aerospace

MDPI

Article

Prediction of 30-h Runway Directions at Jeju International Airport

Jawoon Cho [1], Dohyun Kim [1], Hyehyun Kim [1], Zin Min Paing [2] and Hojong Baik [2,*]

[1] Department of Advanced Air Transportation, Korea Aerospace University, Goyang-si 10540, Gyeonggi-do, Republic of Korea; newhuman20@gmail.com (J.C.); dhkim.7130@gmail.com (D.K.); kmky26200@gmail.com (H.K.)
[2] School of Air Transportation and Logistics, Korea Aerospace University, Goyang-si 10540, Gyeonggi-do, Republic of Korea; zinmin75@kau.kr
* Correspondence: hbaik@kau.ac.kr

Abstract

In Air Traffic Flow Management (ATFM), anticipating runway directions is essential for accurately estimating airport capacity. With this information, operators can proactively manage taxi flows and mitigate reconfiguration disruptions. Motivated by these practical needs, this paper introduces a 30-h runway direction forecasting framework for Jeju International Airport (RWY 07/25) using multi-source data including Terminal Aerodrome Forecasts (TAFs), aerodrome weather reports (METAR/SPECI), and Airport Collaborative Decision Making (A-CDM) records (2022–2024). The framework estimates hourly direction probabilities using a regularized logistic regression, and applies the Viterbi algorithm to decode the sequence, removing most short-lived direction reversals. In an external 2024 evaluation, the framework demonstrated an accuracy of 89.16% (T+1–T+6) and 85.00% (T+1–T+30), exceeding the baseline TAF wind rule by 5.73 and 4.09 percentage points. Although correctly timed predicted transitions fell from 94 to 83, Viterbi decoding reduced predicted direction changes by 27.7% and false alarms from 505 to 355 while leaving hourly accuracy unchanged. Ultimately, this 30-h outlook could serve, after prospective validation, as a decision-support input for capacity planning, traffic coordination, and surface movement management at Jeju International Airport.

Keywords: runway direction prediction; runway configuration prediction; single-runway airport; logistic regression; Viterbi decoding; transition timing; TAF; METAR; air traffic flow management

Check for updates

Academic Editor: Milad Memarzadeh

Received: 14 October 2026
Revised: 16 November 2026
Accepted: 19 November 2026
Published: 23 November 2026

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

1. Introduction

1.1. Research Background and Motivation

Air Traffic Flow Management (ATFM) aims to balance air traffic demand against available airport and airspace capacity to prevent unnecessary congestion and delays. Airport capacity depends on weather, runway configuration, the arrival–departure ratio, and fleet mix [1]. Specifically, the runway operating direction establishes the arrival and departure routes and surface routing flows, serving as a primary determinant of hourly airport throughput. Anticipating runway operating directions is therefore essential for robust capacity planning and advance operational management.

In practice, runway configurations change dynamically, and the transition process itself introduces substantial operational burdens. Runway capacity typically degrades during its configuration transitions, with reciprocal direction reversals causing more severe

throughput disruptions than routine operations. Such transitions also intensify controller and pilot workloads through required traffic reorganization, revised coordination with adjacent sectors, and flight system reprogramming.

These operational disruptions are particularly pronounced at Jeju International Airport. Due to its unique island geography, the airport is routinely subjected to volatile wind direction and speed regimes that directly dictate runway orientation [2]. While Jeju International Airport operates two intersecting runways, flight movements are almost exclusively concentrated on its primary runway, RWY 07/25. As a result, the auxiliary runway (RWY 13/31) cannot absorb traffic during reciprocal direction changes, leaving the airport fully vulnerable to configuration transition bottlenecks. These idiosyncratic operational constraints establish Jeju International Airport as a uniquely compelling and challenging operational benchmark for long-horizon runway operating direction prediction.

1.2. Previous Studies on Runway Configuration and Research Gap

Previous studies have mainly predicted or optimized runway configurations at multi-runway airports, where a configuration combines the active runways, their operating directions, and their arrival–departure roles. For airports operating a single runway or relying heavily on a primary runway, the operating direction remains the sole configuration variable to be determined. Jeju International Airport exemplifies this scenario. In a related study, Lee [3] predicted the hourly direction of Jeju's RWY 07/25 using random forest and LightGBM classifiers based on local numerical weather forecasts and the Terminal Aerodrome Forecast (TAF) as the initial step for an arrival-throughput model. Building upon this foundation, this study advances that methodology by drawing on the broader runway configuration research reviewed below.

Research on runway configuration divides into prescriptive and predictive approaches [4]. Prescriptive studies search for the configuration, or sequence of configurations, that gives the best operational performance and recommend it. Li and Clarke planned the sequence of configurations that maximizes capacity under uncertain wind [5], and Lam et al. used reinforcement learning to set a day's configuration schedule that maximizes capacity at Philadelphia International Airport [6]. These studies show which configuration would be preferable under given conditions. The present study takes the predictive approach rather than this prescriptive one. Robust capacity planning and advance operational management require an estimate of the direction that will actually be in use, not of the direction that would be preferable [4].

Predictive studies learn runway configuration choices from past operations. For instance, Ramanujam and Balakrishnan [7] revealed that both headwind and inertia—the tendency to maintain the current configuration—significantly influence runway selection at LaGuardia and Newark airports. Avery and Balakrishnan [8] predicted runway configurations three hours ahead at San Francisco, LaGuardia, and Newark airports with 81.2%, 81.3%, and 77.8% accuracy. In another study, Puranik et al. [4] predicted configurations 15 min ahead using a machine-learning model, reaching 93% accuracy at Charlotte Airport (which operates only two configurations, similar to a binary direction choice) and 73% at Denver Airport (which operates eleven). These studies highlight the strong persistence of runway states and the critical role of wind; however, they predict at most a few hours ahead using current observations or short-range forecasts.

Looking further ahead requires longer-range forecasts as input. De Leege and Janssen showed that, with weather forecasts as input, runway use can be forecast up to 30 h ahead, giving a probability for each runway combination at Amsterdam Airport Schiphol [9]. Such forecasts, however, become less accurate the further ahead they look: with machine-learning models trained on TAFs at five US airports, Raju et al. found that average accuracy fell

from 72% three hours ahead to about 65% at 14 and 24 h [10]. Whether observations or forecasts are the better input also depends on how far ahead the prediction looks: at eleven European airports, the latest routine aerodrome weather report (METAR) predicted runway allocation better than the TAF up to about four hours ahead, and the TAF was better beyond that [11]. Predicting the whole 30-h horizon therefore calls for combining the forecast with recent observations and runway use, using only what is available at the time of prediction.

A forecast of this kind is also shaped by the temporal characteristics of its inputs, which are distinct from the forecast lead time: how often observations and forecasts are issued, how quickly they are delivered, and over what interval operations are aggregated. Rezo and Steiner showed that the sampling and update frequency of meteorological data affects the information available to aviation operations [12], and verification studies of TAFs emphasize that a forecast describes time windows and conditional ranges rather than a single value at a single time [13,14]. At Madeira International Airport, a deep-learning study used high-frequency wind data to predict the operational status of the runway at one-minute resolution over a 20-min horizon [15]; that setting is tactical and relies on data much finer than the hourly reports and labels available for the 30-h horizon considered here. For the present data, the reporting intervals and the availability delay are described in Section 2.2, and their effect on the results is examined in Appendix B.

Beyond the choice of inputs, a 30-h prediction is a sequence of hourly directions. If each hour is decided separately, the predicted direction can flip back and forth whenever the probability is close to 50%. Rebollo et al. sought to stabilize predicted runway configuration sequences by feeding each step's prediction into the next [16], Dalmau and Herrema predicted six-hour runway-use sequences at Amsterdam Airport Schiphol with an encoder–decoder network, using observed future weather and demand in their evaluation [17], and Wang et al. found that producing a six-hour runway configuration sequence in one pass was more accurate than producing it step by step [18]. Following these studies, the present study develops the earlier work at Jeju [3] in two ways: it adds recent observations and runway use to the forecast inputs, and it selects the whole 30-h sequence at once while accounting for how strongly the runway direction persists from hour to hour.

The transition itself has also been studied: Lau et al. estimated how long a configuration change takes to complete at Philadelphia International Airport [19], a duration target that differs from detecting the hour in which the direction changes. Recent work has further examined the bias and the robustness to noisy inputs of an assistance tool for runway configuration management [20], and Wang and Zhang predicted the runway configurations and airport acceptance rates of three airports in the New York area jointly from gridded weather forecasts [21]. These settings, which use observed weather or gridded weather forecasts, differ from one in which only the forecast issued for the airport, the TAF, is available over its full validity period. Sequence prediction is therefore not itself the contribution of the present study; its contribution lies in the specific setting—the hourly binary direction of a dominant runway over 30 h, predicted from the TAFs actually issued—and in evaluating hourly accuracy and transition timing separately.

1.3. Research Objective and Contributions

Building on these studies, this paper develops a 30-h runway operating direction prediction framework for Jeju International Airport, where flight movements are heavily concentrated on a single primary runway. Four times daily, at fixed forecast origins, the framework predicts the primary runway's operating direction for each of the subsequent 30 h. These predictions rely exclusively on the forecasts, observations, and runway-operation records available at that specific time.

A regularized logistic regression, which is well-suited for binary choice problems and offers high interpretability, estimates the direction probability at each target hour. Subsequently, Viterbi decoding utilizes these probabilities to determine the optimal 30-h sequence as a whole, accounting for the strong persistence of runway directions rather than evaluating each hour in isolation. For an out-of-sample evaluation, the framework was trained on 2022–2023 data and tested on chronologically separated 2024 data, with all settings selected within the 2022–2023 data.

The prediction results show that the framework predicts the runway operating direction more accurately than a baseline rule derived solely from TAF wind direction, and that it does so across the full 30-h horizon rather than only in its first hours. The accuracy is concentrated in hours in which the direction persists, whereas the hour of a direction change remains difficult to predict for every method examined. Although some correctly timed predicted transitions are lost, Viterbi decoding reduces the number of predicted direction changes and false alarms while maintaining accuracy. After training on each airport's own data, the framework may extend to other airports operating a single runway or concentrating operations on one primary runway.

2. Problem Definition, Data and Methodology

2.1. Study Scope and Prediction Task

This study formulates the prediction of the operating direction of the primary runway at Jeju International Airport as a sequential binary classification problem. Although the airport has two intersecting runways, almost all traffic uses the primary runway, RWY 07/25, and RWY 13/31 is used only rarely. The scope is therefore restricted to RWY 07/25, and its hourly operating direction is modeled as a binary state, $y_t \in \{0, 1\}$, where $y_t = 0$ corresponds to RWY 07 and $y_t = 1$ to RWY 25. At each prediction time T (hereafter the forecast origin), the framework produces a sequence of hourly directions over the next 30 h, from the first complete hour after T (i.e., T+1) to T+30.

At each forecast origin T, the framework uses three data sources available at that time: the latest Terminal Aerodrome Forecast (TAF), surface observations reported up to T in routine and special aerodrome weather reports (METAR/SPECI), and records of recent runway use from Airport Collaborative Decision Making (A-CDM).

To solve this sequential classification problem, the proposed framework employs a two-stage methodology as illustrated in Figure 1. Following the extraction of input features, a logistic regression estimates the hourly direction probabilities. Subsequently, Viterbi decoding is applied to select the optimal 30-h sequence. The detailed formulations of these predictive procedures are provided in Section 2.4.

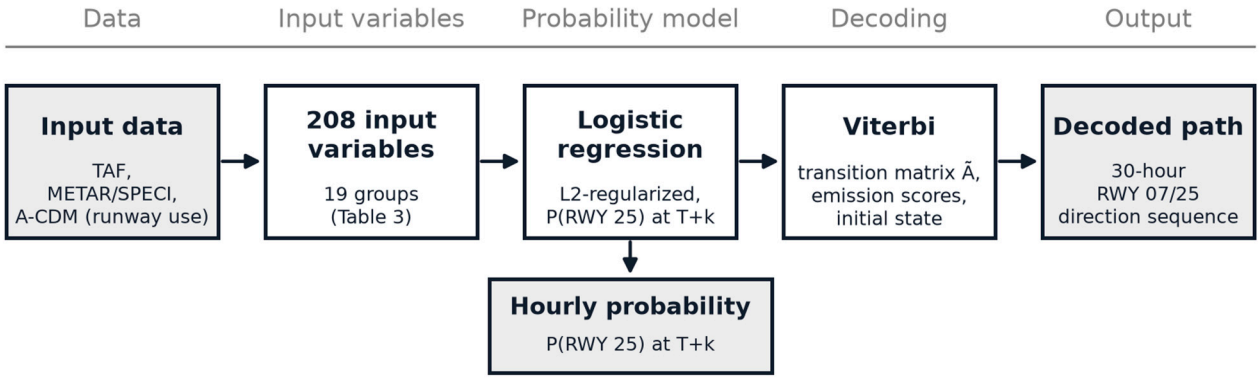

Figure 1. Architecture of the runway-direction prediction framework. Arrows show the flow from the input data to the decoded path; shaded boxes are the inputs and outputs.

2.2. Data Sources and Temporal Alignment

The data cover 1 January 2022 to 31 December 2024 and combine TAFs, METAR/SPECI observations, and A-CDM records (Table 1). The model was trained and validated on the 2022–2023 data, and the whole of 2024 was kept for evaluation.

Table 1. Data sources and variables used in the analysis (1 January 2022–31 December 2024).

Data Source	Variables	Notes
TAF	Issue time; validity period; wind direction; wind speed; gust; visibility; weather	Includes amended (AMD) and corrected (COR) forecasts
METAR and SPECI reports	Observation time; wind direction; wind speed; gust; visibility; temperature; dew point; QNH; weather	Includes corrected (COR) reports; only reports available by the forecast origin were used
A-CDM	Recorded operation time; runway direction (RWY 07 or RWY 25)	Hourly predominant direction for Jeju International Airport; hours without a confirmed direction were excluded

Each TAF was converted into hourly values over its validity period using the change groups defined in ICAO Annex 3 [22] (Table 2). Specifically, an FM group replaces the prevailing forecast from its indicated start time, while a BECMG group is applied from the midpoint of its change period. Because TEMPO and PROB groups describe conditions that may only temporarily occur, they were retained as separate conditional inputs rather than overwriting the prevailing forecast. Amended (AMD) and corrected (COR) forecasts were integrated as soon as they were issued and identified with distinct flags. To reflect real-world operational constraints, all meteorological reports were assumed to become available ten minutes after their issuance or observation time, strictly limiting the model to data accessible at the forecast origin.

Table 2. Processing of TAF change indicators.

Indicator	Processing	Notes
FM	Applied from the stated start time	The new prevailing conditions replace the preceding prevailing state.
BECMG	Applied from the midpoint of the change interval	A deterministic hourly conversion was used for model input.
TEMPO, PROB30, PROB40	Stored as separate conditional variables	Conditional values did not overwrite prevailing values; indicator type was retained.

The three sources differ in reporting interval. Routine METARs were issued every 30 min in 2022 and every hour in 2023–2024, with on average 1.1–1.5 SPECI reports per day; routine TAFs were issued four times a day (05, 11, 17, and 23 UTC), with 253–314 amended TAFs per year. The observation inputs use the latest report available at the forecast origin and at each of the six preceding hourly cut-offs, so that the routine reports used are those issued on the hour in all three years. The forecast origins fall ten minutes after the routine TAF issue times; the assumed ten-minute delay is therefore the longest delay at which the newest TAF is still usable at the origin (Appendix B).

A-CDM movement records were aggregated into hourly intervals, with each hour labeled according to its predominant operating direction. Hours lacking confirmed movements, such as during airport curfews, were excluded, yielding a final dataset of 18,660 la-

beled hours. These A-CDM records serve a dual purpose: data prior to the forecast origin provide recent-runway inputs, while future target hours supply the ground-truth labels. Forecast origins were established at 02:10, 08:10, 14:10, and 20:10 Korea Standard Time (KST), aligning with Jeju's six-hourly routine TAF issuance cycle. The first complete hour following the origin is denoted as T+1 (e.g., 09:00 for the 08:10 origin), with each forecast run predicting the sequence from T+1 to T+30. Because the 30-h horizons overlap across different origins, identical target hours appear in multiple runs. To prevent statistical bias, each row was weighted inversely to the frequency of its target hour's appearance, ensuring equal representation for all target hours during model training.

The hourly label follows the majority of the arrivals and departures recorded on RWY 07/25 within the clock hour. In 2024, 44.0% of the 6689 labeled hours contained movements in both directions, with a median minority share of 6.9%; 21 h (0.3%) had equal counts and took the direction that the counting routine listed first; and movements on RWY 13/31 (1.4% of all movements) were not counted, which left two hours that had movements only on RWY 13/31 without a label. A label therefore identifies the predominant direction of an hour and not the minute at which the direction changed: 80% of the hours in which the label changed contained movements in both directions. Appendix A gives the remaining processing rules.

2.3. Exploratory Analysis and Feature Construction

To construct appropriate input variables, the historical data were statistically analyzed to identify key operational patterns.

Overall, RWY 07 was utilized in 56.10% of the labeled hours, while RWY 25 accounted for 43.90% (Figure 2a). Runway operations exhibited strong persistence: among 17,561 consecutive hourly pairs, 95.39% maintained the same direction (Figure 2b). Furthermore, at the latest forecast origin before each target hour, the direction of the last completed hour $(H-1)$ matched the target-hour direction in 86.54% of cases, and that of the sixth completed hour $(H-6)$ in 76.54%. Consequently, the operating directions of the six completed hours immediately preceding the forecast origin were incorporated as inputs. To capture its diminishing influence over longer forecasting horizons, the most recent direction $(H-1)$ was interacted with the five lead-time bands. Additionally, because the usage share of RWY 25 varied significantly by time of day (38.66–49.29% over 06–23 KST) and by month (14.69–65.62%; Figure 2c,d), the target hour and month were cyclically encoded using sine and cosine transformations to preserve their periodic nature:

$$\sin\left(\frac{2\pi \cdot \text{hour}}{24}\right),\ \cos\left(\frac{2\pi \cdot \text{hour}}{24}\right),\ \sin\left(\frac{2\pi \cdot \text{month}}{12}\right),\ \cos\left(\frac{2\pi \cdot \text{month}}{12}\right)$$

Wind is the primary meteorological determinant in runway selection, as aircraft take off and land into the wind. A simple baseline rule—assigning RWY 25 when the target-hour TAF wind direction lies between 160° and 340°—correctly matched the actual operating direction in 79.60% of the 93,163 forecast-origin–target-hour pairs (Figure 3a,b). However, this rule exhibited significant vulnerabilities near its boundaries and under light wind conditions. Specifically, for the latest forecast available for each target hour (18,649 h; Figure 3c–e), accuracy dropped to 48.04% when the forecast direction was within 10° of the 160° or 340° boundaries (compared to 92.16% at 60° or more), and fell to 70.31% under light winds of 5–7 kt (compared to 93.87% at ≥14 kt) (Figure 3c,d).

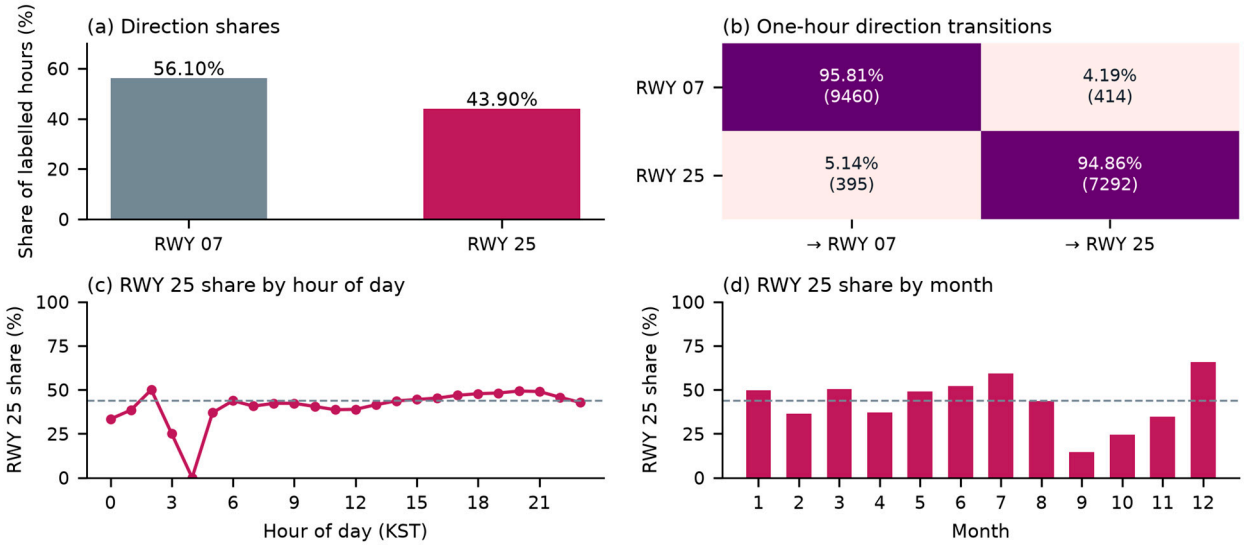

Figure 2. Runway-operating characteristics in the labeled data, 2022–2024: (**a**) share of labeled hours by direction; (**b**) one-hour direction transitions; (**c**) RWY 25 share by hour of day (KST); (**d**) RWY 25 share by month. Dashed lines mark the overall RWY 25 share. In (**b**), each row gives the direction in one hour and each column (→) the direction in the next hour; darker cells indicate larger shares.

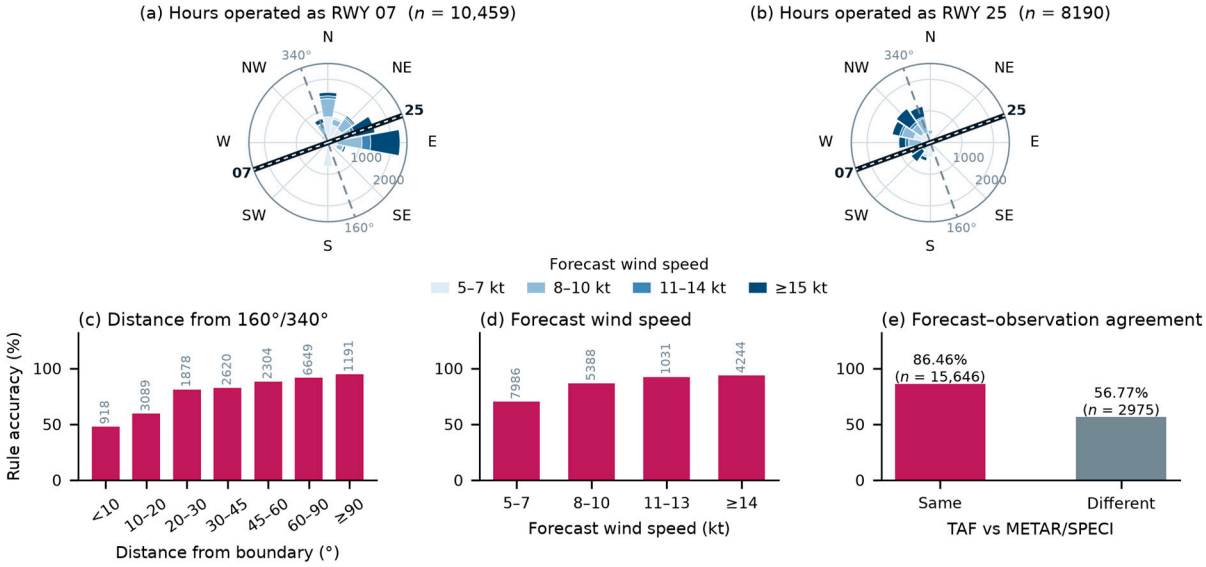

Figure 3. The TAF wind-direction rule, 2022–2024: (**a**,**b**) target-hour TAF wind direction and speed in hours operated as RWY 07 and RWY 25, with the rule boundaries at 160° and 340° (dashed lines) and the runway axis used in this study (solid black line); (**c**–**e**) rule accuracy by distance from the boundary, by forecast wind speed, and by agreement between the TAF and the latest METAR/SPECI. Numbers above the bars are numbers of target hours.

To address these limitations, the wind inputs were expanded beyond raw direction. They incorporate the boundary distance $D(\theta)$—defined as the angle between the forecast wind direction θ and the nearer boundary of 160° or 340° (0° on the boundary and 90° along the runway axis)—as well as the along-runway and cross-runway wind components, wind speed, and gusts. Furthermore, the broader forecast context proved influential: the actual share of RWY 25 operations was only 7.56% when no more than 20% of the surrounding forecast hours favored RWY 25, but surged to 80.65% when over 80% did. Consequently, wind values from the same TAF at offsets ranging from −6 to +6 h around

the target hour were included, while conditional TEMPO and PROB groups were retained as separate inputs.

Recent observations help mitigate forecast uncertainty. When the latest TAF and preceding METAR indicated the same direction under the wind rule, the rule achieved 86.46% accuracy over 15,646 target hours, compared to just 56.77% in the 2975 h when they disagreed (Figure 3e). Observed wind, gusts, visibility, temperature, dew point, and QNH from the latest METAR/SPECI and the six preceding hours were included, alongside the differences between the forecast and observed winds. In total, 208 input variables spanning 19 groups were utilized (Table 3). Skewed variables were log-transformed, and missing values were replaced by the training-period median, with dedicated binary flags recording the availability of each value.

Table 3. Candidate input variables of the logistic regression (208 total).

Variable Group	Variable Subgroup	No.	Variables Included
TAF prevailing forecast	Wind direction and runway-axis wind	15	Wind direction: $\sin(\theta)$, $\cos(\theta)$, $\sin(2\theta)$, $\cos(2\theta)$, and availability (5); RWY 07/25 boundary-distance value–availability pair (2); 070°-axis along-wind, crosswind, and absolute-crosswind value–availability pairs (6); RWY 25-rule and variable-wind flags (2)
TAF prevailing forecast	Wind speed, gust, and visibility	8	Raw wind speed, log-transformed wind speed, gust, and log-transformed visibility as value–availability pairs (4 × 2)
TAF prevailing forecast	TAF issue and change type	6	Regular, amended, and corrected TAF flags (3); active TEMPO, PROB40, and PROB30 condition flags (3)
TAF prevailing forecast	Weather phenomena	5	Rain, snow, thunderstorm, fog/mist, and dust/haze flags
TAF conditional forecast	Conditional wind direction	3	TEMPO/PROB wind-direction $\sin(\theta)$, $\cos(\theta)$, and availability
TAF conditional forecast	Conditional speed, gust, and visibility	6	TEMPO/PROB wind speed, gust, and log-transformed visibility as value–availability pairs (3 × 2)
TAF conditional forecast	Conditional weather phenomena	5	TEMPO/PROB rain, snow, thunderstorm, fog/mist, and dust/haze flags
TAF context around target time	Wind by target-time offset	48	At −6, −4, −2, −1, +1, +2, +4, and +6 h from target: direction (sin, cos, availability), speed (value, availability), and same-bulletin direction-and-speed availability (8 × 6)
TAF context around target time	Directional trend across offsets	5	Across target and offset times: RWY 25-rule fraction and availability (2); RWY 07/25 rule-switch count and availability (2); available-direction fraction (1)
METAR/SPECI observations	Latest and previous observed wind	49	At the latest and H − 1 through H − 6 cutoffs: direction (sin, cos, availability), speed (value, availability), and RWY 07/25 boundary distance (value, availability) (7 × 7)
METAR/SPECI observations	Latest observed weather values	12	Gust, log-transformed visibility, temperature, dew point, QNH, and observation age (minutes) as value–availability pairs (6 × 2)
METAR/SPECI observations	TAF–observation differences	6	Circular direction gap, TAF-minus-observed wind speed, and RWY 25-rule agreement as value–availability pairs (3 × 2)
METAR/SPECI observations	Observation report type	5	Separate flags for METAR, SPECI, COR, METAR_COR, and SPECI_COR reports
METAR/SPECI observations	Observed weather phenomena	5	Rain, snow, thunderstorm, fog/mist, and dust/haze flags
Recent runway operations	Runway use in previous 1–6 h	12	H − 1 through H − 6 modal runway directions and corresponding availability flags (6 × 2)

Table 3. *Cont.*

Variable Group	Variable Subgroup	No.	Variables Included
Recent runway operations	Previous runway × lead-time band	5	Interactions of H − 1 runway direction with T+1–2, T+3–6, T+7–12, T+13–24, and T+25–30 lead-time bands
Forecast horizon and time	Continuous lead-time measures	4	Actual prediction-to-target lead; nominal scheduled-cycle-to-target lead; and square-root-transformed actual lead with availability (1 + 1 + 2)
Forecast horizon and time	Lead-time bands	5	Flags for nominal T+1–2, T+3–6, T+7–12, T+13–24, and T+25–30 lead-time bands
Forecast horizon and time	Target hour and month	4	Cyclic sine/cosine encodings of target hour (2) and target month (2)
Total	19 variable subgroups	208	Total number of candidate input variables

Note: TAF = terminal aerodrome forecast; METAR/SPECI = routine/special aerodrome observations; θ = wind direction; H − k = kth completed hour before prediction; T+a–b = nominal lead-time band; COR = corrected report. Availability is a separate binary data-presence flag. The study-specific RWY 25 rule is 160° ≤ wind direction ≤ 340°.

The statistics in this section describe all three years; recomputed on the 2022–2023 development period alone, they were essentially the same (RWY 07 in 56.06% of labeled hours; the same direction in 95.26% of consecutive hourly pairs).

2.4. Prediction Methodology

The input variables constructed in Section 2.3 are fed into a two-stage predictive framework (Figure 1). In the first stage, a logistic regression (Section 2.4.1) estimates the probability of RWY 25 usage at each target hour. Then, Viterbi decoding utilizes these probabilities to determine the optimal 30-h direction sequence (Section 2.4.2).

2.4.1. Logistic Regression

Logistic regression is employed as the probability model due to its suitability for binary discrete-choice problems and its ability to output direct probabilities. Such models have been extensively utilized in prior runway configuration research [7–9]. The probability of operating RWY 25 at target hour $T + k (k = 1, \ldots, 30)$ is defined in Equation (1):

$$p_{\text{Logistic},T+k} = \sigma\left(\beta_0 + \sum_{j=1}^{208} \beta_j x_{j,T+k}\right) = \left[1 + \exp\left(-\left(\beta_0 + \sum_{j=1}^{208} \beta_j x_{j,T+k}\right)\right)\right]^{-1} \quad (1)$$

Here, $x_{j,T+k}$ denotes the j-th input variable (Table 3), β_j is its corresponding coefficient, β_0 is the intercept, and σ represents the logistic function bounded between 0 and 1. The model retains high interpretability, as the sign of β_j explicitly indicates whether an input shifts the probability toward RWY 25 or RWY 07.

Because many of the 208 inputs describe the same wind in different forms and are strongly correlated, unpenalized coefficients are prone to instability and overfitting. To address this, the coefficients were estimated using $L2$ regularization by minimizing the objective function in Equation (2):

$$\mathcal{L}(\beta, \beta_0) = -\frac{\sum_{i=1}^{N} w_i [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]}{\sum_{i=1}^{N} w_i} + \lambda \sum_{j=1}^{208} \beta_j^2 \quad (2)$$

The objective function in Equation (2) comprises two components. The first term calculates the prediction error via weighted log-loss, penalizing predictions that deviate from the actual operating direction. Here, N is the total number of training instances, $y_i \in \{0, 1\}$ represents the actual direction (1 for RWY 25, 0 for RWY 07), p_i is the predicted

probability of RWY 25, and w_i is the row weight defined in Section 2.2. The second term imposes an $L2$ penalty on the magnitude of the coefficients, scaled by the regularization parameter λ.

Prior to model fitting, all continuous input variables were standardized to a zero mean and unit variance using the training dataset. This standardization ensures that the $L2$ penalty is applied uniformly across all inputs, preventing variables with naturally larger scales from dominating the penalty. The regularization strength was optimized via time-ordered validation within the 2022–2023 development set (Section 2.5.1). Among the four candidate values evaluated ($C \in \{0.01, 0.1, 1, 10\}$; C is the inverse regularization strength of scikit-learn, which corresponds to $\lambda = 1/(2CN)$ in Equation (2) because the row weights average one), the strongest penalty ($C = 0.01$) yielded the optimal predictive performance.

2.4.2. Viterbi Decoding of the Final Direction Sequence

Evaluating each target hour independently (e.g., assigning RWY 25 whenever $p_{\text{Logistic}} > 0.5$) can lead to excessive and unrealistic switching of the predicted direction when the probability fluctuates around the 0.5 threshold. This contradicts the strong operational inertia demonstrated in Section 2.3, where the actual runway direction remains unchanged 95% of the time. To address this, Viterbi decoding [23] is applied to determine the optimal 30-h sequence as a whole, effectively balancing the hourly predicted probabilities against this operational persistence. The algorithm integrates three components: state transition probabilities (the likelihood of keeping or changing direction), emission scores (the model's hourly probability for each direction), and the initial runway state.

The transition matrix A, where each element A_{ij} represents the probability of transitioning from direction i to direction j in the subsequent hour (ordered as RWY 07, RWY 25), was empirically derived from consecutive hourly labels in the 2022–2023 training data. To ensure no transition receives a zero probability, additive smoothing (adding a constant of 5 to each transition count) was applied, yielding Equation (3):

$$\mathbf{A} = \begin{bmatrix} 0.9581 & 0.0419 \\ 0.0566 & 0.9434 \end{bmatrix} \tag{3}$$

Because these empirical persistence probabilities are exceptionally high, they could overly dominate the hourly evidence, preventing the decoded sequence from switching directions even when the meteorological predictions are strong. Consequently, the switching probability, $q_i = A_{i,1-i}$, was relaxed toward 0.5 using a dispersion coefficient γ. Here, $\gamma = 0$ completely eliminates the persistence effect (treating hours independently), while $\gamma = 1$ retains the original matrix A:

$$\tilde{q}_i = \sigma(\gamma \cdot \text{logit}(q_i)) = \left[1 + \exp\left(-\gamma \ln\left(\frac{q_i}{1 - q_i}\right)\right)\right]^{-1}$$

Calibrating with an optimal dispersion coefficient of $\gamma = 0.1$ yielded the regularized transition matrix $\widetilde{\mathbf{A}}$ in Equation (4):

$$\widetilde{\mathbf{A}} = \begin{bmatrix} 0.5776 & 0.4224 \\ 0.4301 & 0.5699 \end{bmatrix} \tag{4}$$

At each target hour, the logistic-regression probability is converted into emission scores for each direction, defined in Equations (5a) and (5b):

$$e_{T+k}(1) = \eta\left[ln p_{\text{Logistic},T+k} - ln\pi\right] \tag{5a}$$

$$e_{T+k}(0) = \eta\left[ln\left(1 - p_{\text{Logistic},T+k}\right) - ln(1 - \pi)\right] \tag{5b}$$

where π represents the baseline prevalence of RWY 25 in the training data, and η serves as a scaling factor that controls the weight of the hourly evidence relative to the transition probabilities. A predicted probability equal to the baseline π indicates neutrality, whereas any probability exceeding π explicitly favors RWY 25. The decoding sequence is initialized using the last observed direction $(H-1)$ with a confidence probability ρ; if this prior direction is missing, it defaults to the baseline share π. At each subsequent time step, the algorithm recursively accumulates the emission score and the logarithm of the transition probability. Based on the regularized matrix in Equation (4), changing the runway direction incurs a logarithmic penalty of approximately 0.3 compared to maintaining it. Consequently, a direction switch is triggered only when the hourly meteorological evidence strictly outweighs this transition cost. The optimal sequence is computed using the dynamic programming recursion in Equation (6):

$$\delta_k(j) = e_{T+k}(j) + \underset{i\in\{0,1\}}{max}\left\{\delta_{k-1}(i) + ln\left[\left(\widetilde{A}^{\Delta_k}\right)_{ij}\right]\right\} \quad (6)$$

where $\delta_k(j)$ denotes the maximum accumulated score for a sequence terminating in direction j at the k-th target hour, and $\widetilde{A}^{\Delta k}$ accounts for a temporal gap of Δk hours (e.g., the interval between $H-1$ and T+1). By backtracking from the optimal final state, $\text{argmax}_j\delta_{30}(j)$, the full 30-h sequence—hereafter referred to as the decoded path—is extracted. Each forecast run is decoded independently.

The optimal hyperparameters ($\gamma = 0.1$, $\eta = 2.0$, and $\rho = 0.70$) were determined via a grid search coupled with time-ordered validation on the 2022–2023 dataset ($\gamma \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0, 1.5\}$, $\eta \in \{0.5, 1.0, 1.5, 2.0, 3.0, 4.0\}$, $\rho \in \{0.70, 0.80, 0.90, 0.95, 0.98, 0.995\}$). The setting with the highest ±1-h transition-timing F1 score on the validation blocks was selected (Section 2.5.1), with ties resolved by the higher path accuracy.

Two further details complete the specification. Because the emission scores compare the probability with the baseline share π (0.438 in the 2022–2023 data) rather than with 0.5, the decoded direction is RWY 25 whenever the probability exceeds π, even without any persistence. Each forecast run is decoded over all 30 target hours, including hours without movements, for which the model still produces a probability, and probabilities are clipped to the interval from 10^{-6} to $1-10^{-6}$ before logarithms are taken.

2.5. Model Development and Evaluation

2.5.1. Time-Ordered Development and External Evaluation

The models were trained and validated exclusively on the 2022–2023 dataset and externally evaluated on the chronologically separated 2024 dataset, simulating real-world operational forecasting [24]. Preserving temporal order is critical in this context: prior research demonstrated that when runway-configuration models at five US airports were evaluated using a forward-rolling temporal validation rather than random sampling, the accuracy of tree-based models dropped significantly from 79–83% to approximately 65% [10]. To prevent any data leakage, all model parameters and hyperparameters were estimated solely from the 2022–2023 data. Following hyperparameter tuning, the model was refitted on the entire 2022–2023 dataset before generating predictions for 2024, ensuring that each forecast run utilized only the historical information available exactly at its specific origin.

For temporal tuning—such as selecting the logistic regression penalty strength C—a time-series cross-validation approach was employed. The forecast runs of the 2022–2023 data were ordered in time; the first third was used only for training, and the remainder was partitioned into four consecutive blocks, validating on each block in turn while restricting the training set to labels strictly preceding the validation block. The Viterbi settings were selected on the last three blocks. Strict programmatic checks were implemented during fitting to

guarantee that no future labels—especially from the 2024 test set—were inadvertently exposed. Consequently, the regularization strength and the Viterbi decoding configuration were fixed before the 2024 external evaluation was scored.

Together with the regularization of the model (Section 2.4.1), this separation limits overfitting to the 208 correlated inputs: no setting was tuned on the evaluation data, and all reported results come from years the models had not seen (2024, and 2023 after training on 2022). The 2024 data had, however, been examined in the originally submitted version of this work, in which a second probability model was chosen with reference to the evaluation years; that model has been removed (Section 4.3), and the present results were obtained with settings selected within the development period only.

2.5.2. Comparison Methods and Performance Measures

Predictive accuracy was evaluated across various forecast horizons (T+1–T+6, T+1–T+30) and five specific lead-time bands (T+1–2, T+3–6, T+7–12, T+13–24, and T+25–30). The proposed framework was benchmarked against four references: (i) the TAF wind-direction rule, (ii) a persistence baseline (maintaining the latest observed direction), (iii) a persistence–hysteresis wind rule, which starts from the latest observed direction and changes direction only when the TAF rule indicates the opposite direction with the forecast wind direction at least δ from the rule boundary and the forecast wind speed at least v ($\delta = 20°$ and $v = 0$ kt, chosen on the development years), and (iv) the hour-by-hour decisions of the logistic regression (RWY 25 when the probability is at least 0.5), which isolate the effect of Viterbi decoding. To ensure strict comparability across methods with varying data availability, accuracy was also computed on a common subset of intersecting rows. An ablation study was also conducted to quantify the contribution of each data source; the complete pipeline, including the selection of the regularization strength and of the Viterbi settings, was repeated using TAF inputs only (114 variables), TAF with METAR/SPECI (191), TAF with recent runway use (131), and the complete feature set (208).

The quality of the predicted probabilities was assessed using the Brier score, log loss, and the expected calibration error (ECE) [25]. The ECE was computed using both ten equal-width bins and ten equal-mass bins (containing equal numbers of predictions); lower values indicate superior calibration across all three metrics. Operational sequence stability was measured by the frequency of direction changes between adjacent hours within each forecast run. The precise timing of these changes was evaluated by matching predicted and actual switches one-to-one, calculating precision (the proportion of predicted changes that actually occurred), recall (the proportion of actual changes successfully predicted), and the F1 score, both at the exact target hour and within a ± 1 h tolerance. Because the 30-h runs overlap, each target hour at which the path of any run changes direction was counted as one predicted transition, issued at the earliest run showing it, and a predicted transition could be matched only to an actual change that occurred after it was issued; the 275 actual changes of 2024 are the distinct hours at which the label changed, and they appear as 1312 changes among the 30,429 adjacent labeled pairs within runs. Accuracy was additionally reported separately for transition hours, in which the direction differs from that of the preceding labeled hour, and for the remaining stable hours.

The statistical significance of paired differences was evaluated using McNemar's test on discordant pairs where two methods disagreed [26]. To account for strong temporal autocorrelation, 95% confidence intervals were constructed using a block bootstrap approach, resampling entire calendar weeks 2000 times [27]. Finally, as a second evaluation, the entire experimental procedure was replicated using 2022 data for training/validation and 2023 data for evaluation.

2.5.3. Implementation and Reproducibility

The analysis used Python 3.13 (Python Software Foundation, Wilmington, DE, USA) with NumPy 2.5.1, pandas 3.0.5, SciPy 1.18.0, scikit-learn 1.9.0, and joblib 1.5.3. On a laptop computer with an Apple M4 processor (Apple Inc., Cupertino, CA, USA; 10 cores, 24 GB memory), fitting the logistic regression on the 59,613 development rows took 0.7 s, the whole selection procedure of Section 2.5.1 about 2 min, and one 30-h forecast 22 ms (17 ms for the probabilities and 5 ms for decoding), given the constructed input variables. The reproducibility package contains the input-variable list, data splits, fitted models and weights, and the scripts that regenerate every table and figure; reproduction requires access to the TAF, METAR/SPECI, and A-CDM records under the conditions of the Data Availability Statement. Generative AI tools were used to assist in writing the research code, including the plotting scripts, and to translate and revise the manuscript text; all figures were produced by the authors by running these scripts on the study data; the authors reviewed, tested, and edited all AI-assisted code and text and take full responsibility for the content of this publication.

3. Computational Results

The 2024 external evaluation generated a total of 43,914 hourly prediction rows across the four daily forecast origins. Model accuracy and probability quality were evaluated exclusively on the 33,442 rows containing a confirmed runway direction. Furthermore, sequence stability was assessed using 30,429 pairs of adjacent labeled hours within individual forecast runs, while the precise timing of runway changes was analyzed across 6306 unique consecutive target-hour pairs, which captured 275 actual operational direction changes.

3.1. Runway-Direction Prediction Performance

In the 2024 test set, the final decoded path achieved an accuracy of 89.16% over the short-term horizon (T+1–T+6) and 85.00% over the full 30-h horizon (T+1–T+30) (Table 4). The proposed framework outperformed the baseline TAF wind-direction rule by margins of 5.73 and 4.09 percentage points, respectively. It also exceeded the persistence–hysteresis wind rule by 3.21 and 2.28 percentage points. By contrast, the persistence baseline (maintaining the most recently observed direction) achieved 86.93% accuracy over T+1–T+6, but its performance deteriorated significantly to 73.46% over the extended T+1–T+30 horizon.

Table 4. Runway-direction prediction performance by forecast horizon in the primary 2024 external evaluation.

Prediction Method	T+1–T+6 Accuracy	Rows	T+1–T+30 Accuracy	Rows
Final path (logistic regression + Viterbi decoding)	89.16%	6689	85.00%	33,442
Hour-by-hour logistic-regression decisions	89.09%	6689	85.02%	33,442
Persistence–hysteresis wind rule	85.95%	6689	82.72%	33,442
Simple rule based solely on TAF wind direction	83.43%	6687	80.91%	33,440
Persistence of the most recent runway direction	86.93%	5456	73.46%	25,649

Note: 'Rows' denotes the number of forecast-origin–target-hour pairs that possess both the requisite inputs for the respective method and a confirmed ground-truth direction label; consequently, sample sizes vary across methods. The 'Final path' is decoded from the logistic-regression probabilities, whereas the 'Hour-by-hour logistic-regression decisions' assign RWY 25 whenever the same probability is at least 0.5.

When evaluated on the 25,647 intersecting rows computable by all methods, the final path achieved an 85.27% accuracy, outperforming the TAF rule (81.39%) and the persistence baseline (73.46%). Specifically, the final path exceeded the TAF rule by 3.88 percentage points (95% CI: +2.64 to +5.21; McNemar's test $p < 0.001$) and surpassed the persistence baseline by 11.81 percentage points (95% CI: +9.86 to +13.79; $p < 0.001$).

Against the persistence–hysteresis rule, the final path was 2.28 percentage points more accurate over all 33,442 rows (95% CI +1.60 to +3.01; McNemar $p < 0.001$) and 1.82 percentage points more accurate on the common rows (95% CI +1.07 to +2.60). The hourly probability had a log loss of 0.3547, a Brier score of 0.1098, and an expected calibration error of 1.396% with equal-width bins and 1.365% with equal-mass bins.

The model relied most on the TAF wind at and around the target hour. The largest standardized logistic-regression coefficients belonged to the along-runway component of the target-hour TAF wind (−1.190), the sine of the target-hour TAF wind direction (0.530), and the preceding runway direction in the T+1–2 lead-time band (0.478). Because the inputs are correlated, these values show what the fitted model used, not causal effects.

3.2. Contribution of Each Information Source

Table 5 compares the complete pipeline built with the four input sets. Over the full horizon, recent runway use was the only source with a detectable gain: adding it to the TAF inputs raised accuracy from 84.62% to 85.02% (+0.40 percentage points; 95% CI +0.01 to +0.83; McNemar $p < 0.001$). Adding METAR/SPECI did not improve accuracy (−0.06 percentage points; 95% CI −0.32 to +0.21; $p = 0.497$), nor did all 208 inputs outperform the 131 TAF-plus-runway inputs (−0.02 percentage points; $p = 0.750$). Over the first four hours (T+1–T+4), however, adding METAR/SPECI to the TAF inputs raised accuracy from 86.92% to 87.73% (+0.80 percentage points; 95% CI +0.27 to +1.36), and adding recent runway use raised it to 90.18% (+3.26 percentage points; 95% CI +1.98 to +4.60); once recent runway use was included, the observations added nothing further (−0.27 percentage points; 95% CI −0.67 to +0.11). All 208 inputs were nevertheless kept, because the input set had been fixed before the evaluation.

Table 5. Contribution of each information source within the complete pipeline in the 2024 external evaluation.

Input Configuration	Inputs	Accuracy	Δ vs. TAF Only	95% CI	*p*	T+1–T+4 Accuracy	Δ vs. TAF Only (95% CI)
TAF only	114	84.62%	—	—	—	86.92%	—
TAF + METAR/SPECI	191	84.57%	−0.06 pp	−0.32 to +0.21	0.497	87.73%	+0.80 pp (+0.27 to +1.36)
TAF + recent runway	131	85.02%	+0.40 pp	+0.01 to +0.83	<0.001	90.18%	+3.26 pp (+1.98 to +4.60)
Full (all three sources)	208	85.00%	+0.37 pp	+0.01 to +0.77	<0.001	89.91%	+2.99 pp (+1.85 to +4.23)

Note: All configurations were evaluated on the same 33,442 labeled 2024 rows, with the inverse regularization strength and the Viterbi settings re-selected for each; C = 0.01 was returned in all cases. Accuracy is that of the decoded path; the T+1–T+4 columns use the 4481 rows of those lead times. Differences and confidence intervals are relative to the TAF-only configuration; intervals come from a weekly block bootstrap with 2000 resamples and *p*-values from McNemar's test. A dash (—) marks the reference configuration.

3.3. Performance by Forecast Lead Time

Accuracy decreased with lead time, but the final path stayed above both the TAF rule and the persistence–hysteresis rule in all five lead-time bands (Figure 4), from 91.59% against 86.89% for the TAF rule at T+1–2 to 81.89% against 80.09% at T+25–30. The margin

over the persistence–hysteresis rule ranged from +0.32 to +4.63 percentage points across the bands.

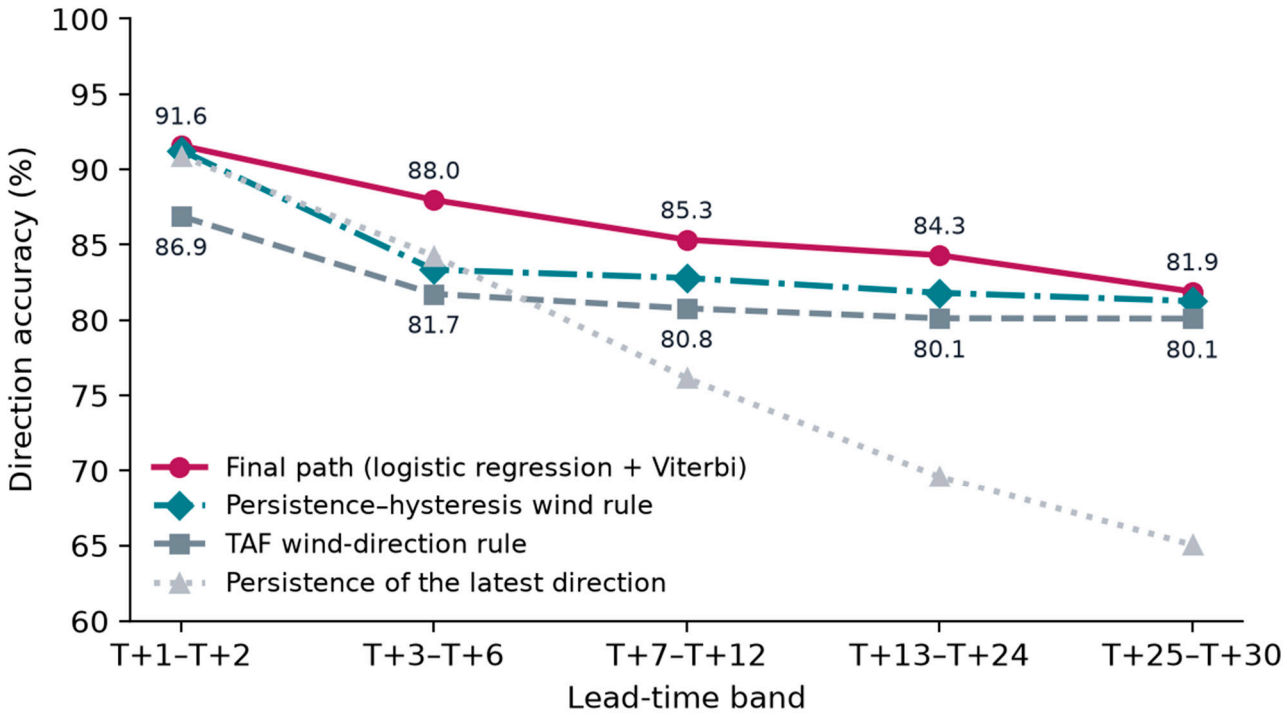

Figure 4. Runway-direction accuracy by lead-time band, 2024 external evaluation.

3.4. Effect of Viterbi Decoding

Evaluating each target hour independently based on the logistic-regression probabilities yielded an accuracy of 85.02%, closely comparable to the 85.00% achieved by the Viterbi decoded path. However, independent evaluation resulted in 722 direction changes across the 30,429 adjacent labeled pairs (a switching rate of 2.37%). In contrast, the decoded path exhibited only 522 changes (1.72%), representing a 27.7% reduction in predicted switching. The difference in overall accuracy was not statistically distinguishable from zero (McNemar's test $p = 0.772$; block-bootstrap 95% CI −0.28 to +0.23), and decoding marginally decreased accuracy in the 2023 robustness check (Section 3.6.2). Consequently, the primary contribution of Viterbi decoding lies not in improving average point-wise accuracy, but in the temporal stability of the predicted 30-h sequence, and its effect is therefore assessed below on the predicted transitions.

The consequence for transition timing is a trade of recall for precision. Decoding reduced the predicted transitions from 599 to 438; the false alarms fell from 505 to 355, and the predicted transitions matched to one of the 275 observed transitions within ±1 h fell from 94 to 83. Precision rose from 15.7% to 18.9%, recall fell from 34.2% to 30.2%, and the F1 score rose from 21.5% to 23.3%; the F1 difference was not statistically distinguishable from zero (+1.8 points; 95% CI −1.3 to +4.6). The same direction of change held at the exact hour and within ±3 h (Table 6). Of the 275 observed transitions, 70 were matched both before and after decoding, 24 only before, 13 only after, and 168 in neither case. Most of those lost (17 of 24) were cases in which the predicted transition moved to two or three hours from the observed one rather than disappearing, and most of those gained (10 of 13) were cases in which it moved closer. Brief transitions, after which the new direction lasted less than three hours, were matched in 16 of 77 cases before decoding and in 12 after, with none gained; this is the cost of removing short reversals.

Table 6. Effect of Viterbi decoding in the 2024 evaluation.

Measure	Before Decoding (Hour-by-Hour Decisions)	After Decoding (Final Path)
Direction accuracy (33,442 rows)	85.02%	85.00%
Correct rows	28,434	28,425
Accuracy in stable hours (30,153 rows)	87.17%	87.10%
Accuracy in transition hours (1375 rows)	42.55%	43.64%
Predicted direction changes within runs	722	522
One-hour reversals	101	9
Predicted transitions	599	438
Predicted transitions matched within ±1 h (275 observed transitions)	94	83
False alarms (±1 h)	505	355
Precision/recall/F1, exact hour	8.2%/17.8%/11.2%	10.0%/16.0%/12.3%
Precision/recall/F1, ±1 h	15.7%/34.2%/21.5%	18.9%/30.2%/23.3%
Precision/recall/F1, ±3 h	23.2%/50.5%/31.8%	29.9%/47.6%/36.7%

Note: Both columns use the same logistic-regression probabilities. Stable and transition hours are defined in Section 2.5.2; the remaining 1914 rows are first hours after a gap (81.71% before and 81.56% after decoding). A predicted transition is a target hour at which the path of any forecast run changes direction; a false alarm is a predicted transition not matched to an observed transition within ±1 h.

Table 7 accounts for every row in which decoding changed the decision, by the kind of edit made to the path. Rows were gained where decoding brought a predicted direction change forward (+137 against −101) and where it flattened reversals of up to three hours (+68 against −42); rows were lost where it delayed a predicted direction change (+58 against −96) and where it overwrote a stretch of four hours or more (+36 against −79). These gains and losses cancel. Nearly all edits (581 of 589 stretches) turned RWY 07 decisions into RWY 25: because the emission scores are computed relative to the baseline share $\pi = 0.438$ (Section 2.4.2), rows with a probability between 0.438 and 0.5 move to RWY 25, and in such rows the actual direction was RWY 25 in 49.9% of cases. As a result, accuracy rose from 82.26% to 84.82% in hours actually operated as RWY 25 and fell from 87.18% to 85.14% in hours operated as RWY 07. This was the only one of the conditions examined (lead time, time of day, season, wind speed, distance of the wind direction from the rule boundary, and distance from a transition) under which decoding changed accuracy detectably.

The size of this exchange depends on the strength of the decoder. With the setting of the originally submitted version ($\gamma = 0.25$, $\eta = 2.0$, $\rho = 0.98$), the final path was 84.96% accurate, with 396 predicted transitions of which 77 were matched and 319 were false (±1-h F1 score 23.0%). Stronger persistence removed more false alarms but also more correct ones and lowered accuracy (for $\eta = 2.0$ and $\rho = 0.70$: 317 false and 77 matched predicted transitions with 84.97% accuracy at $\gamma = 0.25$, and 235 false and 61 matched predicted transitions with 84.80% accuracy at $\gamma = 1.0$). These values are shown for sensitivity and were not used for selection.

Figure 5 illustrates two representative 2024 forecast runs selected based on predefined criteria. In these visualizations, shaded regions denote non-operational hours (e.g., curfew periods without flights), and the plotted traces are offset vertically to enhance legibility.

Table 7. Rows in which Viterbi decoding changed the hour-by-hour decision, by the kind of edit made to the path (2024).

Edit Made by Decoding	Stretches	Rows Changed	Newly Correct	Newly Wrong	Net
Reversal removed (A → B → A flattened to A), up to 3 h	137	110	+68	−42	+26
Reversal removed, 4 h or longer	25	115	+36	−79	−43
Reversal created	3	10	+9	−1	+8
Predicted direction change delayed	166	154	+58	−96	−38
Predicted direction change brought forward	198	238	+137	−101	+36
Predicted direction change removed (no return within the run)	45	103	+50	−53	−3
Predicted direction change created (no return within the run)	15	35	+20	−15	+5
Total	589	765	+378	−387	−9

Note: A stretch is a run of consecutive target hours, within one forecast run, on which the final path differs from the hour-by-hour decisions. Rows are labeled rows; a stretch may include unlabeled hours, so the rows of a category can be fewer than its stretches.

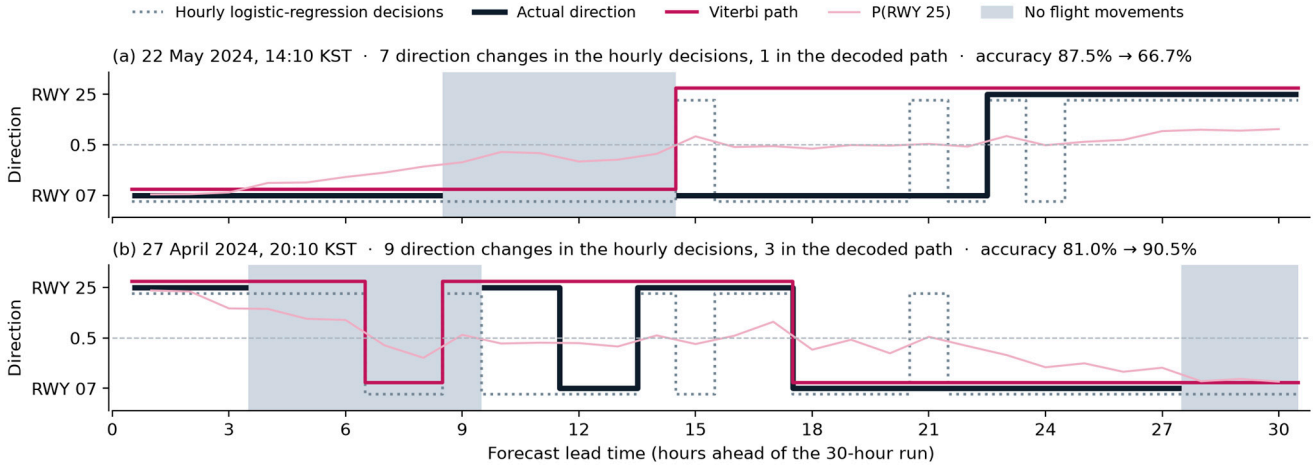

Figure 5. Predicted and actual 30-h direction paths for two 2024 forecast runs: (**a**) run issued at 14:10 KST on 22 May 2024; (**b**) run issued at 20:10 KST on 27 April 2024. In the panel titles, the arrow (→) separates the value for the hourly decisions (left) from that for the decoded path (right).

In the run issued at 14:10 KST on 22 May (Figure 5a), the actual operational direction switched from RWY 07 to RWY 25 at T+23. Because the logistic-regression probability fluctuated around the 0.5 threshold from T+15 onward, independent hour-by-hour decisions changed direction seven times before settling on RWY 25 at T+25. The Viterbi decoded path replaced these with a single transition, but placed it at T+15, eight hours before the actual change, because the probability already exceeded the baseline share from that hour; accuracy over the run fell from 87.5% to 66.7%.

In the 27 April run (Figure 5b), the logistic-regression probability fluctuated between 0.30 and 0.66 during the T+7–T+23 window, causing the independent hour-by-hour evaluation to produce nine direction switches. The decoded path mitigated this instability, executing only three transitions and achieving a 9.5 percentage point increase in accuracy over the entire run. However, this smoothing effect failed to capture a brief, genuine two-hour reversal at T+12–T+13.

These case studies demonstrate an inherent trade-off: while the enforced operational persistence filters probabilistic noise near the 0.5 decision boundary, it can inadvertently suppress short-lived actual changes and mistime transitions driven by weak predictive signals.

3.5. Error Decomposition and Temporal Variability

Predictive accuracy exhibited significant temporal variation throughout the year (Figure 6a), peaking at 92.78% in February and dropping to a nadir of 78.33% in May. Notably, the lowest accuracy levels occurred during May, January, and June—months characterized by a highly balanced utilization of the two operating directions (where the RWY 25 share ranged narrowly between 47.5% and 55.7%). The performance margin of the proposed model over the baseline TAF rule was most pronounced in September and October (gains of 8.44 and 8.07 percentage points, respectively) and most marginal in June and July (0.74 and 1.15 percentage points). These findings suggest that prediction errors are disproportionately concentrated during periods of operational equilibrium—when the direction decision is highly sensitive to marginal shifts in weather—rather than being strictly tied to a specific climatological season. Finally, Figure 6b illustrates the corresponding accuracy stratified by lead-time bands across the common intersecting rows.

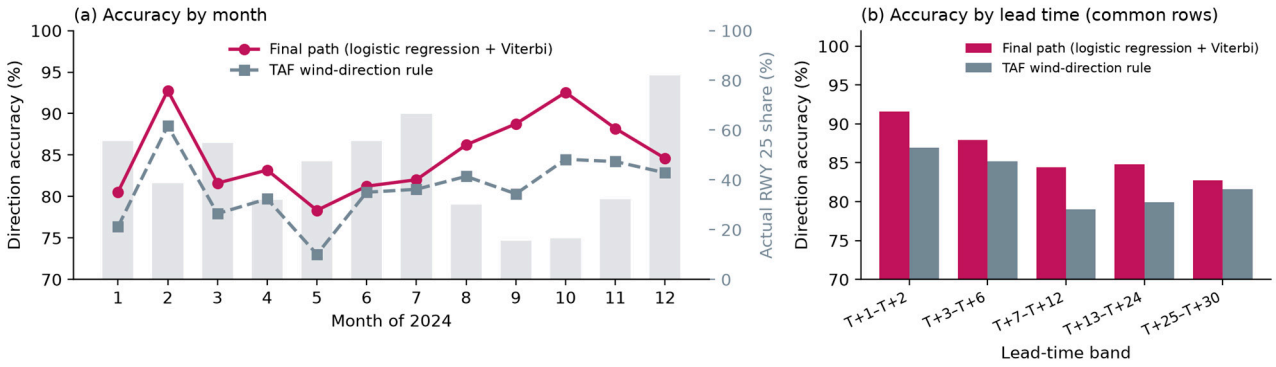

Figure 6. Error decomposition for the 2024 external evaluation: (**a**) accuracy by month, with the actual RWY 25 share as bars; (**b**) accuracy by lead-time band on the rows common to all methods.

3.6. Supplementary Evaluations

3.6.1. Evaluation of Direction-Transition Timing

Table 8 evaluates the temporal accuracy of the predicted transitions against the 275 actual operational transitions. The final decoded path generated 438 transition events, a marked reduction compared to the 599 changes produced by independent hour-by-hour decisions and the 600 by the TAF rule, and comparable to the 417 by the persistence–hysteresis rule.

In terms of timing accuracy, the final path achieved a precision, recall, and F1 score of 10.05%, 16.00%, and 12.34% at the exact target hour. These metrics improved to 18.95%, 30.18%, and 23.28%, respectively, when allowing a ±1-h tolerance window, representing the highest ±1-h precision and F1 scores among the four evaluated methods. However, the performance differences compared to the hour-by-hour decisions were not statistically significant (Section 3.4). Notably, no single method exceeded an F1 score of 25% within the ±1-h window, underscoring the inherent meteorological and operational difficulty of pinpointing the exact hour of runway configuration changes.

Table 8. Evaluation of transition timing derived from runway-direction paths in 2024.

Direction Path	Predicted Transitions	Exact-Hour Precision	Exact-Hour Recall	Exact-Hour F1	±1-h Precision	±1-h Recall	±1-h F1
Hour-by-hour logistic-regression decisions	599	8.18%	17.82%	11.21%	15.69%	34.18%	21.51%
Simple TAF wind-direction rule	600	7.67%	16.73%	10.51%	14.50%	31.64%	19.89%
Persistence–hysteresis wind rule	417	8.63%	13.09%	10.40%	17.03%	25.82%	20.52%
Final path	438	10.05%	16.00%	12.34%	18.95%	30.18%	23.28%

Note: The evaluation is based on 6306 unique consecutive target-hour pairs containing 275 observed operational direction changes. Predicted and actual transitions were matched one-to-one, both at the exact target hour and within a ±1-h tolerance window. Both the hour-by-hour decisions and the final path are derived from the logistic-regression probabilities; predicted transitions of overlapping forecast runs were consolidated as described in Section 2.5.2.

Table 9 reports the ±1-h F1 score by lead-time band, with predicted and actual changes formed from the rows of each band. The final path was ahead of the other methods only in the first six hours (18.01% against 12.29–12.65%); at T+7–12 and T+25–30 it matched fewer transitions than the TAF rule (9.06% against 17.30%, and 4.90% against 12.77%). The scores by band are lower than the overall score because, overall, an actual change counts as matched if any of the overlapping runs anticipates it.

Table 9. ±1-h transition-timing F1 score by lead-time band in 2024.

Direction Path	T+1–6	T+7–12	T+13–24	T+25–30
Hour-by-hour logistic-regression decisions	12.65%	12.20%	17.02%	8.89%
Simple TAF wind-direction rule	12.61%	17.30%	19.23%	12.77%
Persistence–hysteresis wind rule	12.29%	15.29%	16.48%	7.74%
Final path	18.01%	9.06%	17.49%	4.90%

Note: Predicted and actual changes were formed from the rows of each lead-time band only (212 actual changes in each of the three six-hour bands and 275 in T+13–24), and matched as in Table 8.

Table 10 separates direction accuracy by the kind of hour. In the 30,153 stable rows, the final path was 87.10% accurate, against 85.03% for the persistence–hysteresis rule and 83.23% for the TAF rule. In the 1375 transition rows, no method reached 50% (43.64% for the final path and 45.02% for the TAF rule), and within three hours of an actual change the final path and the TAF rule were equally accurate (58.68% and 58.67%).

Table 10. Direction accuracy by kind of hour in 2024.

Prediction Method	Stable Hours (30,153 Rows)	Transition Hours (1375 Rows)	First Hour After a Gap (1914 Rows)	Within ±3 h of a Change (6745 Rows)	Elsewhere (26,697 Rows)
Final path	87.10%	43.64%	81.56%	58.68%	91.65%
Hour-by-hour logistic-regression decisions	87.17%	42.55%	81.71%	58.64%	91.69%
Persistence–hysteresis wind rule	85.03%	42.18%	75.50%	57.79%	89.02%
Simple rule based solely on TAF wind direction	83.23%	45.02%	70.11%	58.67%	86.53%
Persistence of the most recent runway direction	75.21%	39.15%	69.76%	49.81%	79.53%

Note: Labeled rows of the 2024 evaluation. A transition hour is an hour whose direction differs from that of the preceding labeled hour; the first hour after a gap has no labeled preceding hour. The TAF rule cannot be computed for two rows, and persistence is computed on the 25,649 rows with a known direction at the forecast origin.

3.6.2. Second Evaluation Year

Table 11 repeats the comparison with 2022 for training and 2023 for evaluation. The advantage in direction accuracy replicated: the final path reached 81.50%, staying 3.51 percentage points above the TAF rule on the common rows (95% CI +2.10 to +5.03) and 1.83 percentage points above the persistence–hysteresis rule (95% CI +0.57 to +3.01). Decoding reduced predicted direction changes by 20.2% but lowered accuracy from 81.82% to 81.50% (−0.31 percentage points; 95% CI −0.81 to +0.15; McNemar $p = 0.007$). The advantage in transition timing did not replicate: the ±1-h F1 score of the final path (23.39%) did not exceed that of the TAF rule (23.75%) or of the persistence–hysteresis rule (25.32%). The framework therefore improved hourly direction accuracy over the rules in both years, but transition timing in 2024 only.

Table 11. The logistic regression–Viterbi pipeline in two evaluation years.

Measure	2023 (Developed on 2022)	2024 (Developed on 2022–2023)
Labeled evaluation rows	28,925	33,442
Accuracy, hour-by-hour decisions	81.82%	85.02%
Accuracy, final path	81.50%	85.00%
Difference (95% CI); McNemar p	−0.31 pp (−0.81 to +0.15); 0.007	−0.03 pp (−0.28 to +0.23); 0.772
Margin over TAF rule on common rows (95% CI)	+3.51 pp (+2.10 to +5.03)	+3.88 pp (+2.64 to +5.21)
Margin over persistence–hysteresis rule (95% CI)	+1.83 pp (+0.57 to +3.01)	+2.28 pp (+1.60 to +3.01)
Log loss	0.4173	0.3547
Brier score	0.1317	0.1098
ECE (equal mass)	1.63%	1.36%
Direction changes, hour-by-hour → decoded	792 → 632	722 → 522
±1-h transition F1, hour-by-hour/final path	21.39%/23.39%	21.51%/23.28%

Note: Each year was evaluated with a model, Viterbi settings, and a transition matrix obtained only from the preceding development period ($\gamma = 0.1$, $\eta = 4.0$, $\rho = 0.70$ for 2023). Accuracy and the margins are unweighted (the margin over the TAF rule on the rows common to all methods, that over the persistence–hysteresis rule on all labeled rows); log loss, Brier score, and ECE are weighted as in Section 2.2. Arrows (→) give the values before and after decoding. The 2023 timing evaluation contained 213 observed direction changes.

4. Discussion

4.1. Interpretation of Prediction Performance

The final decoded path outperformed the baseline TAF wind-direction rule across the entire 30-h forecast horizon, a predictive advantage that also held on the intersecting common rows (Section 3.1). As demonstrated in Section 3.2, this performance gain did not emerge from a blanket inclusion of all available input sources, but rather from the combination of meteorological forecasts with recent operational runway utilization. Consistent with prior findings for TAF-based configuration predictions at US airports [10], predictive accuracy naturally degraded as the lead time increased (Figure 4). Consequently, the 30-h predicted sequence functions primarily as a strategic probabilistic outlook. Its reliability is intrinsically tied to the forecast horizon, necessitating iterative updates as subsequent weather forecasts and real-time observations become available.

This accuracy largely reflects hours in which the direction persists. In stable hours the final path was 87.10% accurate, whereas in the hours in which the direction changed, no method reached 50% (Table 10), and fewer than a third of the actual changes were matched within ±1 h (Table 8). The advantage of the framework over the rules therefore lies in identifying which direction persists rather than the hour of change.

4.2. Contribution of Forecasts, Observations, and Recent Runway Operations

The baseline TAF rule exhibited acute vulnerabilities near directional boundaries and under light wind conditions (Figure 3c,d). To mitigate this, boundary distance, wind speed, gusts, runway-axis wind components, and surrounding-hour forecasts were integrated into the model space. Interestingly, although the baseline rule deteriorated significantly when the TAF and METAR disagreed (Figure 3e), the formal inclusion of live observations did not yield a statistically significant accuracy improvement over the full horizon (Table 5); it did so over T+1–T+4 only when recent runway use was not among the inputs.

This suggests a rapid temporal decay in the predictive value of current observations; beyond the first few target hours, the TAF intrinsically subsumes any incremental information that a METAR might provide. This finding aligns with Heffar and Dalmau [11], who observed that the latest observation remained the superior input only for lead times up to approximately four hours. By contrast, incorporating recent runway use introduced operational inertia absent from meteorological reports, consistent with the configuration persistence documented at LaGuardia and Newark airports [7]. Because this persistence proved highly accurate in the immediate hours but inadequate over a full 30-h horizon (Table 4), the model was designed to let its influence naturally decay as lead time increased.

These results also place the framework relative to earlier work. Dalmau and Herrema evaluated six-hour sequences using observed future weather and demand, which they describe as a best-case scenario [17], and Wang and Zhang used gridded weather forecasts [21]; the accuracies reported there are therefore not directly comparable with those obtained here from the TAFs as issued. Within the present data, the pipeline that used the TAF inputs alone reached 84.62% over T+1–T+30, within 0.4 percentage points of the full input set, whereas over T+1–T+4, recent runway use raised accuracy from 86.92% to 90.18% (Table 5). The issued TAF thus carries most of the information usable beyond the first few hours, and observed operations matter mainly at short lead times.

4.3. Choice of the Probability Model and Role of Viterbi Decoding

We initially considered stacking [28,29] because LightGBM [30] could capture nonlinear relationships between wind direction and speed and potentially correct some of the logistic regression model's errors. We evaluated this possibility using forward-in-time validation within the 2022–2023 development period. On the same 28,791 validation rows, we applied the Viterbi settings selected for logistic regression to all three variants. The resulting hourly direction accuracies were 81.72% for logistic regression, 81.50% for LightGBM, and 82.09% for the stack. Stacking reduced the log loss from 0.4121 to 0.4003. Its gain in direction accuracy over logistic regression was 0.36 percentage points, with a 95% confidence interval of −0.07 to +0.82 percentage points based on 2000 weekly block bootstrap resamples.

As a supplementary analysis, we tested whether the two coefficients in the stacking model differed. The original month-based validation yielded coefficients of 0.434 for logistic regression and 0.454 for LightGBM. The difference was 0.021 (LightGBM minus logistic regression, 95% CI −0.168 to +0.208, two-sided $p = 0.841$). With forward-in-time validation, the difference was −0.073 (95% CI −0.309 to +0.147, $p = 0.498$). The confidence intervals included zero under both validation schemes. We retained regularized logistic regression as the sole probability model for Viterbi decoding because it was easier to interpret and the

development data provided limited evidence that adding a second model improved the study's primary output, the 30-h sequence of hourly runway directions. The 2024 evaluation data were excluded from these comparisons.

Viterbi decoding optimizes the entire 30-h sequence globally rather than evaluating each hour independently [23]. While it left the average point-wise accuracy largely unchanged, it reduced predicted direction changes and false alarms at the cost of some correctly timed predicted transitions (Section 3.4); smoothing thus removes false and genuine predicted transitions together, and whether this exchange is beneficial depends on the relative cost of a false alarm and a missed transition, which was not measured here. However, a reduction in predicted switches does not guarantee exact temporal precision for actual changes: with a timing F1 score of 23.28% within a ±1-h tolerance (Table 8), the decoded path remains a strategic hourly outlook rather than a precise tactical runway-change alert. Similarly, the reported hourly probability quantifies the directional uncertainty for that specific hour, not the joint probability of the entire sequence or the exact likelihood of a transition event.

The transition matrix is already direction-specific (Equation (3)), and the dispersion coefficient selected on the development data ($\gamma = 0.1$) is weaker than the one used in the originally submitted version ($\gamma = 0.25$), which reduces the over-smoothing of that version. The remaining limit lies less in the decoder than in the probabilities: even without smoothing ($\gamma = 0$), only 98 of the 275 actual changes were matched within ±1 h. This is consistent with the view of a configuration change as a process that takes time to complete [19]: 80% of the hours in which the label changed contained movements in both directions (Section 2.2), so the hour of change is not sharply defined at the resolution of the label. A separate state for the transition period, or transition probabilities that depend on the time of day, would require labels finer than one hour to be estimated reliably, together with wind observations at a comparable resolution such as those used at Madeira [15], and are left for future work.

4.4. Operational Implications and Transferability

The framework provides an input to airport-capacity forecasting, not a capacity estimate; capacity also depends on weather, the arrival–departure ratio, and fleet mix [1]. Linked to direction-specific throughput and demand, the direction path and probabilities could support demand–capacity planning, but capacity, delay, and workload benefits were not measured here. In operation, they would be shown with lead time and input availability, final decisions would remain with controllers, and a predicted transition would mark a period for closer review rather than trigger an action. The framework uses no Jeju-specific rule and could be applied at other airports operating a single runway or one primary runway after the runway-axis variables are redefined and the models are retrained on local data, but its transferability must be shown by airport-specific validation, and it does not extend to airports where configurations combine several runways. Computation is not a constraint for such use (Section 2.5.3).

4.5. Limitations and Future Work

A key limitation is the hourly label itself. Because it represents the direction used by the majority of movements in each hour, it inherently merges mixed-direction operations and short transitions, meaning the timing results apply only at this hourly resolution. Additionally, the inputs currently omit traffic demand, noise restrictions, runway closures, NOTAMs, and airspace congestion. Since demand is known to influence configuration selection [8], its absence likely explains part of the remaining error. The framework therefore predicts the direction that the wind and recent operations make likely, which may differ from the direction a controller would choose; aggregated demand from filed flight plans is

a natural additional input for future work. The exclusion of RWY 13/31 also means that configurations using the crossing runway are not represented, which limits the method to airports with one clearly dominant runway. The main evaluation is also limited to one year at a single airport, and only hours with confirmed directions were scored. Furthermore, because both evaluation years had been examined in the originally submitted version of this work (Section 2.5.1), neither is a first look at unseen data.

Finally, the ten-minute availability delay was assumed rather than empirically recorded. Delays of up to ten minutes left the results unchanged, whereas delays of 15–30 min, at which the newest routine TAF is not yet available at the forecast origin, lowered accuracy by about 1.4–1.5 percentage points (Appendix B). This sensitivity to the timeliness of the inputs is related to, though distinct from, the effect of the reporting frequency of meteorological data examined by Rezo and Steiner [12]. Robustness to degraded inputs, which has been examined for a runway configuration assistance tool [20], was tested here only for delayed availability; noisy or biased reports were not examined. A prospective test using fixed settings under real data delivery delays, missing inputs, and live forecast updates is necessary. Once completed, linking the model outputs to a capacity and traffic-flow model would allow their true operational value to be measured. Improving the detection of short direction changes and conducting tests at other primary-runway airports are important further steps.

5. Conclusions

This study developed and evaluated a predictive framework to forecast the hourly operating direction of a primary runway over a 30-h horizon, using RWY 07/25 at Jeju International Airport as a case study. Executed at four daily forecast origins, the methodology employs a regularized logistic regression to estimate directional probabilities using TAFs, METAR/SPECI observations, and recent runway utilization derived from A-CDM records. Viterbi decoding processes these probabilities to determine the optimal 30-h operational sequence.

In the primary 2024 evaluation, the final decoded path achieved an accuracy of 89.16% over the first 6 h and 85.00% over the full 30-h horizon, outperforming the baseline TAF wind-direction rule by 4.09 percentage points overall and a persistence–hysteresis wind rule by 2.28 percentage points. Recent runway use emerged as the sole additional input source that provided a statistically significant predictive gain beyond the baseline meteorological forecasts over the full horizon. Although correctly timed predicted transitions fell from 94 to 83, the application of Viterbi decoding reduced predicted direction changes by 27.7% and false alarms from 505 to 355 while preserving overall point-wise accuracy.

Because predictive accuracy naturally degrades with increasing lead time, and precisely pinpointing transition times remains inherently challenging (yielding an F1 score of 23.28% within a $\pm$1-h tolerance), the framework's output functions best as a strategic probabilistic outlook rather than a precise tactical alert. Within this context, the proposed model may provide an input to airport-capacity forecasting and ATFM, although these operational benefits were not tested in this study. Subject to localized retraining and validation, this framework may be adapted to support demand-capacity balancing and surface-movement planning at other airports operating a single or primary runway.

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

Funding: This research was supported by a Korea Meteorological Institute (KMI) grant funded by the Aviation Meteorological Office (AMO) (Grant RS-2022-KM220510).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The TAF and METAR/SPECI reports analyzed in this study are openly available from the Korea Meteorological Administration (KMA) open API and from OGIMET. The runway-direction labels were derived from A-CDM records provided by Korea Airports Corporation; these are operational records subject to redistribution restrictions and are therefore not publicly available, and requests for access should be directed to Korea Airports Corporation. The Python code used for preprocessing, feature construction, model training, Viterbi decoding, evaluation, and figure generation is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Code (Anthropic, Claude Opus 5.5 and Claude Fable 5.2) and Codex (OpenAI, GPT-6 Sol and GPT-6 Astra) to assist in writing the research code, including the plotting scripts with which the authors produced the figures from the study data, and to translate and revise the manuscript text, and used Gemini (Google, Gemini 3.1 Pro) to a limited extent to translate and revise the manuscript text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

A-CDM	Airport Collaborative Decision Making
ATFM	Air Traffic Flow Management
ECE	Expected calibration error
KST	Korea Standard Time
LightGBM	Light gradient-boosting machine
METAR	Aerodrome routine meteorological report
NOTAM	Notice to Airmen
QNH	Altimeter setting referenced to mean sea level
RWY	Runway
SPECI	Aerodrome special meteorological report
TAF	Terminal Aerodrome Forecast

Appendix A. Label Construction and Evaluation Details

Labels. Movements were taken from the A-CDM records of arrivals and departures on RWY 07/25 at their actual landing and take-off times, and each clock hour received the direction with the larger number of movements. Hours without a movement on RWY 07/25, including the night curfew, have no label; such hours were not scored and were not used for fitting, and two hours were counted as an adjacent pair only when both were labeled and one hour apart.

Ambiguous hours. In a de-duplicated copy of the 2024 movement records, 17 labeled hours had equal counts in the two directions and 295 had a minority share of at least 25% (the count of tied hours differs from the 21 of Section 2.2 because duplicate records were removed). Excluding the tied hours changed the accuracy of the final path from 85.00% to 85.09%, and excluding the hours with a minority share of at least 25% raised it to 86.05%; within those hours, accuracy was 62.3% for the final path and for the TAF rule alike. Errors are thus concentrated in hours in which both directions were in use, and the treatment of ties does not affect the conclusions.

Wind inputs. When a TAF or a METAR reports calm (0 kt) or variable (VRB) wind, the wind direction is treated as missing: the direction variables take the training-period median, their availability flags are set to zero, and a variable-wind flag is set, while the wind speed is retained. The TAF wind-direction rule is undefined for such hours (two labeled rows in 2024).

Conditional TAF groups. When several TEMPO or PROB groups cover the same target hour, one group is used, in the order TEMPO, PROB40, and PROB30; the conditional values are the prevailing values with the elements stated in that group replaced.

Appendix B. Reporting Interval and Availability Delay

Table A1 repeats the complete procedure of Section 2.5.1 with reports assumed to become available 0, 5, 15, 20, and 30 min after their issue or observation time instead of 10 min. Delays of 0 and 5 min changed the 2024 accuracy by no more than 0.03 percentage points: at origins ten minutes past the hour, the inputs differ only when a SPECI or an amended TAF was issued in the last minutes before the origin. Delays of 15 min or more exclude the routine TAF and the routine METAR issued on the hour, so that the forecast rests on the TAF issued six hours earlier, whose validity period does not cover the last hours of the 30-h horizon: 20% of the rows could not be forecast, and on the remaining rows, accuracy was 1.4–1.5 percentage points lower than with the 10-min delay. Between 15 and 30 min the results were nearly constant. The sensitivity is therefore to whether the newest routine TAF has arrived by the forecast origin, and not to the delay of individual observations.

Table A1. Sensitivity of the 2024 results to the assumed availability delay of the meteorological reports.

Assumed Delay	Rows Forecast	Labeled Rows	T+1–T+30 Accuracy	Δ vs. 10 min, T+1–T+30 (95% CI)	Δ vs. 10 min, T+1–T+6 (95% CI)
0 min	43,914	33,442	85.02%	+0.03 (−0.02 to +0.08)	0.00 (−0.08 to +0.06)
5 min	43,914	33,442	85.01%	+0.01 (−0.03 to +0.06)	0.00 (−0.08 to +0.06)
10 min (main analysis)	43,914	33,442	85.00%	—	—
15 min	35,136	26,756	84.27%	−1.50 (−2.10 to −0.90)	−1.45 (−2.33 to −0.60)
20 min	35,136	26,756	84.31%	−1.47 (−2.07 to −0.86)	−1.39 (−2.29 to −0.51)
30 min	35,136	26,756	84.36%	−1.41 (−2.01 to −0.81)	−1.35 (−2.20 to −0.52)

Note: Each delay went through the complete procedure of Section 2.5.1 ($C = 0.01$ and $\gamma = 0.1$ were selected in all cases, with η between 2.0 and 4.0). Differences are in percentage points on the rows forecast under both delays, with weekly block-bootstrap intervals. A dash (—) marks the reference delay of 10 min.

References

1. Gilbo, E.P. Airport capacity: Representation, estimation, optimization. *IEEE Trans. Control Syst. Technol.* **1993**, *1*, 144–154. [CrossRef] [PubMed]
2. Cho, J.; Lee, K.; Baik, H.; Park, J. A study on the pressure patterns that causes bidirectional tailwind on the runway of Jeju International Airport. *J. Korean Soc. Aviat. Aeronaut.* **2023**, *31*, 93–102. [CrossRef]
3. Lee, K. Prediction of Arrival Throughput for Jeju Airport Using Artificial Intelligence. Master's Thesis, Korea Aerospace University, Goyang, Republic of Korea, 2023. (In Korean)
4. Puranik, T.G.; Memarzadeh, M.; Kalyanam, K.M. Predicting airport runway configurations for decision-support using supervised learning. In *Proceedings of the 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), Barcelona, Spain, 1–5 October 2023*; IEEE: New York, NY, USA, 2023; pp. 1–9. [CrossRef]

5. Li, L.; Clarke, J.-P. A stochastic model of runway configuration planning. In *Proceedings of the AIAA Guidance, Navigation, and Control Conference, Toronto, ON, Canada, 2–5 August 2010*; AIAA: Reston, VA, USA, 2010. [CrossRef]
6. Lam, A.J.G.; Alam, S.; Lilith, N.; Piplani, R. A deep reinforcement learning approach for runway configuration management: A case study for Philadelphia International Airport. *J. Air Transp. Manag.* **2024**, *120*, 102672. [CrossRef]
7. Ramanujam, V.; Balakrishnan, H. Estimation of maximum-likelihood discrete-choice models of the runway configuration selection process. In *Proceedings of the 2011 American Control Conference, San Francisco, CA, USA, 29 June–1 July 2011*; IEEE: New York, NY, USA, 2011; pp. 2160–2167. [CrossRef]
8. Avery, J.; Balakrishnan, H. Data-driven modeling and prediction of the process for selecting runway configurations. *Transp. Res. Rec.* **2016**, *2600*, 1–11. [CrossRef]
9. de Leege, A.; Janssen, C. Probabilistic runway and capacity forecasting using machine learning to support decision making. In Proceedings of the SESAR Innovation Days 2016, Delft, The Netherlands, 8–10 November 2016.
10. Raju, R.; Mital, R.; Wilson, B.; Shetty, K.; Albert, M. Predicting runway configurations and arrival and departure rates at airports: Comparing the accuracy of multiple machine learning models. In *Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 3–7 October 2021*; IEEE: New York, NY, USA, 2021; pp. 1–8. [CrossRef]
11. Heffar, A.; Dalmau, R. A universal machine learning model for better runway allocation predictions. In Proceedings of the 11th International Conference on Research in Air Transportation (ICRAT 2024), Singapore, 1–4 July 2024.
12. Rezo, Z.; Steiner, S. Scaling up sampling and meteorological data update frequency: Evidence from the aviation industry. *WUT J. Transp. Eng.* **2025**, *141*, 43–53. [CrossRef]
13. Mahringer, G. Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions. *Meteorol. Appl.* **2008**, *15*, 113–123. [CrossRef]
14. Sharpe, M.A.; Bysouth, C.E.; Trueman, M. Towards an improved analysis of Terminal Aerodrome Forecasts. *Meteorol. Appl.* **2016**, *23*, 698–704. [CrossRef]
15. Alves, D.; Freitas, D.; Mendonça, F.; Mostafa, S.S.; Morgado-Dias, F. Wind limitations at Madeira International Airport: A deep learning prediction approach. *IEEE Access* **2024**, *12*, 61211–61220. [CrossRef]
16. Rebollo, J.; Khater, S.; Coupe, W.J. A recursive multi-step machine learning approach for airport configuration prediction. In *Proceedings of the AIAA AVIATION 2021 Forum, Virtual Event, 2–6 August 2021*; AIAA: Reston, VA, USA, 2021; AIAA Paper 2021-2406. [CrossRef]
17. Dalmau, R.; Herrema, F. Encoder–decoder approach to predict airport operational runway configuration: A case study for Amsterdam Schiphol Airport. In Proceedings of the 9th SESAR Innovation Days, Athens, Greece, 2–5 December 2019.
18. Wang, X.; Liu, L.; Li, S.; Xu, Y.; Low, K.H.; Wu, W. Prediction of multiple runway configurations for efficient flight operations by hybrid LSTM and transformer model. *Sci. Rep.* **2025**, *15*, 37764. [CrossRef] [PubMed]
19. Lau, M.E.C.; Lam, A.J.G.; Alam, S. Predicting runway configuration transition timings using machine learning methods. In *Proceedings of the 2021 Winter Simulation Conference (WSC), Phoenix, AZ, USA, 12–15 December 2021*; IEEE: New York, NY, USA, 2021; pp. 1–12. [CrossRef]
20. Memarzadeh, M.; Wang, Z.; Masrour Shalmani, F.; Razzaghi, P.; Kalyanam, K.M. Responsible AI for air traffic management: Application to runway configuration assistance tool. *Aerospace* **2025**, *12*, 872. [CrossRef]
21. Wang, Y.; Zhang, Y. Prediction of runway configurations and airport acceptance rates for multi-airport system using gridded weather forecast. *Transp. Res. Part C Emerg. Technol.* **2021**, *125*, 103049. [CrossRef]
22. International Civil Aviation Organization. *Meteorological Service for International Air Navigation, Annex 3 to the Convention on International Civil Aviation*, 20th ed.; ICAO: Montréal, QC, Canada, 2018.
23. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef] [PubMed]
24. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
25. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
26. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
27. Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [CrossRef]
28. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]

29. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
30. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates: Red Hook, NY, USA, 2017.

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