (2019) Representation of synoptic-scale Rossby wave packets and blocking in the S2S prediction project database. *Geophysical Research Letters*, 46(2), 1070–1078. Available from: <https://doi.org/10.1029/2018GL081381>
- Schutte, M.K., Portal, A., Lee, S.H. & Messori, G. (2025) Dynamics of stratospheric wave reflection over the north pacific. *Weather and Climate Dynamics*, 6(2), 521–548. Available from: <https://doi.org/10.5194/wcd-6-521-2025>
- Smith, E.T. & Sheridan, S.C. (2018) The characteristics of extreme cold events and cold air outbreaks in the eastern United States. *International Journal of Climatology*, 38(S1), e807–e820. Available from: <https://doi.org/10.1002/joc.5408>
- Smith, E.T. & Sheridan, S.C. (2020) Where do cold air outbreaks occur, and how have they changed over time? *Geophysical Research Letters*, 47(13), 1–11. Available from: <https://doi.org/10.1029/2020GL086983>
- Thompson, D.W.J. & Wallace, J.M. (2000) Annular modes in the extratropical circulation. part I: month-to-month variability. *Journal of Climate*, 13(5), 1000–1016. Available from: [https://doi.org/10.1175/1520-0442\(2000\)013<1000:AMITEC>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2)
- Tripathi, O.P., Charlton-Perez, A., Sigmond, M. & Vitart, F. (2015) Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environmental Research Letters*, 10(10), 1–8. Available from: <https://doi.org/10.1088/1748-9326/10/10/104007>
- Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C. et al. (2017) The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98(1), 163–173. Available from: <https://doi.org/10.1175/BAMS-D-16-0017.1>
- Wang, L. & Chen, W. (2010) Downward arctic oscillation signal associated with moderate weak stratospheric polar vortex and the cold December 2009. *Geophysical Research Letters*, 37(9), 1–5. Available from: <https://doi.org/10.1029/2010GL042659>
- Wheeler, D.D., Harvey, V.L., Atkinson, D.E., Collins, R.L. & Mills, M.J. (2011) A climatology of cold air outbreaks over North America: WACCM and ERA-40 comparison and analysis. *Geophysical Research Letters*, 47(D12), 1–8. Available from: <https://doi.org/10.1029/2020GL086983>
- White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J.T., Lazo, J.K., Kumar, A. et al. (2017) Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications*, 24(3), 315–325. Available from: <https://doi.org/10.1002/met.1654>
- White, I., Garfinkel, C.I., Gerber, E.P., Jucker, M., Aquila, V. & Oman, L.D. (2019) The downward influence of sudden stratospheric warmings: association with tropospheric precursors. *Journal of Climate*, 32(1), 85–108. Available from: <https://doi.org/10.1175/JCLI-D-18-0053.1>

How to cite this article: Millin, O.T. & Furtado, J.C. (2026) Subseasonal predictability of cold spells in the central United States. *Quarterly Journal of the Royal Meteorological Society*, 152:e70081. Available from: <https://doi.org/10.1002/qj.70081>