

Aplicació de Graph Neural Networks per a la millora de les prediccions meteorològiques als Països Catalans

Treball de Final de Grau

Nil Farrés Soler

Tutors: Jordi Casas Roma, Josep Lladós Canet

Amb la col·laboració de:

Centre de Visió per Computador
Secció de meteorologia de 3Cat

Matemàtica Computacional i Analítica de Dades

Universitat Autònoma de Barcelona
Juny de 2025

Resum

Aquest treball de final de grau explora l'ús de les xarxes neuronals en graf (Graph Neural Networks, GNNs) per a la predicción meteorològica horària als Països Catalans. El projecte parteix de dades meteorològiques recollides entre 2016 i 2024, provinents de centenars d'estacions de fonts fiables. S'ha desenvolupat un pipeline robust de tractament, depuració i codificació de dades per transformar-les en grafs dinàmics temporals aptes per a l'entrenament d'un model predictiu.

El model proposat, anomenat MeteoGraphPC, s'ha implementat mitjançant PyTorch Geometric i entrenat en un entorn amb GPU, utilitzant finestres temporals de 48 i 120 hores i horitzons de predicción de 6 i de 120 hores. Les prediccions a curt i a mitjà termini mostren un aprenentatge estable i una capacitat predictiva limitada però prometedora per modelitzar variables com la temperatura i la pressió atmosfèrica, tot superant la línia base de climatologia. No obstant això, el model encara presenta dificultats de generalització a noves dades meteorològiques, especialment amb variables esporàdiques o no estacionàries, com el vent i especialment la precipitació, on la capacitat predictiva és molt baixa.

El projecte ha posat les bases d'un sistema de predicción meteorològica basada en graf i ha suposat una oportunitat d'aprenentatge tècnic i metodològic en camps com el tractament de dades, la intel·ligència artificial, la meteorologia, la programació avançada en Python, la computació d'alt rendiment i l'aprenentatge profund. Malgrat les dificultats tècniques i computacionals, el projecte ha establert una base sólida per a futures línies de recerca que permeten aprofundir en el potencial de les GNNs en la meteorologia dels Països Catalans.

Abstract

This final degree project explores the use of Graph Neural Networks (GNNs) for hourly weather forecasting in the Catalan Countries. The project is based on meteorological data collected between 2016 and 2024, sourced from hundreds of reliable weather stations. A robust data processing pipeline has been developed to clean, transform, and encode this information into dynamic temporal graphs suitable for training a predictive model.

The proposed model, named MeteoGraphPC, was implemented using PyTorch Geometric and trained in a GPU environment, using temporal windows of 48 and 120 hours and forecasting horizons of 6 and 120 hours. Short- and medium-term forecasts show stable learning and a limited but promising predictive capability for modeling variables such as temperature and atmospheric pressure, outperforming the climatology baseline. However, the model still struggles to generalize to new meteorological data, especially for sporadic or non-stationary variables such as wind and, in particular, precipitation, where predictive performance remains very low.

The project lays the foundation for a graph-based weather forecasting system and has provided a valuable learning experience in areas such as data processing, artificial intelligence, meteorology, advanced Python programming, high-performance computing, and deep learning. Despite technical and computational challenges, the project has established a solid basis for future research lines aiming to further explore the potential of GNNs in weather prediction across the Catalan Countries.

Índex

1	Introducció	3
2	Objectius	3
3	Estat de l'art	4
3.1	Què són les xarxes neuronals en graf?	4
3.2	Models recents en predicció meteorològica: de les GNNs als sistemes híbrids	5
4	Dades inicials i preparació del dataset	7
4.1	Les dades	7
4.1.1	Selecció de fonts fiables	9
4.1.2	Anomalies en les dades	10
4.2	Preprocessament de les dades originals	11
4.2.1	Funcions auxiliars	11
4.2.2	Processament i paral·lelització	11
4.3	Conversió de format: de CSV a objectes Data	12
4.3.1	Paràmetres de normalització pels Països Catalans	13
4.3.2	Generació dels grafs horaris	15
4.3.3	Resum de l'estructura del graf	18
5	Metodologia	19
5.1	Generació de seqüències temporals dinàmiques	19
5.2	Agrupació eficient de seqüències per a l'entrenament massiu	20
5.3	Entrenament, validació i test	21
5.3.1	Preparació dels grups de seqüències	21
5.3.2	Normalització de les dades	21
5.3.3	Inicialització i configuració del model	22
5.3.4	Procés d'entrenament	25
5.3.5	Guardat i restauració del millor model	25
5.3.6	Avaluació sobre el conjunt de test i càlcul de baselines	25
5.3.7	Càlcul i anàlisi de mètriques de rendiment	26
6	Resultats	27
6.1	Predictió a curt termini	27
6.1.1	Analisi de les mètriques	28
6.1.2	Analisi comparativa de les correlacions: predicció vs. dades reals	31
6.1.3	Analisi de les prediccions	32
6.2	Predictió a mitjà termini	36
6.2.1	Analisi de les mètriques	36
6.2.2	Analisi comparativa de les correlacions: predicció vs. dades reals	40
6.2.3	Analisi de les prediccions	41
7	Conclusions	43
7.1	Discussió dels resultats	43
7.2	Aprendentatges adquirits	43
7.3	Reptes, dificultats tècniques i metodològiques	44
8	Treball futur	45
9	Agraïments	46
10	Bibliografia	46
A	Annex: codi font i un dia complet de dades	49

1 Introducció

La meteorologia juga un paper fonamental en la nostra societat, incident en àmbits tan diversos com l'agricultura, la gestió dels recursos naturals, la planificació urbana i la seguretat civil. Els avenços en la recopilació de dades i en els models físics tradicionals han contribuït a millorar la precisió de les prediccions meteorològiques, tot i que encara existeixen reptes significatius, especialment en la resolució espacial i temporal i en la detecció precoç de fenòmens extrems, cada cop més freqüents a causa del canvi climàtic.

En aquest context, l'aparició de la intel·ligència artificial i de les xarxes neuronals en graf (Graph Neural Networks, GNNs) obre noves possibilitats per a l'anàlisi i la predicció dels fenòmens meteorològics. Les GNNs permeten modelar de forma efectiva les relacions complexes entre les dades espacials i temporals, superant certes limitacions dels enfocaments tradicionals. Aquestes arquitectures han demostrat un gran potencial per capturar patrons rellevants en dades d'alta dimensionalitat i complexitat.

El present treball de final de grau pretén desenvolupar un model basat en GNNs que realitzi prediccions meteorològiques, utilitzant les dades obtingudes de les estacions meteorològiques de fonts fiables dels Països Catalans, corresponents al període comprès entre el 2016 i el 2024. Aquestes dades, preprocessades acuradament per garantir-ne la qualitat, inclouen variables clau com la temperatura, la humitat, la pressió atmosfèrica, la direcció i la velocitat del vent, així com la precipitació horària. És per això, que la pregunta que es prenent respondeix és la següent:

L'aplicació de GNNs pot oferir una eina complementària que ajudi a millorar la precisió i la capacitat predictiva dels models meteorològics tradicionals als Països Catalans?

El treball es desenvolupa en diverses fases: primer, el preprocessament i la conversió de les dades originals a un format adequat per a l'entrenament de models de deep learning (*utilitzant torch.geometric*); a continuació, la implementació d'un model meteorològic basat en xarxes neuronals en grafs (que anomenarem MeteoGraphPC: Meteorologia en Grafs pels Països Catalans) i, finalment, la validació del model mitjançant dades no utilitzades durant l'entrenament. Així, aquest estudi pretén aportar noves eines per ajudar a millorar la predicció meteorològica als Països Catalans, amb un especial èmfasi en la detecció precoç d'episodis extrems i en la interpretació de la dinàmica atmosfèrica.

Amb aquesta recerca es pretén aportar una nova eina de suport que, complementant els mètodes tradicionals amb les capacitats d'aprenentatge profund de les GNNs, permet ajudar a obtenir prediccions meteorològiques més ajustades amb l'objectiu de contribuir a una millora dels sistemes de predicció en l'àmbit dels Països Catalans.

2 Objectius

- **Objectiu general:**

Analitzar el potencial de les xarxes neuronals en graf (GNNs) per realitzar prediccions meteorològiques als Països Catalans, a partir de dades horàries obtingudes d'estacions meteorològiques entre 2016 i 2024.

- **Objectius específics:**

1. Recopilar, netejar i preprocessar les dades meteorològiques horàries de diverses estacions dels Països Catalans.
2. Construir una representació en forma de graf dinàmic per integrar relacions espacials i temporals entre estacions.
3. Generar seqüències de grafs preparades per a l'entrenament i validació de models de deep learning.
4. Dissenyar i implementar un model de predicció meteorològica bàsic basat en GNNs adaptat al context meteorològic dels Països Catalans.
5. Realitzar prediccions meteorològiques amb el model implementat i evaluar-ne els resultats a curt i a mitjà termini.

3 Estat de l'art

Els models meteorològics tradicionals, com ara el Weather Research and Forecasting model (WRF), el Global Forecast System (GFS), els models del Centre Europeu de Prediccions Meteorològiques a Mitjà Termini (ECMWF), entre d'altres, han estat durant dècades l'eina de referència en la predicció numèrica del temps. Aquests models es basen en la resolució de les equacions físiques que governen la dinàmica i la termodinàmica de l'atmosfera, oferint una visió global de la situació meteorològica. No obstant això, tot i la seva sofisticació, presenten limitacions importants en termes de resolució espacial i temporal, així com en la detecció precoç d'episodis meteorològics severos.

Amb l'augment de la disponibilitat de dades meteorològiques i el gran creixement dels mètodes d'aprenentatge automàtic, s'ha iniciat una nova etapa en la investigació aplicant tècniques d'aprenentatge profund per complementar els models tradicionals. Inicialment, es van utilitzar arquitectures com les xarxes neuronals convolucionals (CNNs) i les xarxes recurrents (RNNs) per processar imatges de satèl·lit i sèries temporals, millorant la detecció i predicció de certs fenòmens meteorològics. Recentment, però, ha sorgit un nou paradigma d'aprenentatge profund especialment adequat per modelar relacions complexes en dades espacials: les xarxes neuronals en graf (GNNs).

3.1 Què són les xarxes neuronals en graf?

Les xarxes neuronals en graf (Graph Neural Networks, GNNs) són una classe d'arquitectures d'aprenentatge profund dissenyades per treballar directament amb dades que tenen una estructura de graf. A diferència de les xarxes convolucionals (CNNs) que processen imatges sobre una graella regular o les xarxes recurrents (RNNs) que tracten seqüències, les GNNs poden operar sobre estructures arbitràries de nodes i arestes.

El mecanisme central de les GNNs es coneix com a “message passing”. En aquest procés (Figura 1), cada node agrega la informació de les seves característiques pròpies amb els missatges que rep dels seus nodes veïns. Aquests missatges són, en essència, les característiques dels veïns transformades. Aquest procés es repeteix en diverses capes, permetent que un node integri informació de veïns cada vegada més llunyans i capturi dependències espacials a múltiples escales. Cada node d'un graf pot contenir característiques pròpies i cada connexió pot codificar relacions específiques entre nodes.

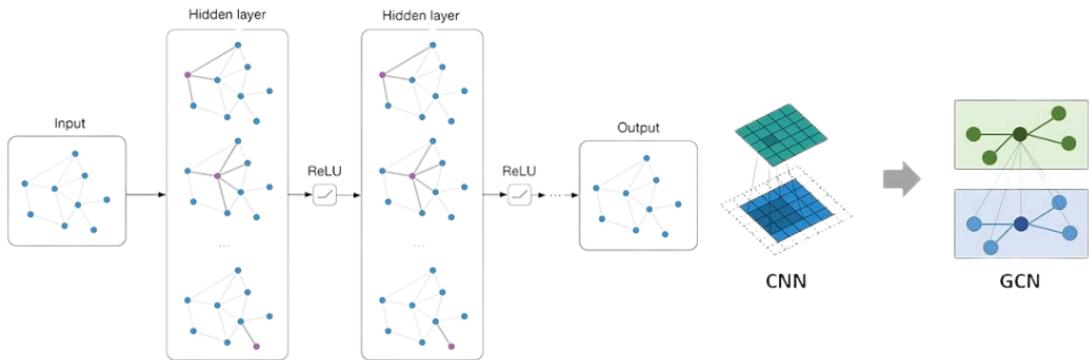


Figura 1: Esquema general d'una xarxa neuronal basada en grafs (GNN) i comparació amb una CNN.

Existeixen diferents variants d'arquitectures GNN, cadascuna amb mecanismes específics per aprendre de les dades estructurades en forma de graf. Les més rellevants, especialment en el context d'aquest treball, són:

- **Graph Convolutional Networks (GCN)**: aquesta és una de les arquitectures més fonamentals. Les GCN generalitzen l'operació de convolució, típicament usada en imatges, a dades amb estructura de graf. El seu funcionament es basa a actualitzar l'estat d'un node (la seva representació vectorial) mitjançant l'agregació de la informació dels seus nodes veïns. Habitualment, aquesta agregació es fa calculant una mitjana ponderada de les característiques dels veïns. Això permet que el model capturi informació de l'entorn local de cada node ([Kipf i Welling \(2017\)](#)).
- **Graph Attention Networks (GAT)**: són una evolució de les GCN que incorporen mecanismes d'atenció. En lloc de ponderar tots els veïns de manera uniforme o amb pesos fixos, les GAT permeten al model aprendre dinàmicament la importància de cada veí. D'aquesta manera, per a un node donat, el model pot prestar més atenció a aquells nodes veïns que siguin més rellevants per a la predicció en un moment concret, millorant la flexibilitat i la capacitat expressiva del model ([Veličković et al. \(2018\)](#)).
- **Temporal Graph Convolutional Network (TGCN)**: aquesta arquitectura és una extensió dissenyada específicament per a dades espaciotemporals, on tant les característiques dels nodes com l'estructura del graf poden canviar amb el temps. La cèl·lula TGCN integra una GCN per capturar les dependències espacials entre les estacions en cada instant de temps, i una xarxa neuronal recurrent (en concret, una GRU) per modelar l'evolució temporal d'aquestes dependències al llarg d'una seqüència ([Zhao et al. \(2019\)](#)).

Aquesta flexibilitat per modelar relacions complexes fa que les GNNs siguin especialment adequades per a problemes on les interaccions entre entitats són crucials, com en xarxes socials, química computacional, sistemes de transport i, com en aquest cas, en la predicció meteorològica basada en xarxes d'estacions disperses en el territori.

3.2 Models recents en predicció meteorològica: de les GNNs als sistemes híbrids

En els darrers anys, aquestes arquitectures han emergit com una eina prometedora per abordar la complexitat de les dades meteorològiques (Figura 2). A diferència dels mètodes tradicionals, les GNNs permeten modelar relacions espacials i temporals en dades que presenten estructures no euclidianes, com és el cas de les estacions meteorològiques distribuïdes geogràficament. Aquestes arquitectures han donat lloc a treballs innovadors com *GraphCast* i *GenCast* (de Google Deep Mind), *FourCastNet* (de NVIDIA), *Aurora* (de Microsoft) i *Pangu-Weather* (de Huawei), entre d'altres¹, que han demostrat una gran capacitat per captar interaccions entre regions atmosfèriques, superant alguns models meteorològics tradicionals i millorant les prediccions d'episodis extrems. De fet, molts d'aquests models basats en GNNs s'estan aplicant de manera experimental al [portal de mapes meteorològics de l'ECWMF](#).

¹**GraphCast**: <https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/>.

Article científic: <https://www.science.org/stoken/author-tokens/ST-1550/full>.

GenCast: <https://deepmind.google/discover/blog/gencast-predicts-weather-and-the-risks-of-extreme-conditions-with-sota-accuracy/>.

FourCastNet: https://docs.nvidia.com/deeplearning/modulus/modulus-v2209/user_guide/neural_operators/fourcastnet.html.

Aurora: <https://www.microsoft.com/en-us/research/blog/introducing-aurora-the-first-large-scale-foundation-model-of-the-atmosphere/>.

Pangu-Weather: <https://www.nature.com/articles/s41586-023-06185-3>.

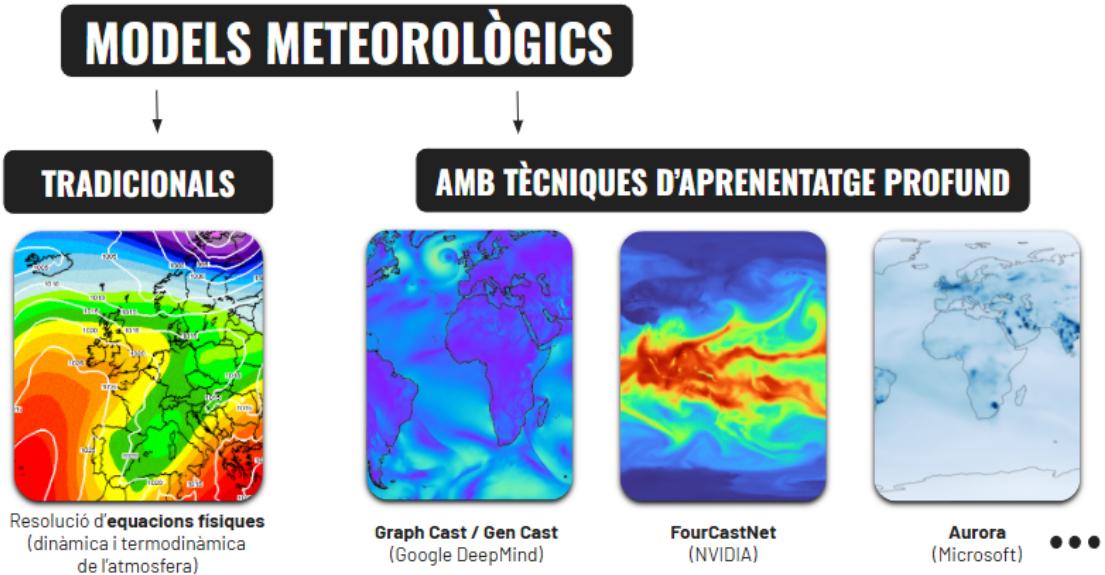


Figura 2: Models meteorològics físics tradicionals vs models recents amb tècniques d'aprenentatge profund.

Paral·lelament, s'està avançant cap a metodologies híbrides que integren l'aprenentatge automàtic amb la simulació física, ja sigui a través de la correcció de sortides dels models físics, l'aprofitament de simulacions sintètiques per entrenar xarxes, o la incorporació de restriccions físiques directament a l'arquitectura del model (com ara els anomenats neural operators o les arquitectures physics-informed). Aquesta combinació permet aprofitar la robustesa dels models físics tradicionals i, alhora, incorporar les capacitats predictives dels mètodes basats en GNNs. D'aquesta manera, s'obtenen sistemes predictius més precisos i amb una millor detecció anticipada de fenòmens extrems. Un exemple seria el model de Google Deep Mind anomenat *NeuralGCM*.

Malgrat els avenços recents, cal tenir present que la majoria de serveis meteorològics nacionals i internacionals continuen utilitzant models físics com a base de la predició operativa. Els models d'aprenentatge profund, tot i haver demostrat una capacitat sorprenent per reproduir i anticipar patrons meteorològics a gran escala, encara presenten desafiaments importants (tot i que molts d'ells ja s'estan intentant resoldre actualment). La integració de dades heterogènies, la gestió de la incertesa (veure *SEEDS* de Google Deep Mind) i la interpretació dels resultats dels models basats en GNNs continuen sent àrees obertes d'investigació. Aquest context posa de manifest la necessitat de seguir explorant noves metodologies que combinin els avantatges dels models numèrics tradicionals amb les possibilitats que ofereixen les tècniques d'aprenentatge profund.

L'estat de la recerca actual, per tant, evidencia un interès creixent per l'aplicació de les GNNs com a eina complementària als models meteorològics tradicionals. La implementació de sistemes híbrids que integrin informació de fonts diverses i abordin de manera efectiva els reptes de la resolució espacial i temporal es presenta com una línia d'investigació favorable per millorar els sistemes predictius.

Dins d'aquest context de canvi accelerat, la recerca en xarxes neuronals en graf aplicades a xarxes d'estacions meteorològiques disperses territorialment, com la que es desenvolupa en aquest treball de final de grau, representa una aportació valiosa i innovadora. Aquesta línia permet abordar la predició meteorològica de forma localitzada i adaptada a la distribució real de sensors, oferint una via complementària i potencialment més eficient per a la millora dels sistemes de predició meteorològica a escala regional.

4 Dades inicials i preparació del dataset

4.1 Les dades

Les dades utilitzades en aquest treball provenen de la secció de meteorologia de 3Cat (la unió de TV3 i Catalunya Ràdio) i recullen informació procedent de diverses fonts fiables dels Països Catalans. Tot i que inicialment es disposava de dades des de mitjans de 2013, s'ha decidit limitar l'anàlisi al període comprès entre el 2016 i el 2024 (Figura 3). Aquesta decisió es basa, entre altres raons, per la manca de la variable de pressió atmosfèrica en els fitxers anteriors al dia 4 de juliol de 2015 a les 7 del matí i en la necessitat de garantir la qualitat de les dades des d'un bon començament.

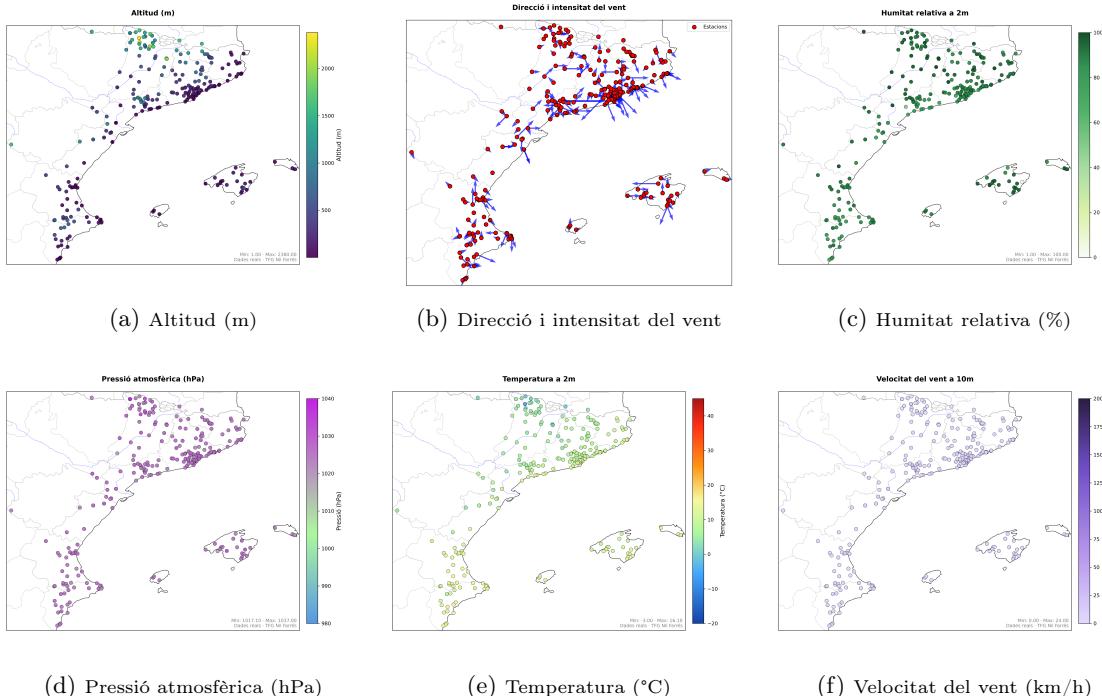


Figura 3: Distribució espacial de les diverses variables meteorològiques als Països Catalans el dia 1 de gener de 2016 a les 00h. Com que en aquell instant no plovia, no hi ha gràfic de precipitació horària acumulada.

A continuació es mostra la matriu de correlació calculada a partir de totes les dades meteorològiques disponibles (veure Figura 4). Aquesta mostra els coeficients de correlació de Pearson entre les principals variables: temperatura, humitat relativa, precipitació, força i direcció del vent, pressió atmosfèrica i altitud.

Si l'analitzem, podem observar-hi el següent:

- La temperatura i la humitat relativa presenten una correlació negativa moderada (-0,40). Aquest patró és típic en meteorologia: a mesura que augmenta la temperatura, la humitat relativa sol disminuir ja que l'aire calent pot contenir més vapor d'aigua abans d'assolir la saturació.
- La temperatura i l'altitud també mostren una correlació negativa (-0,33), fet que reflecteix la disminució progressiva de la temperatura amb l'altitud, una característica ben coneguda de la dinàmica atmosfèrica.
- La pressió atmosfèrica i la temperatura també mantenen una correlació negativa (-0,21). Aquesta relació, si bé és consistent amb la dependència de la pressió respecte a la temperatura i a l'altitud, no permet extreure conclusions físiques directes només a partir d'aquesta anàlisi global.

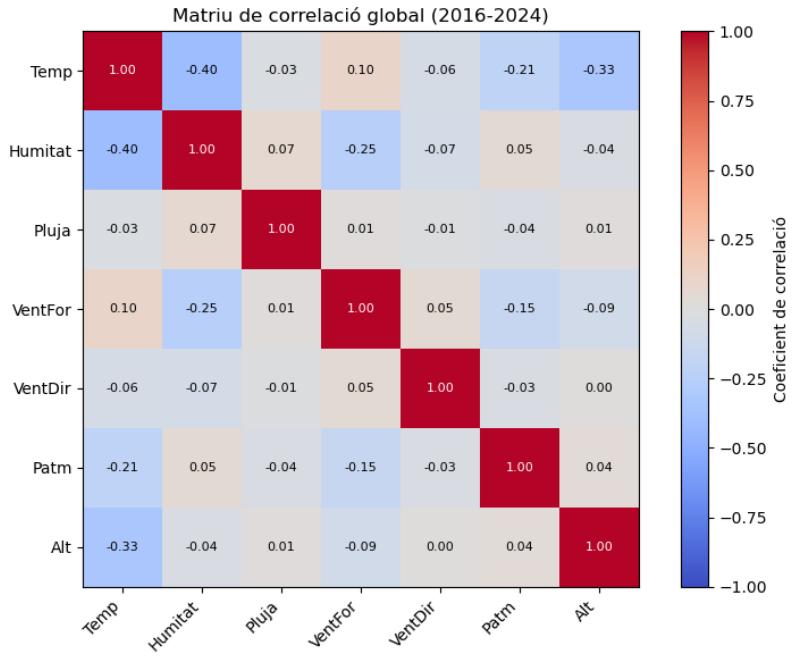


Figura 4: Matriu de correlació calculada a partir de totes les dades meteorològiques disponibles entre 2016 i 2024.

- La precipitació no mostra pràcticament cap correlació amb la resta de variables de manera global, amb valors molt propers a zero. Això s'explica pel seu caràcter altament local i aleatori, ja que la precipitació pot ser molt variable tant en l'espai com en el temps.
- Les variables del vent (força i direcció) presenten valors de correlació baixos amb la resta de variables. En particular, la força del vent manté una correlació lleugerament negativa amb la humitat (-0.25), però cap de les dues variables de vent mostra un patró clar ni una relació estadística forta amb la resta de magnituds meteorològiques, segons aquesta matriu global.

Cal destacar que els valors de correlació, en general, són relativament baixos. Aquesta baixa correlació global és esperable quan s'analitzen dades meteorològiques agregades de tot un territori, estacions i períodes diferents. Els factors climàtics locals, la variabilitat estacional i la diversitat geogràfica del territori tendeixen a diluir les relacions lineals fortes que sí que podrien aparèixer en àmbits més locals, temporals o en situacions meteorològiques concretes. Per tant, per una interpretació física més robusta, caldria una anàlisi específica de casos o estudiar la distribució de correlacions segons la situació meteorològica.

Les dades meteorològiques, geogràfiques i d'identificació de cada estació estan emmagatzemades en fitxers amb format CSV, organitzats en carpetes de manera jeràrquica segons l'any, mes, dia i hora (Figura 5). La nomenclatura de cada fitxer es basa en el patró YYYYMMDDHHdadesPC_utc.csv (per exemple, el fitxer 2023082500dadesPC_utc.csv conté les dades de totes les estacions meteorològiques de fonts fiables dels Països Catalans pel dia 25 d'agost de 2023 a les 00 hores). D'entre tots els fitxers disponibles, només es consideren aquells el nom dels quals acaba amb dadesPC_utc.csv, ja que són els que contenen la informació completa de totes les estacions. Així, per cada hora de cada dia de cada mes, des de 2016 fins a 2024, hi ha un fitxer en aquest format amb totes les dades.

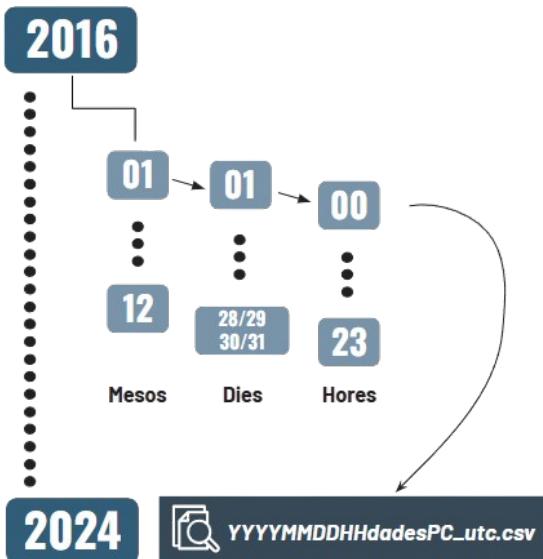


Figura 5: Diagrama de l'estructura jeràrquica d'organització de carpetes i fitxers de dades meteorològiques utilitzats al treball.

D'entre totes les variables que contenen els fitxers originals s'han seleccionat les més imprescindibles per a l'anàlisi. Aquestes són:

- **id**: identificador de l'estació meteorològica.
- **Font**: nom de la font d'informació fiable.
- **Temp**: temperatura (°C).
- **Humitat**: humitat relativa (%).
- **Pluja**: precipitació acumulada diària (mm).
- **VentDir**: direcció del vent (NO, N, NE, E, SE, S, SO, O).
- **VentFor**: velocitat del vent (km/h).
- **Patm**: pressió atmosfèrica (hPa).
- **Alt**: altitud (m).
- **lat i lon**: coordenades geogràfiques.

4.1.1 Selecció de fonts fiables

El conjunt de dades conté dades de diverses fonts meteorològiques. És per això que s'han hagut d'avaluar aquestes fonts, conjuntament amb la secció de meteorologia de 3Cat, per tal de seleccionar únicament aquelles considerades prou fiables. Són les següents:

- **METEOCAT** (Servei Meteorològic de Catalunya): encarregat de gestionar els sistemes d'observació i predicció meteorològics a Catalunya.
- **AEMET** (Agència Estatal de Meteorologia): encarregada de prestar els serveis meteorològics que siguin competència de l'Estat.
- **Vallsdaneu**: meteorologia de les Valls d'Àneu.
- **SAIH** (Sistema Automàtic d'Informació Hidrològica).
- **Avamet** (Associació Valenciana de Meteorologia).
- **MeteoPirineus**: previsions del temps específiques pel Pirineu.
- **Meteoclimatic**: xarxa d'estacions meteorològiques automàtiques distribuïdes per la Península Ibèrica, els dos arxipèlags, Ceuta i Melilla, desitjant oferir informació de qualitat.

- **Meteoprades**: estudi del clima de Muntanyes de Prades i la seva relació amb el medi natural i rural.
- **WLINK_DAVIS**: xarxa d'estacions meteorològiques Davis connectades a la plataforma WeatherLink, que permet consultar dades meteorològiques en temps real de nombroses estacions particulars arreu del territori.

Cal destacar que les dades d'algunes estacions de les fonts *MeteoPirineus*, *Meteoprades*, *SAIH* i *AEMET* presenten alguns errors de precisió en els darrers anys per problemes de funcionament o desconnexió, fet que cal tenir-ho en compte en l'anàlisi dels resultats.

4.1.2 Anomalies en les dades

En la fase inicial d'anàlisi s'ha observat que les dades originals contenen diversos fitxers problemàtics: alguns es troben completament buits i altres no incloïen la capçalera amb els noms de les variables meteorològiques. Per localitzar aquests fitxers de manera sistemàtica, s'han desenvolupat dos scripts específics (veure Annex):

- **fitxers_buits.py**: examina tots els fitxers CSV (directori DADES_METEO_PC) per identificar aquells que estan completament buits o que són illegibles. D'aquesta manera es pot obtenir una llista dels fitxers conflictius.
- **sense_nomcol.py**: comprova la primera línia de cada fitxer per verificar si coincideix amb la capçalera coneuguda. Així s'han identificat els fitxers que no contenen la capçalera amb els noms de les variables.

Un cop s'han conegit els fitxers que presenten aquests problemes, s'ha contactat novament amb 3Cat. En col·laboració conjunta, s'han localitzar aquests mateixos fitxers correctes en una còpia de seguretat (backup), garantint així la integritat i la qualitat de les dades per al preprocessament posterior.

Per altra banda, altres problemes detectats en les dades originals són els següents:

- **Valors NaN**: diversos fitxers contenen estacions amb certes variables que no contenen valors. Per tractar aquests casos, s'ha aplicat un procediment d'imputació mitjançant interpolació amb dades de fitxers adjacents (fins a 8 hores abans i també 8 hores després o, en cas que encara no s'hagi trobat el valor d'aquesta variable, amb el fitxer de la mateixa hora del dia anterior/posterior). Quan aquesta interpolació no és possible, es procedeix a eliminar la fila (estació meteorològica) corresponent. La fila també s'elimina directament si el valor d'alguna de les variables de posició (latitud, longitud o altitud) no hi és present.
- **Tractament de la pluja acumulada**: les dades originals proporcionen la pluja acumulada diària (això vol dir que si a les dues de la matinada han caigut 3mm de pluja i no plou més, aquesta dada s'arrosga fins al final del dia), però per a un entrenament efectiu del model final es necessita la pluja acumulada per hora. Per solucionar-ho, es calcula la diferència entre el valor acumulat actual i el del fitxer de l'hora anterior per a cada estació, assumint un valor previ de 0 per a la primera hora del dia. A més, els valors NaN en la columna Pluja s'han tractat com a 0, suposant que no ha plogut (0 mm).

Aquesta part de preprocessament de les dades és la que ha suposat un major temps de desenvolupament i anàlisi. S'ha dut a terme mitjançant un fitxer anomenat **prep.py** (present a l'Annex), assegurant així que les dades utilitzades per a la construcció i entrenament del model compleixin els requisits de qualitat i fiabilitat necessaris. Tot seguit l'expliquem més detalladament.

4.2 Preprocessament de les dades originals

El principal objectiu del fitxer `prep.py` és realitzar un pre-processament incial de les dades meteorològiques originals (veure Annex). Es vol passar dels fitxers csv originals a fitxers csv preprocessats (Figura 6).

```
"nu","id","Font","Data","Població","lat","lon","Temp","Temp_Max","Temp_Min","Amplitud_Termica","VentDir","VentFor","VentForX","SimbolsEnv","Windchill","Humitat","Pluja","Alt","Patm","Webcams","NomOK","Comarca"
"1","ESCAT0800000008328A","Meteoclimatic","01/01/2016 01:32:00h","Abrera - Can Villalba 175m",41.523393,1.92644,8.5,18.4,8.5,1.9,,6,simbols/.png,8.5,0,10.5,,175,1024,No,Abrera - Can Villalba 175m,Belx Llobregat
"2","ESCAT0800000008328B","Meteoclimatic","01/01/2016 01:38:00h","Alegre - 28m",41.586389,2.278333,11.2,11.3,11.2,,1,E,18,19,simbols/.png,11.2,86,11.2,,289,1025,No,Alella 28m,Maresme
"3","ESCAT0800000008328C","Meteoclimatic","01/01/2016 01:44:00h","Arenys de Mar - 85m",41.592778,2.555,11.4,12.1,11.3,,1,F,18,19,simbols/.png,11.4,87,11.4,,179,1025,No,Arenys de Mar - 85m,Maresme
"4","ESCAT0800000008328D","Meteoclimatic","01/01/2016 01:50:00h","Arenys de Mar - 85m",41.592778,2.555,11.4,12.1,11.3,,1,F,18,19,simbols/.png,11.4,87,11.4,,179,1025,No,Arenys de Mar - 85m,Maresme
"5","ESCAT0800000008328E","Meteoclimatic","01/01/2016 01:56:00h","Arenys de Mar - Riera - 42m",41.594167,2.54667,11.4,12.1,11.3,,2,243611,12.1,12.1,11.9,,2,Nw,,6,simbols/.png,12.1,89,12.1,,79,1025,No,Badalona - Bufals 70m
"6","ESCAT0800000008328F","Meteoclimatic","01/01/2016 01:35:00h","Badalona - Bufals 70m",70m,41.44,2.2175,10.6,11.4,18.6,,8,Nw,,18,simbols/.png,10.6,86,10.6,,69,1025,No,Badalona - Llefià 60m
"7","ESCAT0800000008328G","Meteoclimatic","01/01/2016 01:32:00h","Badalona - Llefià&grave; ",45m,,41.46,2.218056,10.4,11.2,18.4,,8,Wn,,18,simbols/.png,10.4,85,10.4,,45,1028,No,Badalona - Llefià &grave;;
```

(a) Format original de la taula de dades

```
Id,Font,Temp,Humitat,Pluja,Alt,VentDir,VentFor,Patm,lat,lon
ESCAT0800000008328A,Meteoclimatic,11.2,86.0,0.0,280,90.0,10.0,1025.0,41.506389,2.278333
ESCAT0800000008328B,Meteoclimatic,11.1,79.0,0.0,119,225.0,3.0,1027.0,41.496944,2.295278
ESCAT0800000008350A,Meteoclimatic,11.1,87.0,0.0,85,337.5,5.0,1023.0,41.592778,2.565
ESCAT0800000008913B,Meteoclimatic,10.6,86.0,0.0,60,315.0,10.0,1025.0,41.44,2.2175
ESCAT0800000008913C,Meteoclimatic,10.4,85.0,0.0,45,292.5,10.0,1028.0,41.46,2.218056
ESCAT0800000008912B,Meteoclimatic,12.3,80.0,0.0,25,292.5,2.0,1026.0,41.445,2.243333
ESCAT0800000008695A,Meteoclimatic,6.4,85.0,0.0,865,90.0,3.2222222222222223,1025.0,42.257778,1.860278
ESCAT0800000008018D,Meteoclimatic,13.1,69.0,0.0,12,337.5,2.0,1024.0,41.400556,2.193611
ESCAT0800000008014C,Meteoclimatic,13.2,75.0,0.0,61,45.0,2.0,1026.0,41.384167,2.133056
ESCAT0800000008035A,Meteoclimatic,10.9,86.0,0.0,280,337.5,4.153846153846153,1025.0,41.420833,2.1175
```

(b) Format preprocessat de la taula de dades

Figura 6: Comparativa entre un fitxer csv original i un de preprocessat amb `prep.py`.

4.2.1 Funcions auxiliars

En el codi s'hi defineixen diverses funcions que encapsulen operacions comunes:

- `get_file_path_for_timestamp`: construeix el camí complet d'un fitxer CSV a partir del directori arrel i d'un *timestamp*. Utilitza el format `YYYYMMDDHHdadesPC_utc.csv`.
- `load_file`: carrega un fitxer CSV provant primer amb la codificació `utf-8` i, en cas d'error, amb `latin-1`. També implementa una *cache* per evitar recàrregues innecessàries en el procés d'interpolació.
- `get_station_value`: obté el valor numèric d'una variable per a una estació determinada. Filtra les dades per mantenir només les fonts fiables i converteix els valors a format numèric.
- `get_neighbor_value`: cerca, en un marge de 8 hores (cap enrere i cap endavant en el temps), el valor més proper d'una variable per a una estació concreta. L'objectiu és, amb la funció `interpolate_value`, poder substituir el valor `Nan` present mitjançant interpolació. Si no troba cap valor per aquesta estació en els fitxers d'aquest interval de temps (tant endavant com endarrere), s'intenta amb la mateixa hora del dia anterior o posterior. Cal dir que, si després de fer tots aquests passos no ha trobat cap valor, l'estació meteorològica del fitxer que s'està preprocessant s'eliminarà.
- `interpolate_value`: utilitza els valors veïns (obtinguts amb `get_neighbor_value`) per interpolar linealment el valor d'una variable quan aquest és `Nan`. En aquesta operació s'utilitza Cupy per realitzar els càlculs a la GPU, millorant el rendiment.

4.2.2 Processament i paral·leització

La funció `preprocess_csv` és el nucli del codi i executa els passos següents:

1. **Lectura del fitxer:** es carrega el fitxer CSV, provant les codificacions `utf-8` i `latin-1`. Si el fitxer és buit o no conté columnes, es retorna un DataFrame buit.
2. **Filtrat i extracció del timestamp:** es filtra el DataFrame per conservar només les fonts fiables. A més, s'extreu el *timestamp* a partir del nom del fitxer mitjançant expressions regulars, validant que l'hora sigui correcta (entre 0 i 23) ja que a les dades originals hi ha una carpeta anomenada 25 que fa referència a un resum de dades del dia que no utilitzarem.

3. **Selecció i conversió de columnes:** es seleccionen les columnes d'interès (com per exemple `id`, `Font`, `Temp`, etc.) i es converteixen a format numèric.
4. **Càlcul de la pluja:** es calcula la precipitació horària com la diferència entre el valor actual i el valor de l'hora anterior. En aquest càlcul s'utilitza `Cupy` per realitzar la resta de manera accelerada a la GPU.
5. **Interpolació dels valors faltants:** per a les variables com `Temp`, `Humitat`, `VentDir` i `VentFor` (i `Patm` si s'aplica ja que recordem que aquesta variable no apareix fins a mitjans de 2015), si hi ha valors nuls, es realitza una interpolació utilitzant els valors dels fitxers adjacents.
6. **Depuració i ajust de rangs:** es descarten les files on persisteixen valors nuls en les variables clau i es limiten els rangs de certes variables per garantir una coherència amb la meteorologia dels Països Catalans (per exemple, els valors de la variable de la velocitat del vent s'han limitat entre els 0 i els 200 km/h en vents de tramuntana huracanats com a màxim per Catalunya, els valors de la variable d'humitat s'han de trobar entre 0 i 100%, etc.).

La funció `process_all_csvs_parallel` recorre l'estructura de directoris per identificar els fitxers CSV a processar i aplica els següents passos:

- Es descarten directoris que contenen subcadenes no rellevants (per exemple, fitxers d'errors o arxius no pertinents).
- Es compila una llista amb els camins dels fitxers que compleixen els criteris (nom dels fitxers que acaba amb `dadesPC_utc.csv` i que corresponen als anys d'interès).
- S'utilitza `ProcessPoolExecutor` per enviar, en paral·lel, la tasca `process_file` per a cada fitxer, aprofitant diversos nuclis de la CPU.
- La barra de progrés proporcionada per `tqdm` informa de l'estat del processament en temps real.

Finalment, el bloc principal (dins del `if __name__ == "__main__":`) defineix els directoris d'entrada (`root_directory`) i de sortida (`processed_directory`) i crida la funció `process_all_csvs_parallel` amb un nombre màxim de workers (per exemple, 8). Un cop finalitzat el processament, es registra i s'imprimeix un missatge indicant l'acabament del procés.

4.3 Conversió de format: de CSV a objectes Data

Un cop completat el preprocessament dels fitxers amb les dades meteorològiques inicials, el següent pas és convertir aquestes dades a un format adequat per entrenar, més endavant, un model meteorològic basat en xarxes neuronals en graf (GNN). Concretament, es vol passar dels fitxers preprocessats en format CSV a objectes `Data` compatibles amb la llibreria `torch.geometric`², estructurant la informació en forma de grafs dinàmics amb nodes, arestes i atributs d'acord amb la realitat física de les observacions meteorològiques (Figura 7). En aquest cas en concret, quan parlem de grafs dinàmics ens referim a que els nodes i les arestes poden variar a cada pas horari. Aquest fet és degut a diversos factors: hi ha hagut una averia en una estació, el programa informàtic que emmagatzema les dades ha fallat, s'han desabilitat estacions que ja no són funcionals, etc.

Cada node representa una estació meteorològica i les seves mesures puntuals en una hora determinada. Cada fitxer CSV correspon a una instantània horària de totes les estacions disponibles, de manera que es genera un graf per a cada hora entre 2016 i 2024. Durant aquesta conversió també s'assegura una normalització coherent de les característiques dels nodes, pas fonamental per evitar biaixos numèrics i afavorir un aprenentatge robust per part del model.

Per realitzar aquesta conversió es fan servir dos scripts de Python: `compute_PC_norm_params.py` i `toData.py`, els quals s'expliquen amb més detall a continuació.

²<https://pytorch-geometric.readthedocs.io/en/latest/>

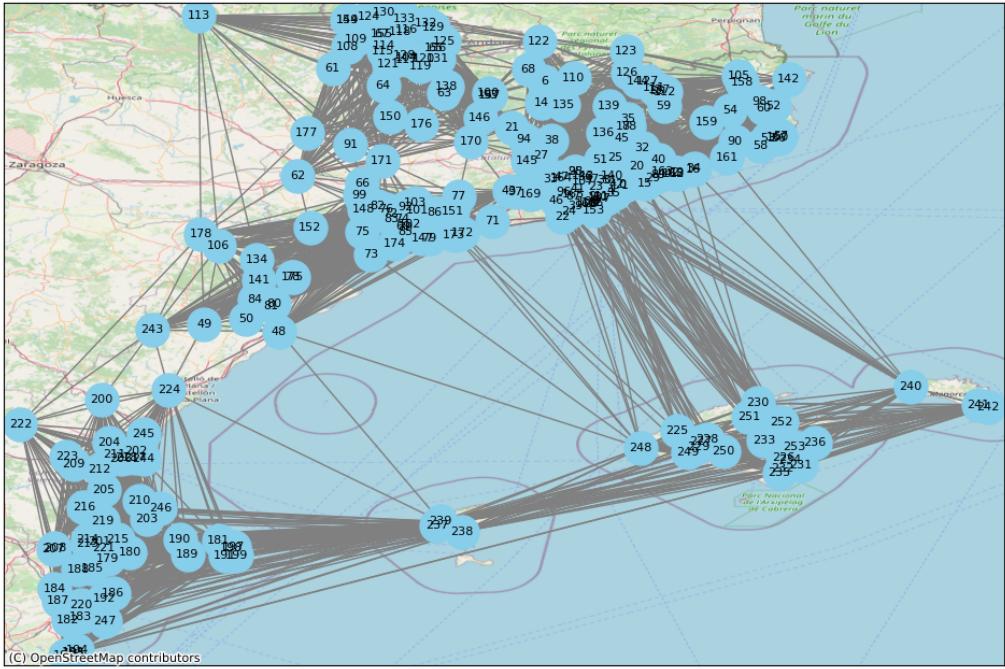


Figura 7: Representació de la xarxa de connexions entre les estacions meteorològiques sobre el mapa dels Països Catalans pel dia 1 de gener de 2016 a les 00h.

4.3.1 Paràmetres de normalització pels Països Catalans

El primer pas per garantir que el model pugui generalitzar correctament sobre totes les estacions meteorològiques dels Països Catalans és normalitzar les variables d'entrada. Aquesta normalització porta totes les característiques a una escala comparable, evitant que variables amb magnituds molt diferents (per exemple, la temperatura respecte a la pluja o la humitat) dominin l'entrenament o provoquin biaixos numèrics.

Per aconseguir-ho, l'script `compute_PC_norm_params.py` recorre tots els fitxers CSV ja preprocessats i calcula la mitjana i la desviació estàndard globals de cada variable només pels anys d'entrenament del model (de 2016 fins a 2022 ja que 2023 s'utilitza per fer validació i 2024 per fer test). Abans de calcular les estadístiques, aplica exactament les mateixes transformacions que després s'executaran durant la generació dels grafs amb `toData.py`: aquestes transformacions inclouen la codificació temporal cíclica, el càlcul de l'angle zenithal solar, la temperatura potencial, el punt de rosada i la conversió de la direcció del vent a components vectorials.

D'aquesta manera, es garanteix que la normalització és coherent i aplicable tant en fase d'entrenament com en fase d'inferència. Els paràmetres de normalització resultants es desen en un fitxer JSON anomenat `PC_norm_params.json`, que facilita la seva posterior aplicació als objectes `Data` de `torch_geometric`.

Codificació temporal cíclica

En meteorologia, molts processos atmosfèrics (com els cicles tèrmics diürns o estacionals) depenen del moment del dia i de l'època de l'any, i presenten una naturalesa clarament cíclica. Per capturar correctament aquestes dinàmiques, és fonamental que el model entengui que les hores i els dies de l'any formen cercles continus i no intervals lineals.

Per evitar discontinuïtats artificials entre les 23 h i les 0 h, o entre el 31 de desembre i l'1 de gener, es codifiquen les hores i els dies sobre un cercle unitari:

$$\begin{aligned} (\text{hora_sin}, \text{hora_cos}) &= \left(\sin\left(\frac{2\pi h}{24}\right), \cos\left(\frac{2\pi h}{24}\right) \right) \\ (\text{dia_sin}, \text{dia_cos}) &= \left(\sin\left(\frac{2\pi(d-1)}{N}\right), \cos\left(\frac{2\pi(d-1)}{N}\right) \right) \end{aligned} \quad (1)$$

on h és l' hora en UTC i d és el dia julià (sent $N = 365$ o 366 segons si l'any és de traspàs o no).

Aquesta representació preserva la distància angular real entre instants propers, permetent al model capturar de manera suau els cicles diürns i estacionals sense introduir errors artificials deguts a la numeració discreta del temps.

Angle zenithal solar

La radiació solar és la font d'energia primària que impulsa la majoria dels processos meteorològics, ja sigui de manera directa (a través de l'escalfament superficial) o indirecta (afavorint gradients de pressió, convecció atmosfèrica i cicles d'evaporació-condensació). Per tal que el model pugui captar aquesta variabilitat diürna i estacional de la lluminació, es calcula el cosinus de l'angle zenithal solar, $\cos \theta_{sza}$:

$$\cos \theta_{sza} = \sin \varphi \sin \delta + \cos \varphi \cos \delta \cos \text{HRA}, \quad (2)$$

on φ és la latitud de l'estació, δ la declinació solar i $\text{HRA} = 15^\circ(t_{\text{solar}} - 12)$ l'angle horari. La declinació solar s'aproxima amb la relació $\delta = 23.44^\circ \sin \frac{2\pi(284+d)}{365}$, sent d el dia julià de l'any, i l' hora solar local t_{solar} s'obté corregint el temps en UTC amb $\lambda/15$ hores (λ és la longitud).

S'utilitza el cosinus de l'angle, i no pas l'angle mateix, perquè $\cos \theta_{sza}$ és proporcional al flux de radiació solar incident sobre una superfície horitzontal. Aquesta informació farà que el model pugui detectar els cicles d'escalfament superficial, de refrigeració nocturna i d'estratificació atmosfèrica, processos fonamentals en la formació de boires, inversions tèrmiques i l'evolució de la convecció.

Temperatura potencial

La temperatura potencial θ s'utilitza per comparar l'estat tèrmic de masses d'aire que es troben a pressions diferents. Es defineix com:

$$\theta = T \left(\frac{P_0}{P} \right)^{R/c_p} \quad \text{amb} \quad \frac{R}{c_p} \simeq 0.286, \quad (3)$$

on T és la temperatura absoluta (en Kelvin), P és la pressió atmosfèrica local, i $P_0 = 1013$ hPa és una pressió de referència estàndard.

A diferència de la temperatura real, la temperatura potencial és conservativa en processos adiabàtics secs: és a dir, si una parcel·la d'aire s'eleva o descendeix sense intercanvi de calor amb l'entorn, la seva θ es manté constant. Això la converteix en un traçador natural per estudiar l'estabilitat de l'atmosfera, la presència d'inversions tèrmiques i la tendència a la convecció.

Incloure θ com a característica d'entrada permet al model inferir propietats importants de l'estratificació vertical de manera implícita, sense haver d'explícitar l'equació termodinàmica completa.

Punt de rosada

El punt de rosada T_d és la temperatura a la qual caldria refredar una parcel·la d'aire, a pressió constant, perquè arribés a la saturació (humitat relativa del 100 %). Aquest paràmetre resumeix de manera compacta la combinació no lineal entre temperatura i humitat relativa.

Es calcula a partir de la fórmula empírica de Magnus:

$$\alpha = \ln \left(\frac{H}{100} \right) + \frac{a T_c}{b + T_c}, \quad T_d = \frac{b \alpha}{a - \alpha}, \quad (4)$$

on T_c és la temperatura en graus Celsius ($T_c = T - 273.15$), H és la humitat relativa (en %), i els paràmetres $a = 17.27$ i $b = 237.7$ s'ajusten per a condicions típiques de temperatura moderada.

Incorporar T_d com a característica d'entrada permet al model detectar amb més facilitat situacions properes a la saturació, com la formació de boires, rosades o el potencial de precipitació convectiva, sense haver de deduir explícitament la relació complexa entre temperatura i humitat.

Components del vent

La representació del vent és crítica per capturar els fluxos horitzontals d'aire i les adveccions de propietats atmosfèriques. En meteorologia, la direcció del vent θ es defineix segons la convenció "d'on bufa", és a dir, l'angle mesurat des del nord geogràfic en sentit horari.

Per transformar aquesta informació angular en components cartesianes útils per a xarxes neuronals, es calcula:

$$u = -V \sin \theta, \quad v = -V \cos \theta, \quad (5)$$

on V és la intensitat del vent, u és la component zonal (positiva cap a l'est) i v la component meridional (positiva cap al nord).

Aquesta representació vectorial elimina les discontinuïtats artificials associades a l'angle (per exemple, entre 0° i 360°) i facilita l'aprenentatge de patrons d'advecció i de transport horitzontal per part del model.

A més, es conserven les codificacions $\sin \theta$ i $\cos \theta$ originals, aportant una redundància angular útil: d'aquesta manera, el model pot combinar tant la representació vectorial com la cíclica per optimitzar la inferència de patrons de vent locals i regionals.

Després d'aplicar totes les transformacions derivades a cada fitxer, l'script concatena totes les files en un únic conjunt de dades global. Això permet calcular de manera robusta la mitjana (μ) i la desviació estàndard (σ) per a cada variable, assegurant que la normalització sigui coherent a escala dels Països Catalans i de totes les hores considerades.

Els paràmetres resultants es desen en un fitxer JSON anomenat `PC_norm_params.json`, que es farà servir posteriorment durant la generació dels grafs per aplicar la mateixa normalització de forma consistent entre l'entrenament i la inferència.

A més, de manera opcional, l'script pot generar un conjunt d'histogrames per a cada variable normalitzada. Aquesta inspecció visual ajuda a detectar distribucions anòmals, asimetries marcades o valors atípics que podrien afectar el rendiment del model si no es tractessin adequadament.

4.3.2 Generació dels grafs horaris

L'script `toData.py` converteix cada fitxer horari de dades meteorològiques en un objecte `Data` de `torch_geometric`, adequat per a l'entrenament de xarxes neuronals en graf.

Cada objecte conté:

- **Nodes:** corresponents a les estacions meteorològiques actives a aquella hora, amb les seves característiques físiques i meteorològiques corresponents.
- **Arestes:** representen relacions espacials entre estacions basades en la seva posició geogràfica i altitud, i codifiquen gradients o diferències de variables meteorològiques.
- **Atributs d'aresta,** que inclouen distàncies, pendents, diferències meteorològiques i angles relatius.
- **Metadades:** com el grau mitjà del graf o el radi efectiu mitjà de connexió.

Aquests grafs horaris permeten modelar no només les condicions locals de cada estació, sinó també la seva interacció amb l'entorn, fet que és clau per capturar fenòmens meteorològics que depenen tant de factors locals com regionals.

Construcció dels vectors de característiques

Cada estació meteorològica passa primer per totes les transformacions descrites a la secció anterior, incloent-hi la codificació temporal cíclica, el càlcul de l'angle zenithal solar, la temperatura potencial, el punt de rosada i la codificació del vent.

A més, durant la construcció dels vectors de característiques s'apliquen els passos addicionals següents:

- **Normalització de l'altitud:** es normalitza l'altitud respecte d'una mitjana $\mu_z = 454.3$ m i una desviació típica $\sigma_z = 175.6$ m, de manera que la variable `Alt_norm` té valors centrats i comparables. Aquesta informació és clau per captar inversions tèrmiques, convecció orogràfica o distribució de precipitacions.
- **Transformació logarítmica de la pluja:** s'aplica una funció $\log(1 + \text{Pluja})$ a la precipitació acumulada per hora, reduint l'asimetria extrema entre hores sense pluja o episodis intensos i evitant problemes associats a valors nuls (ja que $\log(1 + 0) = 0$). Això facilita l'aprenentatge del model en prediccions relacionades amb la pluja.
- **Normalització consistent:** utilitzant els paràmetres de mitjana i desviació estàndard prèviament calculats amb `compute_PC_norm_params.py`, es normalitzen totes les variables de manera coherent entre hores, dies i anys diferents. Això permet que el model operi amb escales estables i millori la seva capacitat de generalització davant condicions meteorològiques diverses.

Posició dels nodes

Per defecte, cada node guarda la seva posició geogràfica (φ, λ, z) , on φ és la latitud, λ la longitud i z l'altitud sobre el nivell del mar.

Tanmateix, quan cal calcular distàncies físiques entre estacions, aquestes coordenades es transforment a un sistema de coordenades cartesianes tridimensionals assumint un radi terrestre mitjà $R = 6371$ km. Aquesta conversió permet calcular distàncies euclidianes de manera més precisa:

$$(x, y, z) = R (\cos \varphi \cos \lambda, \cos \varphi \sin \lambda, \sin \varphi)$$

El pas a coordenades cartesianes evita problemes de convergència de meridians a altes latituds i assegura que les distàncies entre estacions siguin proporcionals a la seva separació física real. Aquesta transformació és fonamental per a la construcció correcta dels veïnatges i pel càlcul d'atributs d'aresta basats en la distància.

Estructures de veïnatge

Per construir el graf de connexions entre estacions meteorològiques, es combinen diverses estratègies complementàries que garanteixen una estructura robusta i físicament raonable:

1. **Backbone Delaunay 2-D:** es construeix una triangulació de Delaunay sobre les coordenades $(\lambda \cos \varphi, \varphi)$, on λ és la longitud i φ la latitud. Aquesta transformació simula una projecció conforme, minimitzant distorsions a latituds moderades. El resultat és una grella planar i natural que assegura veïnatges físicament coherents sense arestes creuades.
2. **Radi adaptatiu local basat en KNN:** per cada node i , es calcula la distància al seu k -èssim veí més proper ($d_{k(i)}$, amb $k = 4$) utilitzant el concepte de veïns més propers (KNN). A partir d'aquesta distància, es defineix un radi adaptatiu $r_i = \kappa d_{k(i)}$ amb $\kappa = 1.3$, i es connecten totes les estacions situades dins d'aquest radi. Aquest mètode permet capturar la variabilitat local de densitat d'estacions, evitant la sobreconnexió en zones d'alta densitat i la manca de connexions en zones aïllades.
3. **Multiescala:** s'afegeixen connexions supplementàries entre estacions separades per una distància inferior al quantil q de la distribució global de distàncies (normalment $q = 0.65$). Aquesta estratègia permet al graf capturar teleconnexions regionals i fluxos de llarga distància, rellevants en la dinàmica atmosfèrica.

A més, per millorar la fidelitat física dels grafs:

- Es filtren totes les arestes entre nodes amb una diferència d'altitud superior a 150 metres ($\Delta z > 150$ m), evitant connexions irrealistes entre vessants separats per barreres orogràfiques.
- S'assegura que cap node quedí aïllat: si un node no té veïns després del filtratge, es connecta al veí més proper que compleixi els criteris, dins d'un radi màxim de 80 km.

Aquestes estratègies permeten obtenir grafs meteorològics que representen tant l'estructura local del relleu com la connectivitat regional, facilitant que el model aprengui patrons meteorològics a múltiples escales.

Atributs d'aresta

Per a cada aresta dirigida ($i \rightarrow j$), es concatena un vector d'atributs que codifica la relació física entre les dues estacions. Aquesta informació complementa la topologia del graf i permet al model entendre com varien les condicions meteorològiques entre localitzacions veïnes. Els atributs concatenats són:

1. **Distància normalitzada** $\frac{d_{ij}}{d_0}$: la distància euclidiana entre i i j , escalada per una distància de referència $d_0 = 100$ km. Permet quantificar la separació espacial entre nodes.
2. **Diferències de variables escalars**: $x_i - x_j$ en temperatura (Temp), humitat relativa (Humitat), pluja transformada (Pluja), velocitat del vent (VentFor) i pressió anòmala (Patm). Aquestes diferències proporcionen informació sobre els gradients atmosfèrics locals.
3. **Desnivell i pendent**: es calcula el desnivell absolut $|\Delta z|$ i el pendent $|\Delta z|/d_{ij}$ entre les estacions. Aquestes magnituds ajuden a capturar efectes orogràfics, com ara canals de vent o gradients de temperatura associats al relleu.
4. **Diferències en variables angulars i termo-dinàmiques**: es concatenen les diferències entre i i j en la codificació sin / cos de la direcció del vent (VentDir_sin, VentDir_cos), així com en el punt de rosada (DewPoint) i la temperatura potencial (PotentialTemp). Aquestes variables reflecteixen canvis en la direcció del flux horitzontal i en l'estabilitat atmosfèrica.
5. **Rumb entre nodes**: es calcula el rumb (bearing) entre les coordenades geogràfiques de i i j , és a dir, l'angle que forma la línia $i \rightarrow j$ amb el nord geogràfic, mesurat en sentit horari. Per evitar discontinuitats a $0^\circ/360^\circ$, es codifica aquest rumb mitjançant els seus valors sin i cos. Aquesta informació proporciona al model una representació contínua de la direcció relativa entre estacions, facilitant l'aprenentatge de patrons direccionals associats a vents predominants, efectes orogràfics o adveccions meteorològiques.
6. **Pes d'esmorteïment exponencial** $w_{ij} = \exp(-d_{ij}/L)$ amb $L = 75$ km: aquest pes redueix progressivament la importància de les connexions més llunyanes, reflectint que en meteorologia les interaccions locals són sovint més rellevants que les remotes.

Aquest conjunt d'atributs permet al model captar tant gradients locals de variables meteorològiques com la separació física i la direcció entre estacions, afavorint una representació rica i físicament fonamentada de les interaccions atmosfèriques.

Post-processament

Abans de desar els grafs, es realitzen diverses operacions de neteja i optimització:

- **Eliminació de bucles propis**: s'eliminen les arestes que connecten un node amb ell mateix, atès que no aporten informació rellevant en el context de la modelització i podrien introduir biaixos no desitjats durant l'entrenament.
- **Coalesce**: es fa servir la funció `coalesce` per agrupar arestes duplicades entre una mateixa parella de nodes, combinant-ne els atributs si s'escau. Això simplifica l'estructura del graf i assegura que no hi hagi més d'una aresta directa entre dos nodes en el mateix sentit.
- **Escriptura del fitxer**: cada graf horari es desa com un fitxer binari amb extensió `.pt` mitjançant `torch.save` i preservant:
 - L'índex d'estació, la font de les dades i la data-hora corresponent.
 - Els paràmetres de normalització aplicats, per garantir coherència en futurs processos d'inferència.
 - Metadades del graf, com ara el grau mitjà (nombre mitjà d'arestes per node) i el radi efectiu (distància mitjana de connexió).

Aquest format binari permet carregar de manera eficient cada graf com un objecte `Data` de `torch_geometric` directament a memòria, preparat per ser utilitzat en processos d'entrenament o validació.

4.3.3 Resum de l'estructura del graf

Nodes

Cada node del graf representa l'estat d'una estació meteorològica en una hora concreta. Les característiques que s'assignen a cada node poden agrupar-se en tres categories principals:

- **Variables meteorològiques observades:** temperatura T (K), humitat relativa H (expressada entre 0 i 1), pluja transformada com a $\log(1 + \text{mm})$, velocitat del vent V (m s^{-1}), pressió anòmala ΔP (hPa) (definida com la desviació de la pressió atmosfèrica respecte d'una pressió de referència estàndard de 1013 hPa) i altitud normalitzada centrant-la a la mitjana i escalant-la amb la desviació estàndard d'altitud de totes les estacions meteorològiques del domini.
- **Variables derivades:** punt de rosada T_d (K) i temperatura potencial θ (K), que encapsulen informació termodinàmica essencial per entendre processos de condensació i estabilitat atmosfèrica.
- **Codificacions cícliques i direccionals:** codificació sin / cos de la direcció del vent, de les hores i els dies de cada any, així com el cosinus de l'angle zenithal solar $\cos\theta_{\text{sza}}$. També s'inclouen els components zonal (u) i meridional (v) del vent, derivats de la intensitat i la direcció.

Arestes

Cada connexió entre dues estacions meteorològiques es representa amb dues arestes dirigides, una en cada sentit ($i \rightarrow j$) i ($j \rightarrow i$), de manera que el graf global és no dirigit. Cada aresta transporta informació sobre la relació espacial i meteorològica entre les dues estacions connectades.

Per defecte, cada aresta conté 14 atributs, que poden arribar a ser 15 si s'inclou el pes exponencial d'esmorteïment. Els atributs concatenats són:

1. **Distància normalitzada** d_{ij}/d_0 : distància física entre i i j escalada per una distància de referència $d_0 = 100$ km, que proporciona una mesura relativa de la separació espacial.
2. **Diferències signades en variables meteorològiques:** diferències $x_i - x_j$ en temperatura (T), humitat relativa (H), pluja transformada, velocitat del vent (V) i pressió anòmala (ΔP). Aquestes diferències ajuden a capturar gradients atmosfèrics horizontals.
3. **Desnivell absolut i pendent:** mòdul de la diferència d'altitud $|\Delta z|$ i pendent $|\Delta z|/d_{ij}$ entre estacions. Reflecteixen l'impacte del relleu en les condicions meteorològiques locals.
4. **Diferències en variables angulars i termo-dinàmiques:** diferències entre i i j en la codificació sin / cos de la direcció del vent, així com en el punt de rosada (T_d) i la temperatura potencial (θ). Permeten capturar canvis en la direcció del flux i en l'estabilitat atmosfèrica de l'aire.
5. **Rumb entre nodes:** sin i cos del bearing (rumb) de i cap a j , codificant de manera contínua la direcció geogràfica relativa entre estacions.
6. **Pes exponencial d'esmorteïment** $w_{ij} = \exp(-d_{ij}/L)$ amb $L = 75$ km (opcional): redueix progressivament la influència de connexions més llunyanas, prioritant interaccions locals que són habitualment més rellevants en la dinàmica atmosfèrica.

En conjunt, el graf resultant capta tant la **topologia espacial** —és a dir, com es relacionen físicament les estacions meteorològiques entre si— com la **dinàmica atmosfèrica** codificada en les seves variables mesurades i derivades.

Aquesta estructura rica permet representar fenòmens meteorològics a múltiples escales espacials: des de la **microescala** (efectes locals de relleu i gradients tèrmics) fins a la **macroescala** (adveccions regionals i teleconnexions). Cada graf correspon a una instantània horària, cobrint el període comprès entre 2016 i 2024.

5 Metodologia

5.1 Generació de seqüències temporals dinàmiques

Un cop generats els grafs horaris amb totes les transformacions descrites anteriorment, el següent pas és agrupar-los en seqüències temporals que puguin ser utilitzades com a entrada per al model basat en xarxes neuronals aplicat a grafs dinàmics. Per fer-ho, s'ha utilitzat el codi `generate_seq.py`, que permet generar seqüències de múltiples grafs horaris consecutius mitjançant una finestra temporal mòbil.

L'script recorre els fitxers acabats en `.pt` (grafs horaris) generats amb `toData.py`, ordenats cronològicament, i construeix seqüències de longitud `window_size`, amb un pas temporal (`stride`) entre seqüències consecutives (Figura 8). Per a cada seqüència, s'inclou la matriu de característiques dels nodes, la informació de lesarestes i dels atributs d'aresta, les metadades, i les etiquetes corresponents al moment `t+horizon` (on `horizon` és l'horitzó de predicció especificat).

Per validar el funcionament correcte del procés de generació, s'ha fet un primer experiment amb els paràmetres següents:

- **window size** = 48 hores
- **stride** = 12 hores
- **horizon** = 6 hores

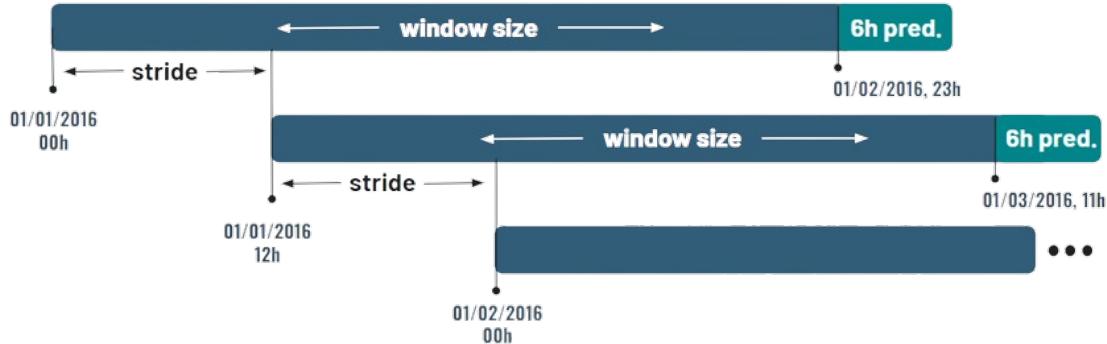


Figura 8: Representació de les primeres seqüències generades amb `window size` = 48h, `stride` = 12h i `horizon` = 6h.

Un cop verificada la correcta generació de les seqüències, sempre mantenint la consistència temporal i preservant les metadades dels grafs individuals, s'ha procedit a generar altres conjunts de seqüències amb finestres i horitzons més amplis, amb l'objectiu que el model pogui aprendre patrons meteorològics a mitjà termini (per exemple, a 5 dies vista):

- **window size** = 120 hores (5 dies)
- **stride** = 12 hores
- **horizon** = 120 hores

Aquestes seqüències permeten capturar tant la dinàmica diària com l'evolució sinòptica dels sistemes meteorològics, afavorint una millor comprensió de processos estacionals, recurrents o d'escala més gran. A més, el pas de 6 hores (`stride`) assegura una certa redundància entre seqüències consecutives i això fa que es mantingui una cobertura temporal densa i s'aprofitin millor les dades disponibles.

L'script també garanteix que només es conserven les seqüències completes, descartant automàticament aquelles en què manca l'etiqueta `y` a l'horitzó `t+horizon` per algun dels passos de la finestra. Tota la informació es desa en fitxers `.pt` (PyTorch) per facilitar una càrrega eficient durant l'entrenament del model.

Estructura interna de les seqüències

Cada seqüència generada per `generate_seq.py` conté un conjunt ordenat de T grafs horaris consecutius (on $T = \text{window_size}$) i una o més etiquetes de predicció a l'horitzó `horizon`. Cada seqüència s'emmagatzema en un únic fitxer binari amb extensió `.pt` i inclou la informació següent:

- **x_seq**: llista de tensors amb les característiques dels nodes que ja hem descrit per a cada pas temporal.
- **mask_seq**: llista de màscares booleanes que indiquen quins nodes són presents a cada pas (útil per tractar estacions intermitents).
- **edge_index_seq**: llista d'arestes de cada graf, amb índexs globals per a cada moment temporal.
- **edge_attr_seq**: atributs associats a cada aresta en cada pas temporal.
- **y_seq**: llista de tensors amb les etiquetes a predir per cada pas de l'horitzó (normalment el tensor **x** del mateix node però `horizon` hores després).
- **y_mask_seq**: màscara que indica si l'etiqueta corresponent està disponible.
- **timestamps**: llista amb l'estampa temporal de cada pas de la seqüència.
- **fonts_seq, norm_params_seq, meta_seq, pos_seq, year_seq**: metadades útils per mantenir la traçabilitat i la consistència entre seqüències, com la font de les dades, els paràmetres de normalització utilitzats, informació geogràfica i l'any.
- **id_seq**: llista dels identificadors dels nodes presents en cada timestamp, per tal de garantir una traçabilitat precisa, atès que es treballa amb grafs dinàmics en què els nodes i les arestes poden variar en cada graf horari dins d'una seqüència.

5.2 Agrupació eficient de seqüències per a l'entrenament massiu

La generació de seqüències temporals meteorològiques produceix milers de fitxers individuals, cada-un representant una única seqüència temporal horària (per exemple, `2023030100_2023030223.pt`), i pot arribar a ocupar desenes de gigabytes d'emmagatzematge. Aquesta organització fragmentada genera problemes pràctics a l'hora d'entrenar models en entorns computacionals intensius, com el clúster de GPUs del Centre de Visió per Computador (CVC). En concret, la necessitat d'obrir i llegir milers de fitxers petits genera colls d'ampolla d'entrada/sortida (I/O bottlenecks), incrementa de manera substancial el temps de càrrega i pot superar els límits de descriptors de fitxer permesos pel sistema operatiu, provocant errors i aturades inesperades.

Per abordar aquest repte, s'ha desenvolupat l'script `all_sequences.py`, que automatitza l'agrupació de les seqüències generades per `generate_seq.py` en fitxers més grans anomenats *chunks*. Cada chunk agrupa un nombre configurable de seqüències temporals (per exemple, 25 o 50), i es desa com un únic fitxer binari. Això redueix dràsticament el nombre total de fitxers i minimitza la penalització associada a les operacions d'I/O, facilitant una càrrega molt més eficient del dataset durant l'entrenament al clúster.

Concretament, l'script genera dos tipus de fitxers per a cada agrupació de seqüències:

- **Fitxer de chunk (`chunk_XXX.pt`)**: conté una llista de `XXX` seqüències temporals, emmagatzemades com a objectes de Python serialitzats amb `torch.save`. Aquest fitxer és el que s'utilitza directament per a la càrrega massiva durant l'entrenament.
- **Fitxer de metadades (`chunk_XXX_meta.pt`)**: recull la llista dels noms base (`.pt`) de les seqüències originals incloses en el chunk `XXX`. Aquesta informació facilita la traçabilitat i permet identificar fàcilment l'origen de qualsevol incidència o error associat a una seqüència concreta.

5.3 Entrenament, validació i test

L'entrenament, la validació i el test constitueixen les tres etapes fonamentals del procés d'avaluació de qualsevol model d'aprenentatge automàtic. El flux general d'aquest procés consisteix a dividir el conjunt total de dades en tres subconjunts: un conjunt d'entrenament (*train*), que s'utilitza per ajustar els paràmetres del model; un conjunt de validació (*validation*), que permet monitoritzar el rendiment del model durant l'entrenament i detectar possibles situacions de sobreajustament (*overfitting*); i, finalment, un conjunt de test (*test*), que serveix per mesurar la capacitat real de generalització del model un cop finalitzat tot el procés d'entrenament.

En aquesta secció es descriu en detall tot el procés d'entrenament, validació i test del model **MeteoGraphPC** (Meteorologia en Grafs dels Països Catalans), així com la gestió dels conjunts de dades, la selecció d'hiperparàmetres i la comparació amb baselines. S'explica el funcionament de l'script `MeteoGraphPC.py` desenvolupat juntament amb la configuració dels arguments, la preparació dels `DataLoader`, l'arquitectura del model implementat i la metodologia emprada per avaluar el rendiment del model sobre dades meteorològiques dels Països Catalans. L'execució de `MeteoGraphPC.py` es fa amb el codi bash `run_MeteoGraphPC.sh` que és, a la vegada, el fitxer principal de configuració. Aquesta secció també inclou la definició de les mètriques utilitzades i el procediment seguit per intentar aconseguir el màxim de robustesa en els resultats.

5.3.1 Preparació dels grups de seqüències

El primer pas fonamental del procés consisteix en la càrrega de les dades meteorològiques, estructurades en grups de seqüències tal i com s'ha vist ens els apartats anteriors. La gestió de les dades es realitza principalment mitjançant la classe `GraphSeqDataset` que gestiona la lectura eficient i l'organització de seqüències temporals de grafs meteorològics, permetent la selecció flexible de variables d'entrada i sortida, la neteja estructural dels grafs i la càrrega a demanda de les seqüències necessàries per a l'entrenament i la inferència del model.

El conjunt complet de dades es divideix en tres subconjunts temporals no solapats. La divisió cronològica estricta, implementada mitjançant la funció `split`, assegura que no hi hagi correlació artificial entre entrenament i test, i reflecteix un escenari realista de predicció meteorològica on sempre es prediu el futur a partir del passat.

- **Conjunt d'entrenament:** inclou totes les seqüències des de l'any 2016 i fins al 2022 (inclosos) i s'utilitza per ajustar els paràmetres del model.
- **Conjunt de validació:** correspon a les seqüències de l'any 2023 i serveix per monitoritzar el rendiment del model durant l'entrenament i regular la complexitat per evitar un sobreajustament.
- **Conjunt de test:** recull els grups de seqüències de l'any 2024 i permet avaluar la capacitat de generalització del model un cop finalitzat tot l'entrenament.

5.3.2 Normalització de les dades

Per tal de garantir una convergència estable durant l'entrenament i evitar que les diferències d'escala entre variables meteorològiques introduixin biaixos en el procés d'aprenentatge, s'aplica una **normalització estandarditzada** a les dades d'entrada i, especialment, a les variables objectiu. Aquesta normalització consisteix a restar la mitjana i dividir per la desviació estàndard de cada variable (valors calculats mitjançant `compute_PC_norm_params.py`). D'aquesta manera el rang resultant està centrat en zero i té una dispersió comparable per a totes les variables.

Aquesta transformació s'aplica tant a les variables d'entrada com a les de sortida segons correspongui, assegurant així que el model pugui aprendre patrons reals i no simplement relacions degudes a escales diferents. Durant el càlcul de les mètriques finals i per a la interpretació dels resultats, es realitza la **desnormalització** de les prediccions, invertint la transformació.

La normalització facilita l'estabilitat numèrica, permet una optimització més ràpida i fiable, i garanteix que totes les variables meteorològiques tinguin un pes comparable durant l'entrenament i l'avaluació del model.

5.3.3 Inicialització i configuració del model

Un cop preparades i normalitzades les dades, el següent pas del procés consisteix en la **definició, inicialització i configuració del model** que s'utilitzarà per a l'experimentació. Al present treball, s'ha implementat el model **MeteoGraphPC**, que permet la predicción multivariable i multiestació sobre grafs dinàmics mitjançant arquitectures de tipus Graph Neural Network.

La configuració dels hiperparàmetres del model (com la mida de les capes ocultes, l'horitzó de predicción, el percentatge de *dropout* o les variables d'entrada i sortida, etc.) es realitza mitjançant arguments específics que es passen a l'script `run_MeteoGraphPC.sh`.

L'arquitectura del model **MeteoGraphPC** es pot dividir conceptualment en quatre blocs principals:

- **Entrada i processament de seqüències de grafs dinàmics:**

El model rep com a entrada els grups de seqüències temporals de grafs dinàmics explícits en els apartats anteriors. Per cada pas temporal, es mantenen la topologia i les característiques dels nodes i les arestes segons la informació disponible.

- **Processament seqüencial spaitemporal amb TGCN:**

El nucli del model (Figura 9) és una cèl·lula recurrent de tipus *Temporal Graph Convolutional Network* (TGCN), que integra de forma conjunta l'aprenentatge de patrons temporals (GRU) i espacials entre estacions (GCN). Aquesta cèl·lula actualitza de manera recurrent l'estat intern de cada node al llarg de la seqüència d'entrada, combinant la informació pròpia de cada node, la dels seus veïns i l'evolució temporal. No hi ha un encoder explícit; tota la representació seqüencial s'aprèn directament dins la TGCN, que actua alhora com a capa de grafs i de memòria recurrent.

- **Decodificació autoregressiva i generació de prediccions:**

Un cop processada la finestra d'entrada, el model utilitza l'estat intern final de cada node per predir, de manera autoregressiva, els valors futurs de les variables meteorològiques. A cada pas de l'horitzó de predicción, la sortida prèvia s'utilitza com a part de la nova entrada, permetent capturar la dependència temporal a diversos passos vista durant l'entrenament.

Les sortides finals són tensors que recullen les prediccions multivariables per a cada node (estació meteorològica) i per a cada pas de l'horitzó futur escollit. Així, el model farà una predicción per a cadascuna de les seqüències del conjunt de test seleccionat.

- **Entrenament i optimització:**

L'entrenament es fa de manera supervisada, comparant les seqüències de prediccions generades amb les observacions reals corresponents mitjançant una funció de pèrdua (tipus MSE), i optimitzant els pesos del model amb l'optimitzador Adam i tècniques modernes de regularització.

MeteoGraphPC

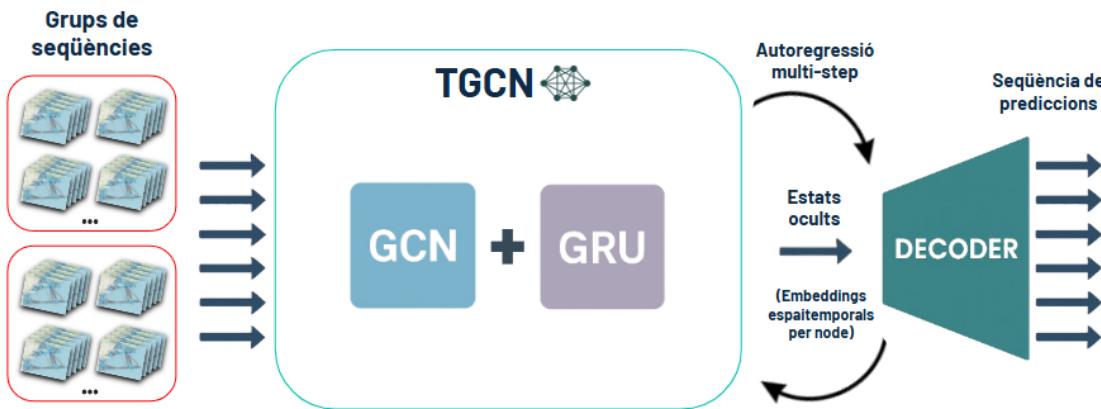


Figura 9: Arquitectura general del model MeteoGraphPC.

A continuació es descriuen tots els paràmetres de configuració i hiperparàmetres que es poden modificar per adaptar l'entrenament del model MeteoGraphPC. Aquests paràmetres es poden passar directament com a arguments a l'script `MeteoGraphPC.py` o bé editar-se al script `run_MeteoGraphPC.sh`.

Paràmetres de sistema i recursos

- `ulimit -v` (**Límit de RAM**): quantitat màxima de memòria RAM (en bytes) per al procés. Ex: 100GB.
- `CUDA_VISIBLE_DEVICES`: identificador de GPU a utilitzar (0, 1, etc).
- `OMP_NUM_THREADS, MKL_NUM_THREADS`: nombre de fils/threads de CPU per a operacions paral·leles.
- `PYTHONUNBUFFERED`: si val 1, força la sortida de Python a ser "no bufferitzada" (útil per veure logs en temps real).

Paràmetres principals de l'entrenament (arguments CLI)

- `seq_dir`: directori amb les seqüències de dades (.pt). *Permet escollir amb quins grups de seqüències ja generats entrenar el model.*
- `batch_size`: mida del batch (int, ex: 4, 8, 16). *Controla quantes seqüències es processen simultàniament abans de fer una actualització dels pesos. Un valor alt pot accelerar l'entrenament (amb prou RAM/VRAM) però pot fer que l'optimització sigui menys "fina". Un valor baix afavoreix una optimització més "sorollosa", però consumeix menys memòria.*
- `epochs`: nombre màxim d'époques (int, ex: 50). *Defineix quantes vegades es recorren totes les dades d'entrenament. Si el model ja ha après (o si hi ha early stopping), potser no arriba a aquest valor màxim. Valors més alts permeten un entrenament més prolongat, però també poden portar a sobreajustament (overfitting) si no es regula.*
- `lr`: learning rate de l'optimitzador (float, ex: $1 \cdot 10^{-4}$). *Controla la velocitat d'aprenentatge del model. Un valor massa alt pot fer que el model no convergeixi; un valor massa baix pot fer que l'entrenament sigui molt lent.*
- `lr_scheduler`: tipus de scheduler pel learning rate (es pot escollir entre "onecycle" o "plateau"). *Permet ajustar automàticament el learning rate durant l'entrenament. OneCycleLR sol donar bons resultats en entrenaments moderns; ReduceLROnPlateau redueix el learning rate si la pèrdua no millora.*
- `hidden_dim`: dimensió oculta de les capes internes del model (int, ex: 128). *Controla la capacitat interna del model. Valors més alts permeten captar patrons més complexos però també augmenten el risc d'overfitting i el consum de memòria.*

- **grad_clip**: valor de clipping pel gradient (ex: 1.0). *Limita la norma màxima dels gradients durant el backpropagation, cosa que ajuda a prevenir explosions de gradient i fa més estable l'entrenament, especialment amb models seqüencials.*
- **patience**: nombre d'èpoques sense millora abans de fer *early stopping*. *Permet aturar l'entrenament automàticament si el model no millora en X èpoques consecutives, evitant perdre temps i sobreajustament.*
- **min_delta**: millora mínima a assolir per resetear la *patience* (ex: $1 \cdot 10^{-4}$). *Defineix el mínim canvi que s'ha de produir en la mètrica de validació perquè es consideri una millora real. Si la millora és menor, es compta com a "no millora".*
- **device**: dispositiu d'entrenament ("cuda" o "cpu"). *Indica si l'entrenament es farà a GPU (molt més ràpid) o CPU.*
- **seed**: llavor per garantir la reproductibilitat (**int**, ex: 42). *Permet obtenir sempre els mateixos resultats d'entrenament a partir dels mateixos paràmetres i dades.*
- **std_eps**: valor petit per evitar divisions per zero a la normalització (ex: $1 \cdot 10^{-6}$). *Important per l'estabilitat numèrica quan alguna variable té una desviació estàndard molt baixa, com per exemple la precipitació.*
- **save_dir**: directori on es guardaràn els checkpoints i resultats. *Permet organitzar els resultats i recuperar models entrenats més endavant.*
- **log_csv**: fitxer on es registrà l'entrenament. *És útil per fer seguiment, comparar models i traçar mètriques a posteriori.*
- **dl_num_workers**: nombre de processos/threads per al *DataLoader* (ex: 2, 4, 8). *Permet accelerar la càrrega de dades quan es disposa d'una CPU amb molts nuclis. Pot fer l'entrenament més eficient, sobretot amb fitxers grans.*
- **input_indices**: índexs de les columnes de les característiques d'entrada a utilitzar (ex: 0 1 2 ... 16). *Permet entrenar el model només amb un subconjunt de variables. Tot i així, sempre és recomanable entrenar el model amb totes les variables disponibles.*
- **target_indices**: índexs de les columnes de variables a predir (ex: 0 1 2 3 4 15 16). *Especifica quines variables es vol que el model predigui. Permet fer predicció multi-variable o centrar-se en una de sola. No es recomana incloure les variables Alt_norm, VentDir_sin, VentDir_cos, hora_sin, hora_cos, dia_sin, dia_cos, cos_sza, DewPoint i PotentialTemp, ja que són derivades, cícliques o calculades a partir d'altres variables i no tenen sentit físic com a targets de predicció directa. És millor limitar els targets a les variables meteorològiques observables com temperatura, humitat, pluja, vent, pressió, etc.*
- **use_edge_attr**: si s'activa, inclou atributs de les arestes com a entrada al model. *Millora la capacitat del model per tenir en compte relacions entre nodes més enllà de la connexió topològica. És recomanable utilitzar-la.*
- **use_mask**: si s'activa, inclou màscares de nodes absents com a entrada. *Ajuda el model a ignorar nodes inexistentes en alguns timestamps, fent-lo més robust a dades incomplides o seqüències irregulars. És recomanable utilitzar-la.*
- **norm_json**: fitxer JSON amb mitjanes i desviacions estàndard per a cada feature per a la normalització per tal d'evitar el càlcul cada vegada. *Agilitza l'entrenament i garanteix la mateixa normalització entre runs diferents.*
- **dropout**: proporció de neurones disconnectades durant l'entrenament, per regularitzar el model i reduir el sobreajustament. *Valors més alts (ex: 0.3) fan el model més robust però poden frenar l'aprenentatge; valors baixos (ex: 0.1) afavoreixen que s'ajusti molt a les dades. Cal modificar-lo directament al codi de *MeteoGraphPC.py*.*
- **weight_decay**: paràmetre de regularització L2 de l'optimitzador (en aquest treball s'utilitza Adam donada la seva eficiència computacional i la seva capacitat per adaptar dinàmicament la taxa d'aprenentatge de cada pes). Penalitza pesos molt grans, ajudant a controlar el sobreajustament. *Valors típics: $1 \cdot 10^{-5}$, $1 \cdot 10^{-4}$. Cal modificar-lo a la creació de l'optimitzador, al codi de *MeteoGraphPC.py*.*

Aquest conjunt d'elements assegura que el model MeteoGraphPC estigui degudament inicialitzat i que l'entrenament es pugui realitzar de manera robusta, flexible i eficient, independentment de la configuració escollida.

5.3.4 Procés d'entrenament

El procés d'entrenament del model MeteoGraphPC s'estructura com un bucle iteratiu d'èpoques, durant el qual el model aprèn a optimitzar els seus paràmetres per minimitzar l'error de predicció sobre el conjunt d'entrenament.

A cada època, les seqüències de dades s'organitzen en lots mitjançant el `DataLoader`, i per cada lot es realitzen les següents operacions principals:

- **Propagació endavant (forward pass):** el model processa les seqüències d'entrada, genera prediccions per a totes les estacions i passos temporals futurs de l'horitzó definit.
- **Càcul de la pèrdua:** es calcula l'error (*loss*) entre les prediccions del model i els valors reals normalitzats, utilitzant la funció d'error quadràtic mitjà (MSE).
- **Propagació enrere (backward pass):** es calcula el gradient de la pèrdua respecte als paràmetres del model.
- **Actualització dels paràmetres:** s'actualitzen els paràmetres del model utilitzant l'optimitzador (Adam) i el planificador de taxa d'aprenentatge, amb control addicional de l'escala dels gradients per evitar desbordaments numèrics.

Durant l'entrenament, a més, es monitoritza periòdicament el rendiment del model sobre el conjunt de validació, calculant mètriques com l'RMSE, MAE, R² i SMAPE. Aquest monitoratge permet regular l'entrenament i activar mecanismes d'aturada anticipada com *early stopping* si el rendiment sobre validació no millora durant un nombre predefinit d'èpoques consecutives, evitant així el sobreajustament.

A cada època, es registren totes les mètriques i pèrdues en fitxers de log per facilitar l'anàlisi posterior de la corba d'aprenentatge i la comparació entre diferents configuracions del model.

5.3.5 Guardat i restauració del millor model

Per garantir que el model final seleccionat sigui el que presenta el millor rendiment en el conjunt de validació, s'implementa un sistema de **guardat de punts de control** (*checkpoints*) durant el procés d'entrenament.

Al final de cada època, es compara el valor de la mètrica de validació principal (habitualment l'RMSE) amb el millor resultat obtingut fins al moment. Si el model actual supera el rendiment anterior, s'emmagatzema l'estat complet del model, així com l'estat de l'optimitzador i el planificador de taxa d'aprenentatge, en un fitxer de checkpoint. Aquest procediment assegura que, encara que el model pugui sobreajustar-se en èpoques posteriors, sempre es conserva la versió òptima identificada durant la validació.

Un cop finalitzat l'entrenament, ja sigui per esgotament de les èpoques o per activació d'*early stopping*, es **restaura** l'estat del model corresponent al millor checkpoint guardat. D'aquesta manera, totes les evaluacions posteriors (test, càcul de mètriques finals i comparació amb línies base) es realitzen amb la versió del model que ha demostrat una millor capacitat de generalització sobre dades no vistes durant l'entrenament.

5.3.6 Avaluació sobre el conjunt de test i càlcul de baselines

Un cop finalitzat l'entrenament i restaurat el millor model, s'avalua la seva capacitat de generalització aplicant-lo sobre el **conjunt de test**, el qual conté dades completament noves i no utilitzades durant les fases prèvies. Aquesta avaluació proporciona una mesura objectiva del rendiment real del model en situacions no vistes, i permet validar la robustesa dels resultats.

Per tal de contextualitzar i interpretar correctament el rendiment del model MeteoGraphPC, es calculen dues línies base (*baselines*) amb les quals comparar els resultats obtinguts:

- **Persistència:** consisteix a considerar que el valor de cada variable meteorològica en cada estació i instant futur és igual al darrer valor conegut (estratègia habitual en predicció meteorològica a curt termini).
- **Climatologia:** consisteix a utilitzar la mitjana històrica de cada variable i estació com a predicció per a l'horitzó futur, representant una línia base estacionària que no té en compte la variabilitat recent.

Aquesta comparació permet quantificar fins a quin punt el model aprèn patrons reals de les dades més enllà d'una simple persistència temporal o d'un comportament mitjà.

Finalment, es desen totes les prediccions i mètriques en fitxers específics per a la seva posterior anàlisi, visualització i comparació entre experiments i arquitectures.

5.3.7 Càcul i anàlisi de mètriques de rendiment

Per quantificar de manera objectiva el rendiment del model MeteoGraphPC i de les línies base, s'utilitzen diverses **mètriques d'avaluació** sobre el conjunt de test, a partir de les prediccions obtingudes i els valors reals. Les mètriques principals utilitzades són:

- **RMSE (Root Mean Squared Error):** mesura l'arrel quadrada de la mitjana dels errors quadràtics, proporcionant una idea clara de la magnitud mitjana de l'error en les unitats originals de cada variable.
- **MAE (Mean Absolute Error):** calcula la mitjana dels valors absoluts de les diferències entre predicció i realitat, sent menys sensible a valors extrems que l'RMSE.
- **R² (Coeficient de determinació):** indica la proporció de la variabilitat total explicada pel model; valors propers a 1 impliquen millor ajust.
- **SMAPE (Symmetric Mean Absolute Percentage Error):** és una versió simètrica del MAPE que normalitza l'error absolut mitjançant la mitjana entre valors reals i predicts, i és especialment útil quan hi ha valors petits o propers a zero.

Aquestes mètriques es calculen tant de forma global com desglossades per variable meteorològica i horitzó de predicció, així com per cada línia base.

6 Resultats

A continuació s'expliquen amb detall les prediccions dutes a terme amb el model MeteoGraphPC a curt i a mitjà termini. L'entrenament s'ha realitzat utilitzant la infraestructura GPU del *Centre de Visió per Computador* i s'ha automatitzat amb l'script `run_MeteoGraphPC.sh`.

Els paràmetres d'entrenament utilitzats en les prediccions són els següents:

Paràmetre	Valor
Mida del batch (<code>batch_size</code>)	8
Ritme d'aprenentatge (<code>learning_rate</code>)	5×10^{-6}
Scheduler d'aprenentatge (<code>lr_scheduler</code>)	<code>onecycle</code>
Dimensió oculta del model (<code>hidden_dim</code>)	128
Paciència per early stopping (<code>patience</code>)	6
Millora mínima per early stopping (<code>min_delta</code>)	1×10^{-4}
Dispositiu d'entrenament (<code>device</code>)	<code>cuda</code>
Límit de gradient (<code>grad_clip</code>)	1.0
Tolerància numèrica per la normalització (<code>std_eps</code>)	1×10^{-6}
Número de subprocessos DataLoader (<code>dl_num_workers</code>)	2
Fitxer de paràmetres de normalització (<code>norm_json</code>)	<code>PC_norm_params.json</code>

Table 1: Paràmetres d'entrenament utilitzats en l'entrenament de MeteoGraphPC (predicció a curt termini).

El model s'ha entrenat durant 30 èpoques, en el cas de la predicció a curt termini, i amb 20 èpoques, en el cas de la predicció a mitjà termini, amb l'objectiu de minimitzar l'error quadràtic mitjà (MSE). El procés inclou:

- Entrenament amb l'optimitzador Adam, utilitzant un weight decay de 1×10^{-5} i el scheduler OneCycleLR.
- Aplicació d'un dropout de 0.3 al model.
- Early stopping basat en la mètrica RMSE de validació.
- Guardat automàtic del millor model segons el Root Mean Squared Error (RMSE) de validació.

S'han utilitzat totes les variables disponibles com a features d'entrada per a l'entrenament del model en els dos casos. Concretament:

- Temperatura, Humitat, Pluja, VentFor, Patm, Alt_norm, VentDir_sin, VentDir_cos, hora_sin, hora_cos, dia_sin, dia_cos, cos_sza, DewPoint, PotentialTemp, Vent_u, Vent_v

En canvi, les **variables objectiu** que s'ha volgut predir són les següents:

- Temperatura, Humitat, Pluja, VentFor, Patm, Vent_u, Vent_v.

Aquesta selecció de variables a predir s'ha fet ja que són les més rellevants per a la interpretació meteorològica i disposen de registres fiables a la majoria d'estacions, mentre que la resta de variables (per exemple, noves variables calculades, codificacions temporals o d'altitud) s'utilitzen només com a informació auxiliar per ajudar a la predicció però no són d'interès com a targets independents.

També s'han utilitzat atributs d'aresta (`-use_edge_attr`) i màscares de nodes (`-use_mask`) per tractar la dinàmica espacial i l'absència temporal d'estacions meteorològiques.

6.1 Predicció a curt termini

Primer de tot, es presenten els resultats obtinguts amb el model MeteoGraphPC executat amb grups seqüències per a la predicció a curt termini.

Les seqüències utilitzades durant l'entrenament han estat construïdes de la següent manera:

- Finestres temporals de 48 hores, un stride de 12 hores i un horitzó de predicció 6 hores.
- Grups de 50 seqüències per tal d'optimitzar la càrrega i la gestió de memòria.

6.1.1 Anàlisi de les mètriques

A continuació, s'analitza l'evolució de les principals mètriques d'avaluació (loss, MAE, R^2 , RMSE i SMAPE) durant l'entrenament i la validació del model durant la predicció a curt termini, amb l'objectiu d'identificar el comportament d'aprenentatge i el punt òptim de parada.

Loss: A la Figura 10 s'observa una disminució progressiva de la loss tant per al conjunt d'entrenament com de validació, indicant una correcta convergència del model. El valor mínim de la loss en validació s'assoleix a l'època 30 amb un valor de 0.2339, moment en què es selecciona el millor model segons el criteri d'early stopping. L'evolució suau de la loss, sense increments sobtats ni divergències, posa de manifest una dinàmica d'aprenentatge estable.

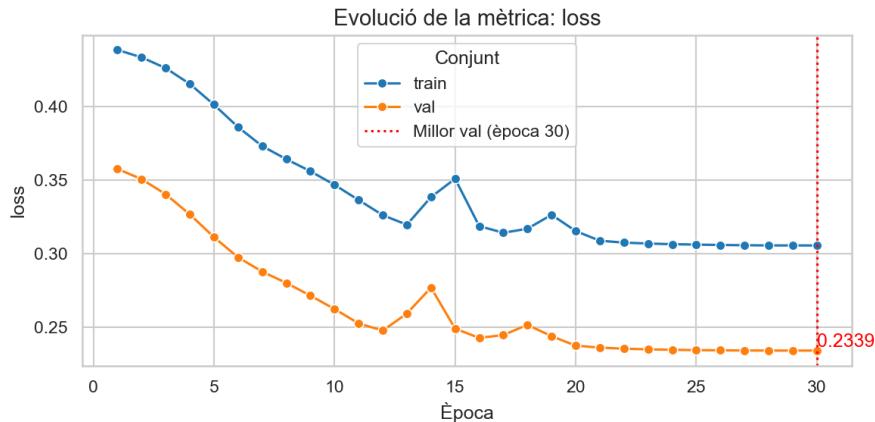


Figura 10: Evolució de la mètrica de loss durant l'entrenament i la validació en la predicció a curt termini.

MAE: La Figura 11 mostra l'evolució del MAE (Mean Absolute Error) en els conjunts d'entrenament i validació al llarg de les èpoques. Tant en el train com en el val es produeix una disminució progressiva del MAE fins a una estabilització als valors mínims, sense increments sobtats ni senyals d'overfitting entre aquests dos conjunts.

Tanmateix, el MAE obtingut en el conjunt de test (representat amb una línia discontinua) es troba significativament per sobre dels valors de validació i entrenament, cosa que pot indicar certa dificultat de generalització del model cap al test (any 2024). Aquest comportament podria ser degut a una possible diferència de distribució entre els conjunts, a una complexitat insuficient del model o a la presència de casos atípics al test.

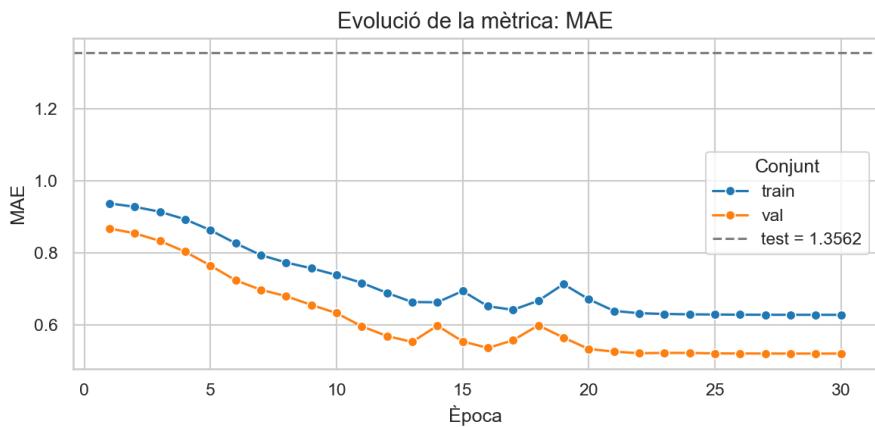


Figura 11: Evolució de la mètrica MAE durant l'entrenament i la validació en la predicció a curt termini.

Coeficient de determinació R^2 : La Figura 12 mostra l'evolució del coeficient de determinació R^2 durant l'entrenament i la validació. Tant en el conjunt de train com en el de val, s'observa

una clara tendència creixent, que indica que el model incrementa progressivament la seva capacitat explicativa a mesura que aprèn. A les últimes èpoques, els valors de R^2 s'estabilitzen al voltant de 0.37.

En canvi, el valor de R^2 obtingut en el conjunt de test (línia discontinua) és lleugerament inferior als màxims assolits en validació, però es manté proper i dins el mateix rang. Això suggereix que el model generalitza de forma raonablement robusta, tot i que, com és habitual, presenta un petit descens de rendiment en test, possiblement per la presència de casos més difícils o per lleugeres diferències de distribució entre conjunts.

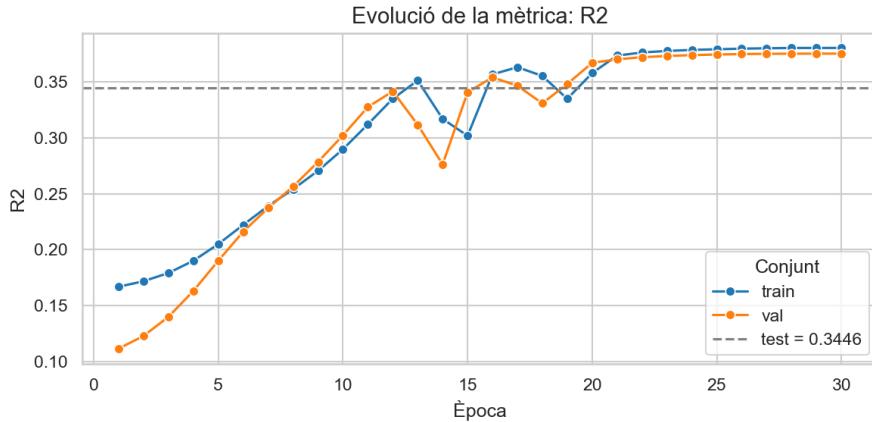


Figura 12: Evolució del coeficient de determinació R^2 durant l'entrenament i la validació en la predicció a curt termini.

RMSE: La Figura 13 presenta l'evolució de la mètrica RMSE (Root Mean Squared Error) durant l'entrenament i la validació. En ambdós conjunts s'observa una disminució progressiva de l'error fins a una estabilització als valors mínims, sense evidència d'overfitting entre train i val.

Tanmateix, el valor del RMSE obtingut en el conjunt de test (representat amb línia discontinua) se situa clarament per sobre dels valors finals de validació i entrenament. Aquest fet posa de manifest un cert desajust en la capacitat de generalització del model, que pot ser atribuïble a una diferència de distribució entre conjunts, a una possible manca de capacitat del model per captar determinades dinàmiques del test, o a la presència de casos atípics. Per tal de millorar aquest comportament, caldria aprofundir en l'anàlisi dels errors específics en test i considerar estratègies addicionals, com l'augment de dades, la regularització o un millor ajust d'hiperparàmetres.

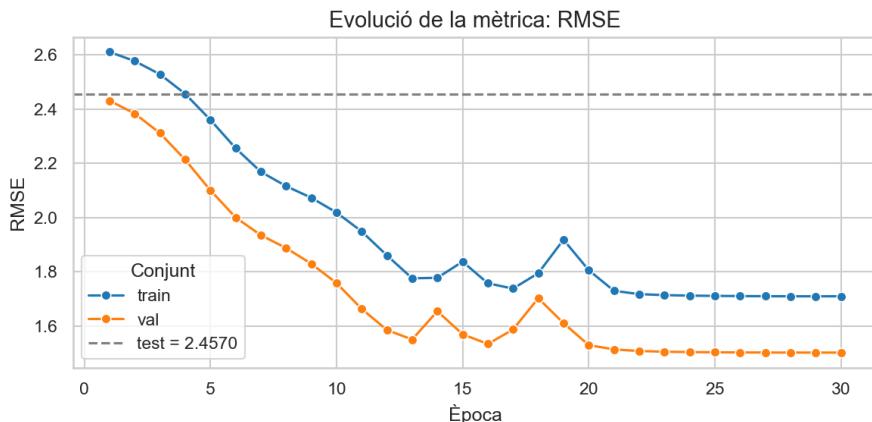


Figura 13: Evolució de la mètrica RMSE durant l'entrenament i la validació en la predicció a curt termini.

SMAPE: Finalment, la Figura 14 mostra l'evolució de la mètrica SMAPE (Symmetric Mean Absolute Percentage Error) en els conjunts d'entrenament i validació. S'hi observa una disminució

progressiva i una estabilització als valors mínims, sense increments sobtats ni indicis d'overfitting entre train i val.

En canvi, el valor de SMAPE en el conjunt de test (línia discontinua) és superior als valors finals d'entrenament i validació, la qual cosa posa de manifest una dificultat significativa del model per generalitzar aquesta mètrica percentual en el test. Aquesta situació pot deure's a la sensibilitat extrema del SMAPE als errors en valors petits de les variables objectiu, a la presència de moltes prediccions zero (per exemple, en la precipitació), o bé a la presència de casos atípics o canvis de distribució en el conjunt de test.

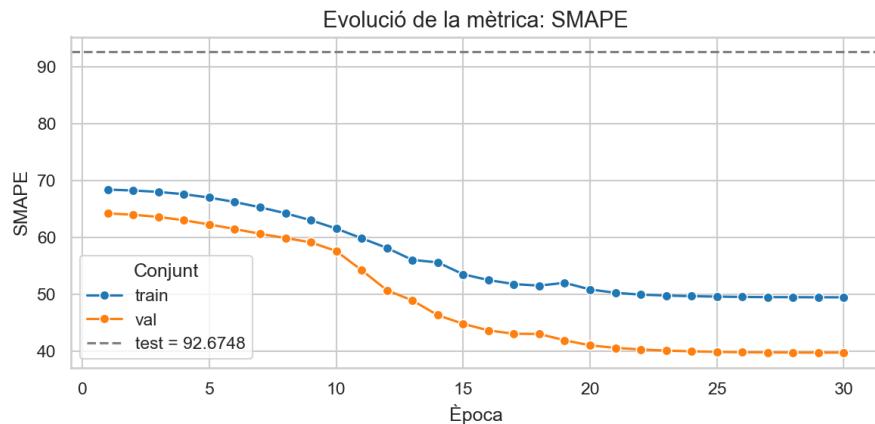


Figura 14: Evolució de la mètrica SMAPE durant l'entrenament i la validació en la predicció a curt termini.

En conjunt, l'evolució de les mètriques durant l'entrenament i la validació mostra un procés d'aprenentatge estable, sense signes evidents d'overfitting. Pel que fa a la capacitat de generalització del model, però, s'observa un desajust clar entre els valors de test i els de validació en la majoria de mètriques, especialment en el MAE, el RMSE i el SMAPE, on el valor de test resulta ser significativament més elevat.

Seguidament es fa una comparativa de les mètriques per a cada variable objectiu durant el test:

Variable	RMSE	MAE	R ²	SMAPE	RMSE (pers.)	RMSE (clima)
Temp	3.96	2.95	0.709	1.02	1.94	7.46
Humitat	0.14	0.11	0.461	17.64	0.08	0.19
Pluja	0.20	0.06	0.039	197.76	0.22	0.20
VentFor	1.63	1.17	0.157	60.11	1.12	1.78
Patm	4.24	2.86	0.679	76.60	1.72	7.50
Vent_u	1.71	1.19	0.204	143.92	1.43	1.92
Vent_v	1.73	1.15	0.164	151.68	1.41	1.90

Table 2: Mètriques de rendiment per a cada variable. S'hi inclouen els valors del model i de les línies base de persistència i climatologia.

A la Figura 15 es presenten les mètriques d'avaluació (MAE, R² i SMAPE) per cadascuna de les variables meteorològiques. S'observa que el model obté els millors resultats en la predicció de la temperatura (*Temp*) i la pressió atmosfèrica (*Patm*), amb valors elevats de R² (0.709 i 0.679, respectivament) i errors baixos (MAE al voltant de 3 °C i 2.9 hPa). Per la variable d'*Humitat*, tot i que l'error absolut (MAE) és baix, la correlació (R²) també ho és, probablement a causa de la menor variabilitat absoluta d'aquesta variable.

Cal destacar la dificultat inherent en la predicció de la precipitació (*Pluja*), com posa de manifest el valor extremadament alt del SMAPE (197.76%), així com un R² pràcticament nul. Aquest comportament és habitual en models automàtics, atesa l'alta esporadicitat i la dominància de valors zero en les sèries temporals de precipitació per a la majoria d'estacions meteorològiques.

Per a les components del vent (*VentFor*, *Vent_u*, *Vent_v*), els errors absoluts són moderats i la capacitat explicativa del model (R^2) és limitada, tot i que millor que en el cas de la precipitació.

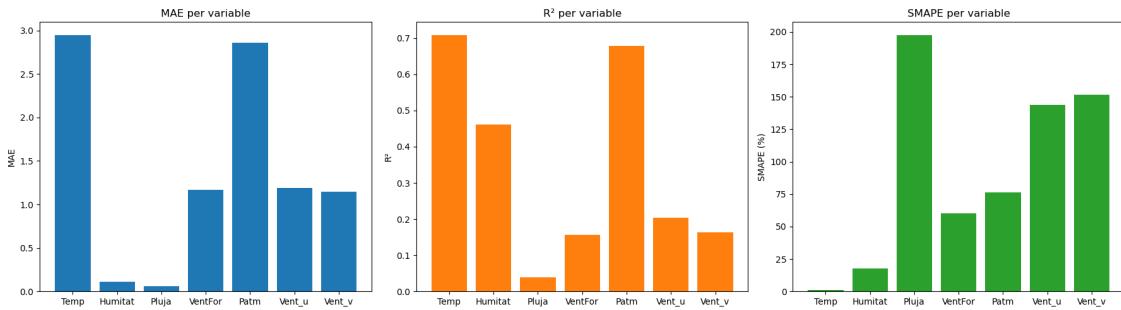


Figura 15: Mètriques d'avaluació per variable meteorològica: MAE, R^2 i SMAPE (%).

A la Figura 16 es mostra la comparativa directa de l'error RMSE del model respecte als baselines de persistència i climatologia per a cada variable. Es pot observar que, per totes les variables, el model supera clarament la climatologia (la barra verda, que representa el RMSE del baseline climàtic, és sempre la més alta), la qual cosa indica que el model aprofita informació temporal i espacial rellevant. No obstant això, la persistència (barra taronja), que consisteix a predir que el valor futur serà igual al present, resulta ser un baseline molt competitiu: el model no aconsegueix superar la persistència en cap de les variables analitzades, sent especialment evident en la temperatura i la pressió atmosfèrica.

Aquesta situació és habitual en la predicció meteorològica a molt curt termini, on la persistència sovint proporciona una aproximació difícil de millorar, especialment per a variables estacionàries o amb canvis graduals. Només en variables més erràtiques, com la precipitació, el model i la persistència, tenen errors similars però amb valors absoluts molt baixos a causa de la baixa incidència de precipitació.

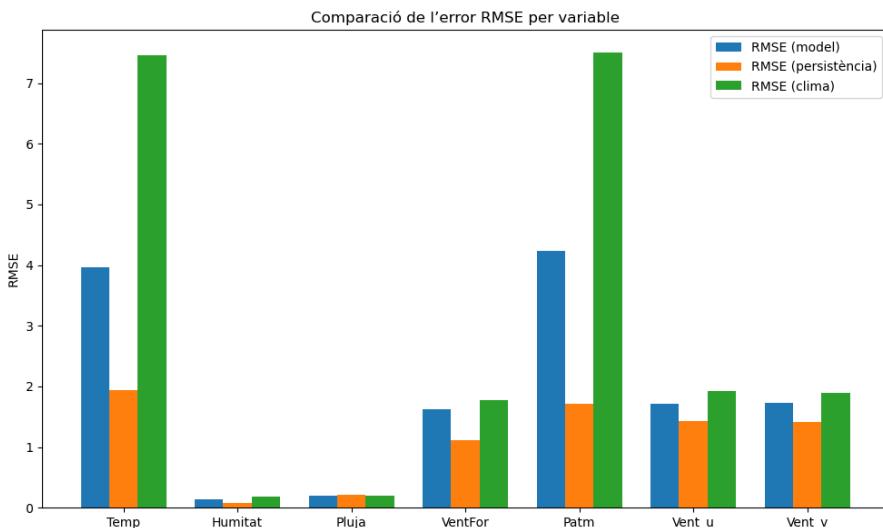
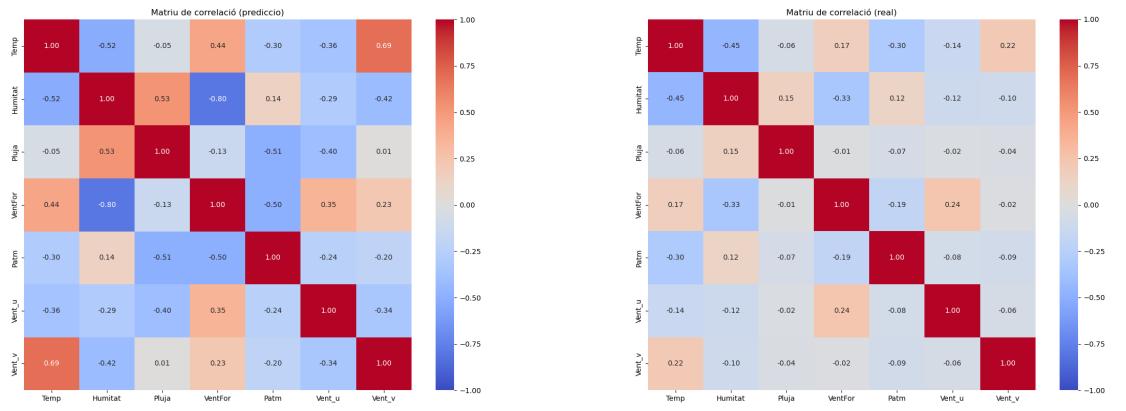


Figura 16: Comparació de l'error RMSE del model amb els baselines de persistència i climatologia per variable.

6.1.2 Anàlisi comparativa de les correlacions: predicció vs. dades reals

A continuació es presenten dues matrius de correlació entre variables meteorològiques: una obtinuda a partir de les prediccions generades pel model MeteoGraphPC i l'altra a partir de les observacions reals. Aquestes matrius, generades amb el codi `matriu_corr.py` (veure Annex), permeten avaluar en quina mesura el model és capaç de reproduir les interdependències estadístiques entre les diferents variables meteorològiques.



(a) Matriu de correlació (predicció)

(b) Matriu de correlació (real)

Figura 17: Comparativa de les correlacions entre variables meteorològiques en les prediccions i en les dades observades.

Podem observar-hi el següent:

- **Temperatura i humitat:** el model reproduceix adequadament la correlació negativa entre temperatura i humitat relativa, tot i que tendeix a sobreestimar la intensitat d'aquesta relació (-0.52 a la predicció vs. -0.45 en les dades reals), suggerint una certa amplificació del comportament invers.
- **Pluja i humitat:** la correlació entre pluja i humitat és més forta a la predicció (0.53) que en les dades reals (0.15), cosa que pot indicar que el model associa de manera més directa situacions humides amb precipitació, tot i que en la realitat aquesta relació és més débil a escala global.
- **Velocitat del vent i humitat:** la correlació negativa entre la velocitat del vent i la humitat relativa és molt intensa a la predicció (-0.80), mentre que és moderada en la realitat (-0.33).
- **Components del vent (u, v):** els components zonal i meridional del vent presenten correlacions més acusades amb la temperatura i amb altres variables en les prediccions (per exemple, correlació Temp-Vent_v = 0.69) que no pas en les dades observades (0.22), suggerint que el model pot estar capturant dinàmiques associades al transport d'aire però potser de manera simplificada o sobredimensionada.
- **Pressió atmosfèrica:** en general, les correlacions entre pressió atmosfèrica i les altres variables també són diferents en ambdós casos excepte la correlació negativa amb la temperatura (-0.30 en tots dos casos).

6.1.3 Anàlisi de les prediccions

Per tal d'analitzar de manera preliminar les prediccions a curt termini generades pel model, s'han generat diversos mapes per cada variable meteorològica per tal de facilitar-ne la interpretabilitat. Aquests mapes s'han generat amb el codi `mapa_preds.py` després d'haver creat un fitxer en format NetCDF (Network Common Data Form: és un tipus de fitxer dissenyat per emmagatzemar i compartir dades multidimensionals, especialment útil meteorologia) mitjançant el codi `inferencia_meteographpc.py` (veure Annex).

A continuació s'analizan les prediccions de MeteoGraphPC a curt termini.

Temperatura: A la Figura 18 s'hi observen les prediccions de temperatura per a la primera seqüència de test de 2024. Concretament, es preduuen les primeres 6 hores del dia 3 de gener de 2024. A la Figura 19, en canvi, s'hi observen les dades reals de les primeres 6 hores del dia 3 de gener de 2024.

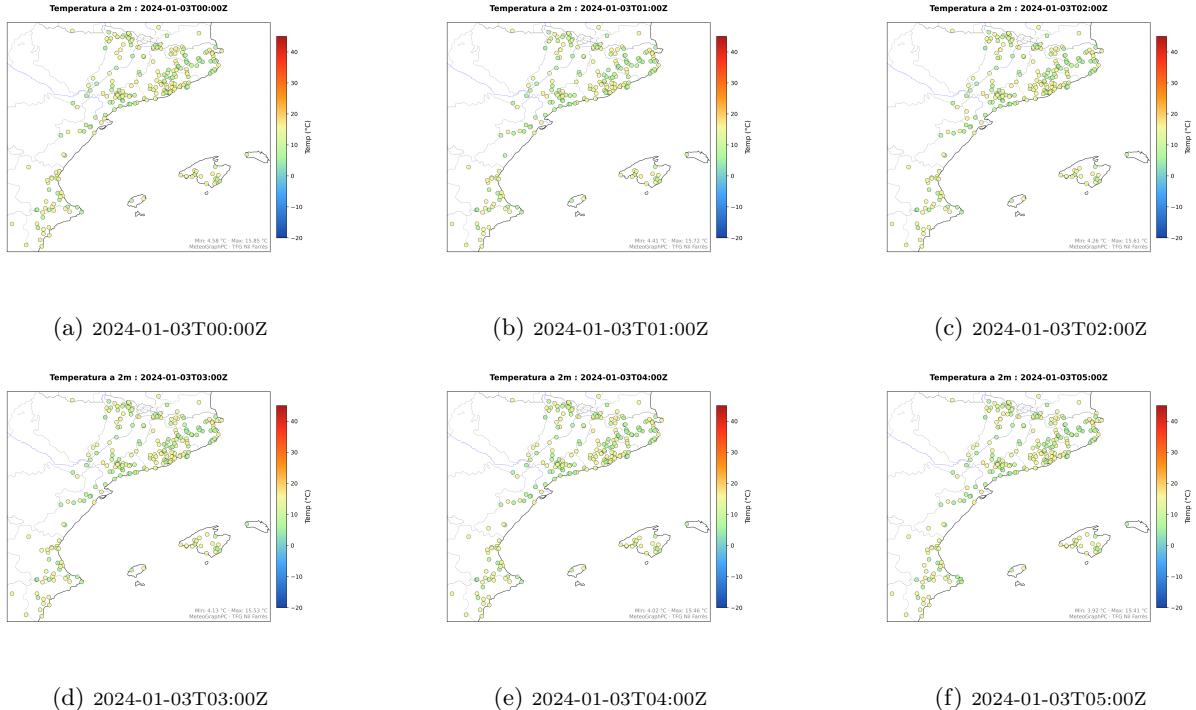


Figura 18: Evolució horària de les prediccions de temperatura a 2 metres realitzades pel model MeteoGraphPC el dia 3 de gener de 2024.

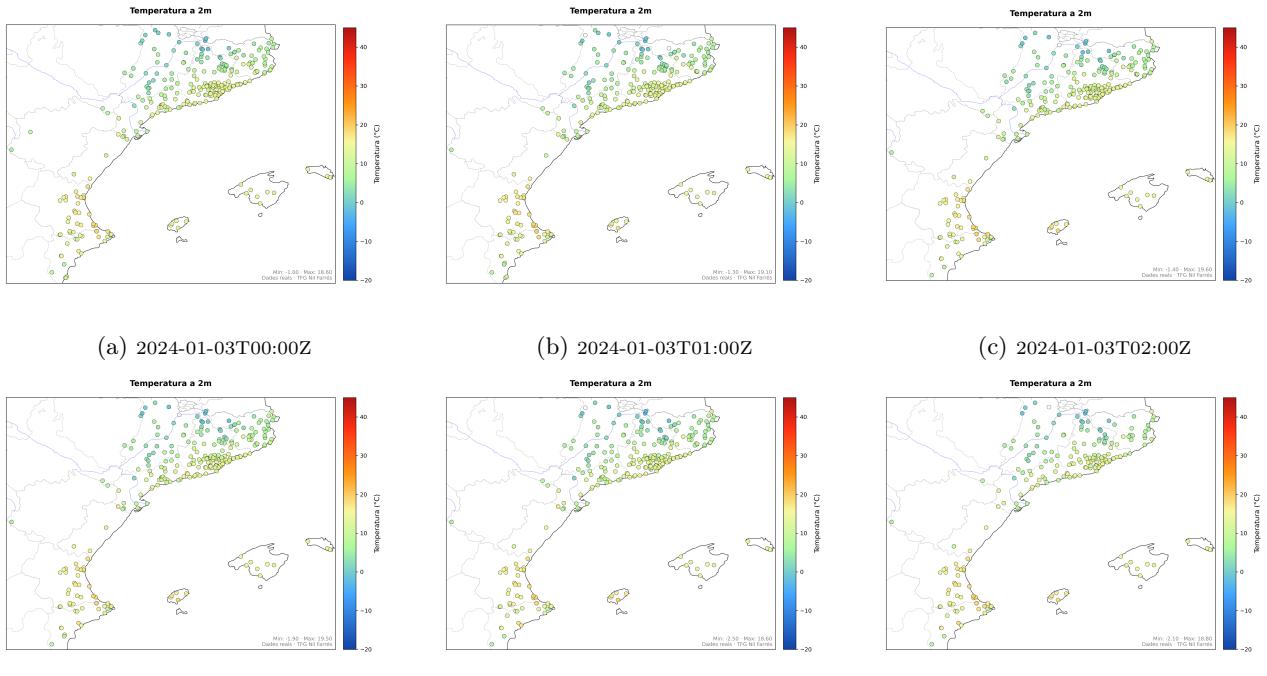


Figura 19: Evolució horària de les dades reals de temperatura a 2 metres el dia 3 de gener de 2024.

Analitzant la primera hora de predicció del model i comparant-la amb la temperatura observada a la mateixa hora, s'observa que la temperatura mitjana regional es troba dins d'uns valors realistes, tal com reflecteixen els colors predominants als dos mapes (tons verdosos i grogosos). Això indica que el model és capaç de captar l'ordre de magnitud global de la temperatura als Països Catalans.

Tanmateix, la distribució espacial de la temperatura a la predicció no reflecteix de manera prou acurada la correlació que s'espera entre temperatura i altitud. Al mapa de prediccions, la disposició

dels colors mostra menys estructura i els patrons associats a l'orografia i als gradients altitudinals són menys evidents que a les dades reals. En conseqüència, la resposta del model a les variacions d'altitud i als factors geogràfics locals sembla limitada, i els valors apareixen més dispersos en comparació amb la realitat.

Les estacions meteorològiques amb el valor mínim (Min: 4.58 °C predicción vs. Min: 1.60 °C observat) i màxim (Max: 15.85 °C vs. Max: 18.66 °C) mostren que el model tendeix a acostar les prediccions cap al valor mitjà.

A més, si s'analitzen les diferents hores consecutives de predicción, es constata que els canvis de temperatura d'una hora a la següent són poc apreciables en la majoria d'estacions, tant en les prediccions com en les dades reals. Aquest comportament és esperable durant una matinada d'hivern sense precipitacions a la vista, on la inèrcia tèrmica fa que els patrons siguin relativament estables. Malgrat tot, aquesta estabilitat horària també podria estar relacionada amb la dificultat del model per captar i predir canvis sobtats o fenòmens locals a curt termini. També es pot apreciar que algunes estacions meteorològiques visibles als mapes de predicción varien amb cada nova hora de predicción. Això es deu a que el model només prediu per a les estacions que tenen dades recents disponibles. Les que no en tenen queden filtrades mitjançant una màscara i no entren al graf d'aquell moment, per evitar prediccions poc fiables.

Humitat: A la Figura 20 s'hi observen les prediccions de la variable d'humitat i les dades reals per a la primera seqüència de test de 2024. Concretament, es prediuen les primeres 6 hores del dia 3 de gener de 2024, però aquí només es mostra la primera ja que els valors d'aquesta variable no canvien de manera significativa al llarg de cada hora de predicción.

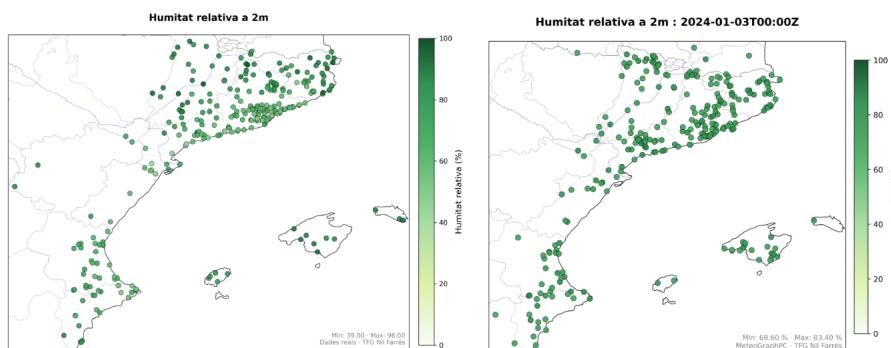


Figura 20: Comparativa de la humitat observada (esquerra) i predita (dreta) a les 00 UTC del dia 3 de gener de 2024.

A nivell general, es constata que el model reproduceix de manera raonable la distribució espacial mitjana de la humitat relativa. Tant les observacions com les prediccions mostren valors elevats de humitat (tons verdosos i més aviat foscos) a gran part del territori, i una distribució de valors que correspon amb la situació meteorològica pròpia d'una matinada d'hivern.

No obstant això, s'identifiquen algunes diferències i limitacions. El rang de valors predicts pel model és més estret (*Min: 68.60%*, *Max: 83.40%*) respecte al rang observat a les dades reals (*Min: 38.00%*, *Max: 96.69%*), evidenciant que la predicción tendeix a suavitzar tant els mínims com els màxims. Aquesta moderació fa que no es captin adequadament situacions d'humitat molt baixa (per exemple, a l'interior de la Vall de l'Ebre o algunes zones costaneres), ni tampoc condicions d'humitat molt elevada pròpies de valls o zones enclotades.

Aquest comportament de “centratge” al voltant de la mitjana és típic dels models de machine learning quan la variable presenta poca variabilitat absoluta i una distribució asimètrica o amb valors extrems poc freqüents.

Pressió atmosfèrica: A la Figura 21 s'hi observen les prediccions de la variable de pressió atmosfèrica i les dades reals per a la primera seqüència de test de 2024. Concretament, es prediuen

les primeres 6 hores del dia 3 de gener de 2024, però aquí només es mostra la primera ja que els valors d'aquesta variable no canvién de manera significativa al llarg de cada hora de predicción.

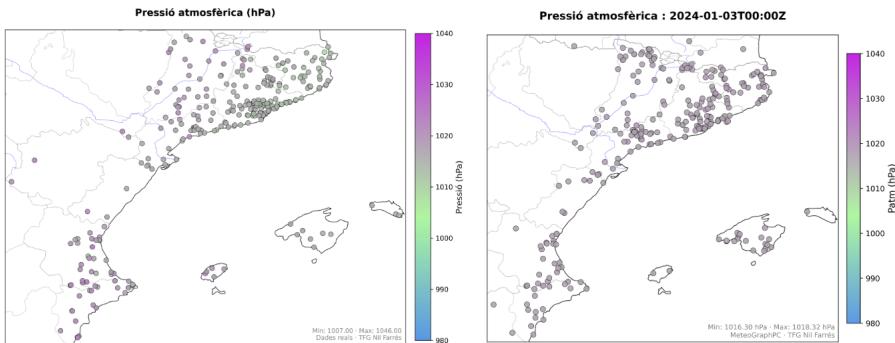


Figura 21: Comparativa de la pressió atmosfèrica observada (esquerra) i predicta (dreta) a les 00 UTC del dia 3 de gener de 2024.

S'observa que el model reproduceix de manera raonablement acurada la distribució mitjana de la pressió atmosfèrica. Tant el rang de valors com la gradació cromàtica de la pressió són força similars entre el mapa real i el de la predicción, i no es detecten discrepàncies notables en el patró regional.

No obstant això, cal destacar que el model tendeix a reduir la variabilitat extrema. La predicción mostra un rang més estret de valors (Min: 1016.30 hPa, Max: 1018.32 hPa) en comparació amb les observacions reals (Min: 1007.00 hPa, Max: 1046.00 hPa). Aquesta moderació fa que els màxims i mínims absoluts de pressió no quedin reflectits de manera fidedigna a la predicción, tot i que el patró general i els gradients són similars.

Aquest comportament és habitual en models automàtics quan la variable mostra una variabilitat limitada i una forta dependència de la situació sinòptica general, com és el cas de la pressió atmosfèrica a escala regional.

Velocitat de vent: A la Figura 22 s'hi observen les prediccions de la velocitat del vent i les dades reals per a la primera seqüència de test de 2024. En aquest cas, però, com que el dia 3 de gener de 2024 no es va registrar cap episodi de vent intens, s'ha escollit el dia 20 de novembre ja que va ser un dia ventós a Barcelona. S'ha escollit, concretament, les 17 hores per a fer-ne l'anàlisi.

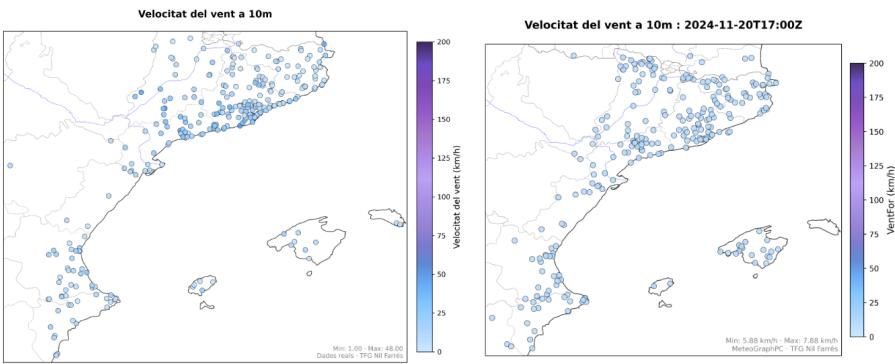


Figura 22: Comparativa de la velocitat del vent observada (esquerra) i predicta (dreta) a les 00 UTC del dia 3 de gener de 2024.

En primer lloc, es constata que el model tendeix a infraestimar la variabilitat i els valors extrems de la velocitat del vent. El rang de valors de la predicción (Min: 5.88 km/h, Max: 7.88 km/h) és molt més estret que el de les dades observades (Min: 1.00 km/h, Max: 48.00 km/h), i la majoria de punts predictius es concentren en valors baixos i similars.

Aquest tipus de comportament és típic dels models estadístics en la predicción de variables fortament esporàdiques i asimètriques, com és el cas del vent, especialment quan la base de dades està

dominada per valors baixos o nuls i només de forma esporàdica hi ha pics elevats.

Precipitació: A la Figura 23 s’hi observen les prediccions de la variable de precipitació acumulada horària i les dades reals per a una seqüència de test de 2024. En aquest cas, però, com que el dia 3 de gener de 2024 no es va registrar precipitació, s’ha escollit el dia 29 d’abril ja que va ser el dia més plujós de l’any 2024 a Barcelona. Només es mostra la primera d’aquest dia ja que la predicció que en fa el model és sempre molt propera a 0mm.

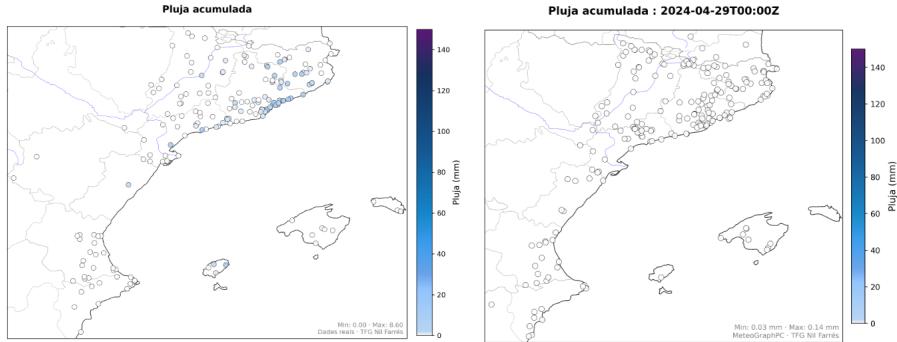


Figura 23: Comparativa de la precipitació acumulada observada (esquerra) i predita (dreta) a les 00 UTC del dia 29 d’abril de 2024.

Aquí es posa de manifest una limitació molt clara del model en la predicció d’episodis de precipitació. A les dades observades, es registren diversos punts amb acumulacions superiors als 8 mm i una distribució espacial molt irregular, especialment al litoral i prelitoral central i nord-est. En canvi, el model pràcticament no prediu precipitació significativa en cap estació: el màxim de la predicció arriba a 0.14 mm i la gran majoria de punts es mantenen a valors pràcticament nuls.

Això evidencia que el model, tal com es pot veure també a l’anàlisi numèrica de mètriques, no és capaç de captar ni la localització ni la intensitat dels episodis de pluja, probablement a causa de la forta esporadicitat i la preeminència de zeros en la sèrie de precipitació, així com la manca d’informació suficient sobre patrons dinàmics o condicions prèvies.

Aquest comportament és habitual quan la variable objectiu és escassa, altament asimètrica i amb valors extrems poc freqüents. Com a resultat, el model tendeix a preveure que no plou per defecte i només excepcionalment s’arrisca a predir valors positius.

6.2 Predicció a mitjà termini

A continuació es presenten els resultats obtinguts amb el model MeteoGraphPC executat amb grups seqüències per a la predicció a mitjà termini.

En aquest segon cas, les seqüències utilitzades durant l’entrenament han estat construïdes de la següent manera:

- Finestres temporals de 120 hores (és a dir, 5 dies en total), un stride de 12 hores i un horitzó de predicció de també 120 hores (5 dies).
- Grups de 25 seqüències i no de 50 (ja que es tracta de seqüències més llargues que ocupen molts més gigabytes) per tal d’optimitzar la càrrega i la gestió de memòria.

6.2.1 Anàlisi de les mètriques

A continuació, s’analitza l’evolució de les principals mètriques d’avaluació (loss, MAE, R^2 , RMSE i SMAPE) durant l’entrenament i la validació del model durant la predicció a mitjà termini.

Loss: Durant les 20 èpoques d'entrenament la *loss* (Figura 24) disminueix de $\approx 0,415$ a $\approx 0,378$ i en validació passa de $\approx 0,332$ a un mínim de **0,3005** a l'època 19 (model escollit). La major reducció, però, es produeix fins a l'època 10 i després la corba s'aplana.

En comparació amb la predició a curt termini (horitzó de 6 h), on el millor *loss* de validació és de 0,2339 a l'època 30, l'augment fins a 0,3005 era esperable atès que predir a 120 h (a 5 dies vista) introduceix més incertesa. Per tant, l'optimització és estable en ambdós casos (curt i mitjà termini), sense signes clars de sobreajustament, i la dificultat per reduir la *loss* creix amb l'horitzó de predició.

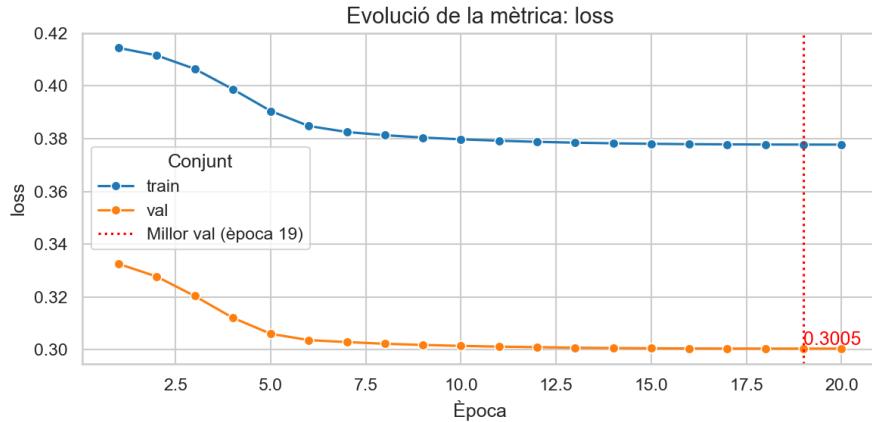


Figura 24: Evolució de la mètrica de loss durant l'entrenament i la validació en la predició a mitjà termini.

MAE: El gràfic d'evolució del MAE (Figura 25) mostra com aquest disminueix de $\approx 0,35$ fins a $\approx 0,31$ durant l'entrenament i de $\approx 0,75$ a $\approx 0,66$ durant la validació, amb la reducció més intensa fins a l'època 7 i posterior estabilització.

La línia discontinua a 1,8604, indica un desajust important entre la validació i el test. De fet, en comparació amb la predició a curt termini, on el MAE de validació és clarament inferior (0,52), l'augment fins a 0,66 és coherent amb l'increment de la incertesa al preveure a 120 h. Per tant, l'aprenentatge és estable i sense signes evidents de sobreajustament, però el MAE creix amb l'horitzó de predició i el test revela un marge de generalització limitat.

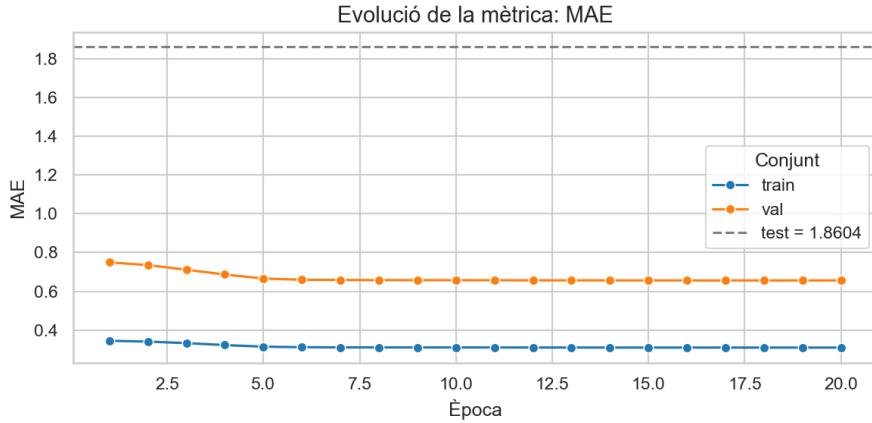


Figura 25: Evolució de la mètrica MAE durant l'entrenament i la validació en la predició a mitjà termini.

Coefficient de determinació R^2 : En aquest cas (Figura 26), el valor de R^2 augmenta de $\approx 0,408$ a $\approx 0,429$ durant l'entrenament i passa de $\approx 0,366$ a $\approx 0,392$ al llarg de la validació. La generalització al test, però, és limitada.

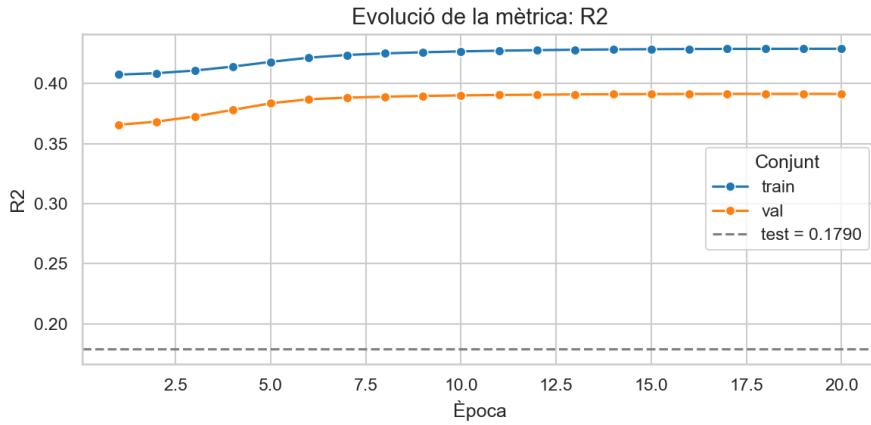


Figura 26: Evolució del coeficient de determinació R^2 durant l'entrenament i la validació en la predicció a mitjà termini.

RMSE: La Figura 27 El gràfic de l'RMSE mostra una convergència estable sense signes d'overfitting. La diferència entre validació i test, però, indica (tal i com s'ha vist a les altres mètriques) la necessitat de reforçar la robustesa del model.

Tant en la predicció a curt com a mitjà termini, el valor de RMSE de test es troba sempre molt per sobre dels valors de validació i entrenament, cosa que indica un cert desajust en la capacitat de generalització del model.

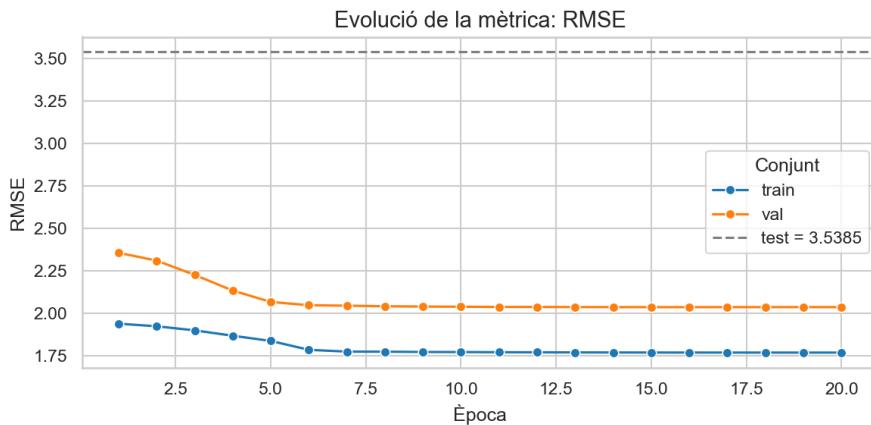


Figura 27: Evolució de la mètrica RMSE durant l'entrenament i la validació en la predicció a mitjà termini.

SMAPE: La disminució inicial de la SMAPE (Figura 28) és notable, però s'aplana a partir de l'època 12. El valor elevat en test indica que caldria considerar estratègies addicionals (com per exemple un augment de les dades d'entrenament) per tal de millorar la generalització percentual.

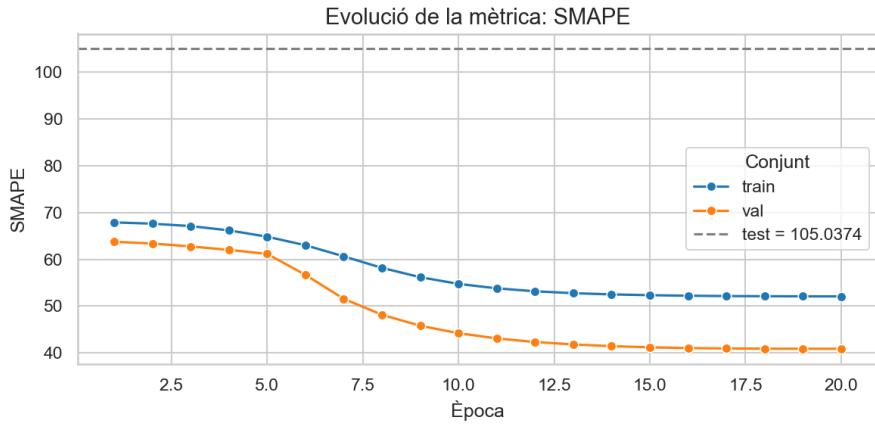


Figura 28: Evolució de la mètrica SMAPE durant l'entrenament i la validació en la predicció a mitjà termini.

Per tant, en conjunt, l'evolució de les mètriques durant la predicció a mitjà termini és similar a les de curt termini, on el model no és capaç de generalitzar correctament a noves dades meteorològiques durant el test. De tota manera, en aquest cas els errors són més elevats a causa d'un horitzó de predicció més llarg que fa que s'introdueixi més incertesa.

Seguidament es fa una comparativa de les mètriques per a cada variable objectiu durant el test de la predicció a mitjà termini:

Variable	RMSE	MAE	R ²	SMAPE	RMSE (pers.)	RMSE (clima)
Temp	5.59	4.34	0.397	1.50	5.84	7.29
Humitat	0.18	0.15	0.052	22.11	0.22	0.18
Pluja	0.20	0.06	0.0003	198.69	0.27	0.20
VentFor	1.73	1.22	0.005	64.67	2.05	1.73
Patm	7.22	5.47	0.082	111.29	7.04	7.53
Vent_u	1.82	1.24	0.026	163.23	2.38	1.85
Vent_v	1.80	1.18	0.016	173.77	2.34	1.82

Table 3: Mètriques de rendiment per a cada variable en la predicció a mitjà termini (model vs. línia base de persistència i climatologia).

A la Figura 29 es presenten les mètriques d'avaluació (MAE, R² i SMAPE) per cadascuna de les variables meteorològiques. S'hi observa que el model obté els millors resultats en la predicció de la temperatura (*Temp*), amb un R² moderat i un error relativament baix. Per a la resta de variables, la capacitat explicativa cau dràsticament (R² < 0.1), especialment per a *Pluja* i components de vent, els quals presenten errors percentuals elevats (SMAPE > 60)

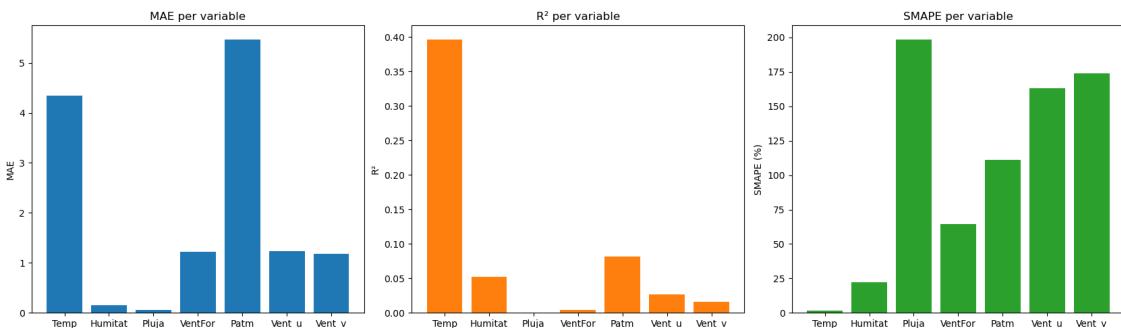


Figura 29: Mètriques d'avaluació per variable meteorològica: MAE, R² i SMAPE (%).

A la Figura 30 es mostra la comparativa directa de l'error RMSE del model respecte als baselines de persistència i climatologia. Es pot observar que:

- El model supera o iguala la línia base de climatologia per a totes les variables (la barra verda és la més alta).
- També supera la persistència en totes les variables excepte la pressió atmosfèrica ($Patm$), on l'RMSE del model és lleugerament superior al de persistència.

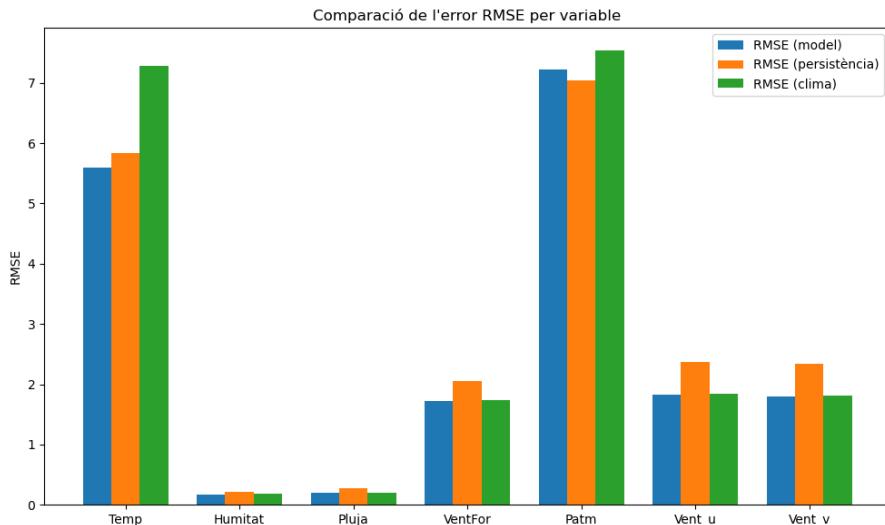


Figura 30: Comparació de l'error RMSE del model amb els baselines de persistència i climatologia per variable.

Aquestes resultats confirmen que, tot i que el model és capaç de capturar força bé les variables contínues (temperatura, humitat), li costa generalitzar en fenòmens dispersos com la precipitació o amb dinàmiques complexes com el vent. Mentre que a la prediccio a curt termini el model no superava la persistència en cap variable, a mitjà termini sí que la supera a totes excepte $Patm$, reflectint una millora relativa malgrat la major dificultat temporal.

6.2.2 Anàlisi comparativa de les correlacions: prediccio vs. dades reals

A continuació es presenten les dues matrius de correlació entre variables meteorològiques per a la prediccio a mitjà termini: una obtinguda a partir de les prediccions generades pel model MeteoGraphPC i l'altra a partir de les observacions reals.

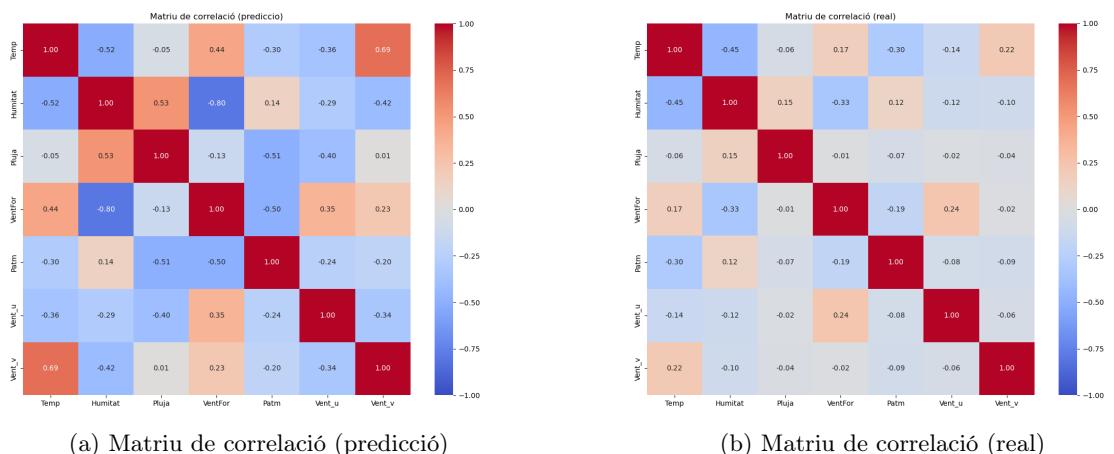


Figura 31: Comparativa de les correlacions entre variables meteorològiques en les prediccions i en les dades observades durant la prediccio a mitjà termini.

En aquest cas, podem observar que les correlacions de les prediccions han augmentat respecte a la prediccio a curt termini. En comparacio amb el curt termini, les matrius de la prediccio a mitjà termini mostren una tendència general a valors absoluts més elevats en les correlacions de

les prediccions. Així, correlacions com la de Temp–Vent _v o la de Humitat–Pluja apareixen més fortemet correlacionats en el model que en les dades reals.

Tot i que el model manté el signe correcte de la majoria de relacions (per exemple, la negativa entre temperatura i humitat, o la positiva entre pluja i humitat), l'amplificació de la magnitud suggerix una certa suavització excessiva i una reducció de l'heterogeneïtat temporal. Un ajust a la regularització o incrementar la diversitat de mostres d'entrenament possiblement permetria reduir aquesta amplificació i millorar la fidelitat a les correlacions observades.

6.2.3 Anàlisi de les prediccions

Per tal d'analitzar de manera preliminar les prediccions a **mitjà termini** (horitzó de 120 h) generades pel model *MeteoGraphPC*, s'han creat, de nou, els diversos mapes de comparació *predicció vs. observació* per a cada variable meteorològica objectiu. Tot i que l'horitzó de predicció és molt més llarg, els patrons espacials predictius continuen sent semblants als predictius a curt termini; però, tal com esperàvem, l'error comès en les prediccions augmenta ja que hi ha més incertesa. La variable que varia més respecte a la predicció a curt termini, és la pressió atmosfèrica, que augmenta considerablement el seu error en aquest cas (RMSE) i prediu uns valors més baixos del que tocaria en general, per exemple, la primera hora del dia 6 de gener de 2024.

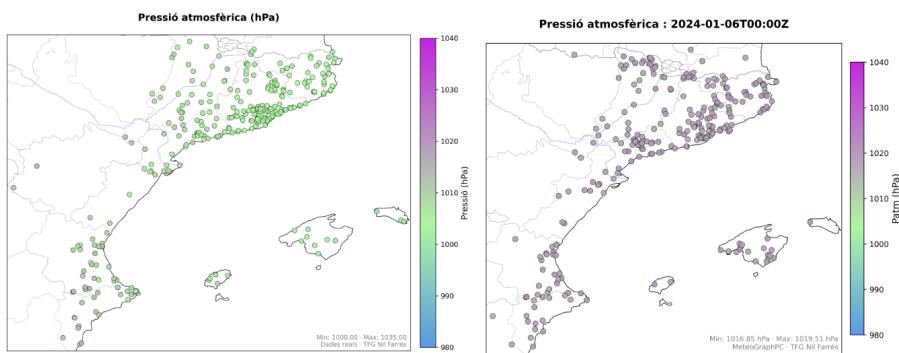


Figura 32: Comparativa de la pressió atmosfèrica observada (esquerra) i predita (dreta) a les 00 UTC del dia 6 de gener de 2024.

La resta de variables segueixen la mateixa tendència que ja observàvem a la predicció a curt termini on la variable de Pluja és la que té un error en la predicció més elevat, cosa que fa que predigi sempre valors molt propers a zero.

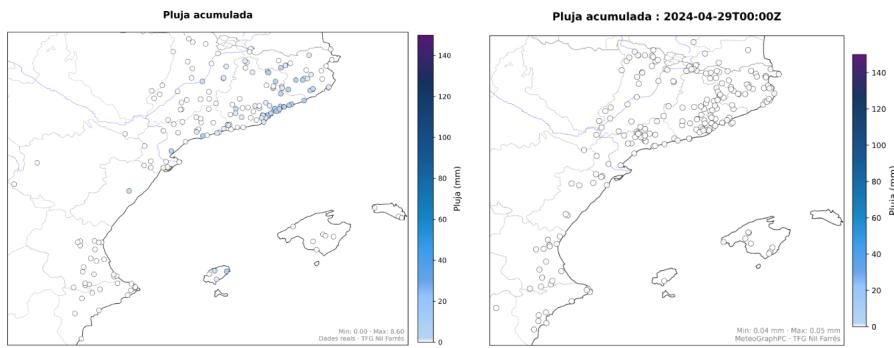


Figura 33: Comparativa de la precipitació observada (esquerra) i predita (dreta) a les 00 UTC del dia 6 de gener de 2024.

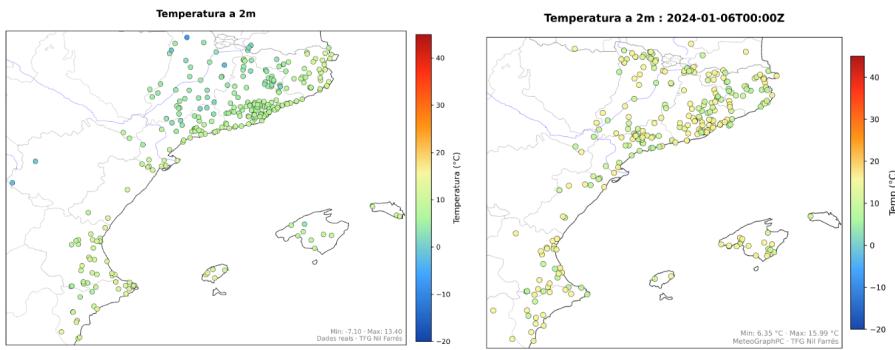


Figura 34: Comparativa de la temperatura observada (esquerra) i predita (dreta) a les 00 UTC del dia 6 de gener de 2024.

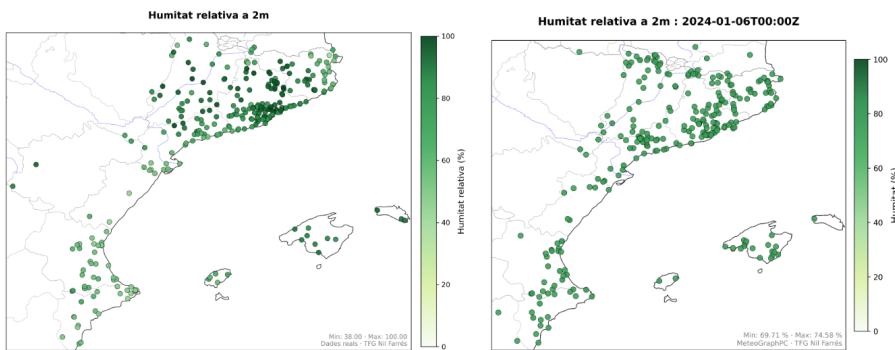


Figura 35: Comparativa de la humitat observada (esquerra) i predita (dreta) a les 00 UTC del dia 6 de gener de 2024.

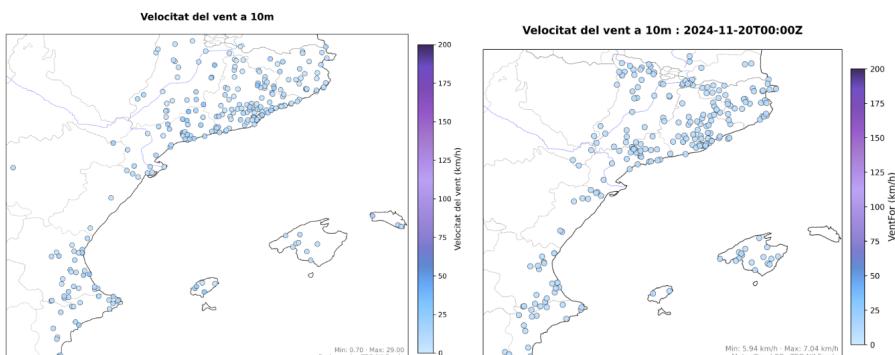


Figura 36: Comparativa de la velocitat del vent observada (esquerra) i predita (dreta) a les 00 UTC del dia 6 de gener de 2024.

7 Conclusions

7.1 Discussió dels resultats

Les prediccions dutes a terme en aquest treball han permès avaluar el rendiment del model **MeteoGraphPC** a curt i a mitjà termini per a diverses variables meteorològiques. L'evolució de les mètriques d'entrenament i validació mostra un *aprenentatge estable*, sense evidències d'overfitting, però també posa de manifest algunes limitacions en la capacitat de generalització cap al conjunt de test (any 2024).

- **Temperatura i pressió atmosfèrica:** el model mostra el millor comportament predictiu en aquestes dues variables, amb valors elevats de R^2 i errors absoluts baixos. Tot i això, el model no supera la línia base de persistència en la predicció a curt termini, que és especialment difícil de batre, però sí que millora clarament la de climatologia tant en curt com en mitjà termini, evidenciant que aprofita informació temporal i espacial rellevant.
- **Humitat relativa:** s'obté un error absolut baix, però la correlació també és limitada, possiblement per la menor variabilitat absoluta d'aquesta variable.
- **Vent i components del vent:** els errors són moderats i la capacitat explicativa del model és limitada, però lleugerament superior a la línia base de climatologia. S'observa una tendència a infraestimar els valors extrems i a predir valors més propers a la mitjana.
- **Precipitació:** es confirma la dificultat inherent a la predicció d'aquesta variable. El model mostra un SMAPE extremadament alt i un R^2 pràcticament nul, la qual cosa posa de manifest la incapacitat per captar la localització i la intensitat dels episodis de pluja. Això es deu principalment a l'alta esporadicitat i a la dominància de valors zero a les sèries de precipitació.
- Un horitzó de predicció més elevat implica un augment de l'error del model, tal i com s'ha vist a la predicció a mitjà termini, ja que s'incrementa més la incertesa a les prediccions.
- El model no té prou en compte l'orografia dels Països Catalans a l'hora de realitzar les prediccions i això fa que es facin prediccions pròximes a la mitjana a totes les estacions meteorològiques.

Per tant, les prediccions a curt i a mitjà termini demostren que el model MeteoGraphPC pot extreure i aprendre informació meteorològica rellevant, però encara mostra dificultats per tenir en compte correctament l'orografia i per generalitzar adequadament sobre dades noves, especialment en variables esporàdiques o amb valors extrems poc freqüents. Per totes les variables, tant en predicció a curt com a mitjà termini, el model supera la línia base de climatologia. Tot i així, la línia base de persistència només és superada per poc pel model quan realitza prediccions a mitjà termini, exceptuant la variable de pressió atmosfèrica. A molt curt termini, la persistència representa una base molt competitiva, especialment en variables estacionàries o amb canvis graduals.

7.2 Aprendentatges adquirits

Al llarg de l'elaboració d'aquest treball, s'ha assolit una formació tècnica i metodològica notable en àmbits diversos de la ciència de dades i la meteorologia computacional. S'ha après a gestionar grans volums de dades meteorològiques provinents de diverses fonts, implementant estratègies robustes de filtratge, depuració, imputació de valors i paral·lelització de tasques per optimitzar el temps de processament.

- **Graph Neural Networks (GNNs):** s'ha aprofundit en el funcionament d'aquest tipus d'arquitectures, especialment per a la predicció de dades espacials i temporals, valorant-ne la seva capacitat per modelitzar relacions complexes entre nodes d'una xarxa meteorològica.
- **Tractament i estructuració de dades meteorològiques:** s'ha adquirit experiència en la gestió de grans conjunts de dades meteorològiques reals amb alguns valors amb errors i estacions inactives, establint protocols de depuració, interpolació i validació amb fonts de confiança.

- **Processament temporal i generació de seqüències:** s'ha treballat amb finestres mòbils, horitzons de predicción i estructures temporals dinàmiques per generar seqüències adaptades a l'entrenament de models seqüencials sobre grafs.
- **Programació avançada en Python:** s'han millorat competències amb biblioteques com per exemple pandas, NumPy, matplotlib, netCDF4, cartopy i especialment PyTorch Geometric, així com en l'organització modular i reutilitzable del codi.

A més, s'ha adquirit experiència pràctica en l'ús d'infraestructures computacionals avançades i s'han desenvolupat habilitats per a la resolució d'incidències relacionades amb incompatibilitats de llibreries, gestió de recursos i automatització de processos mitjançant scripts i fitxers bash. També s'ha adquirit una visió més crítica sobre els límits, avantatges i reptes de l'aprenentatge profund aplicat a la meteorologia, especialment pel que fa a la interpretació dels resultats i la gestió de la incertesa.

7.3 Reptes, dificultats tècniques i metodològiques

Com en tot projecte de gran abast, del qual no se'n tenen gaires precedents, és molt difícil poder dimensionar inicialment i de manera correcte un marc de treball i unes prestacions computacionals adequades. Per aquest motiu i d'altres, en el desenvolupament d'aquest treball, s'han produït diversos contratemps i inconvenients que han condicionat el ritme, l'abast i l'exactitud dels resultats de la recerca. Aquestes dificultats han estat de caràcter tècnic, metodològic i també vinculades a la naturalesa de les dades utilitzades. De tota manera, moltes d'elles s'han pogut resoldre.

A continuació, es detallen els principals obstacles detectats:

1. Filtratge i adequació de les dades meteorològiques

És freqüent en l'àmbit de les ciències de dades que el processament ocipi la gran part del temps de desenvolupament i aquest treball no n'és l'excepció. Les dades meteorològiques utilitzades presentaven diferents mancances que n'han dificultat el tractament i l'anàlisi. Per una banda, en els fitxers CSV amb les dades originals, s'ha observat una presència notable de camps sense valors, especialment els que fan referència a variables de períodes clau per l'entrenament del model. D'altra banda, segurament per problemes tècnics o de desconexions, en determinats períodes desapareixen les files de les estacions meteorològiques afectades. També s'han identificat errors en les coordenades geogràfiques (latituds i longituds incorrectes) i valors anòmals en les mesures, fet que ha obligat a realitzar una depuració exhaustiva i fer imputació de valors.

2. Limitacions de recursos computacionals

La limitació en els recursos computacionals, especialment pel que fa a la memòria RAM, a la capacitat d'emmagatzematge i a la velocitat de transferència dels discs disponibles a les màquines inicialment assignades, ha suposat un dels principals entrebancs per al processament de grans volums de dades meteorològiques i ha penalitzat l'entrenament del model creat. Per pal·liar aquestes mancances, ha calgut ajustar a la baixa els requeriments necessaris per a prediccions meteorològiques més fiables.

3. Incompatibilitats entre llibreries i sistemes

L'entorn d'alguns clústers utilitzats inicialment ha presentat incompatibilitats entre la versió del sistema operatiu Ubuntu i les llibreries de Python o CUDA necessàries per a l'execució òptima dels diversos scripts que conformen aquest treball. En conseqüència, s'ha hagut d'invertir temps a resoldre dependències, escalar a clústers amb més capacitat de càlcul i reconfigurar part del codi per adaptar-lo als requeriments específics de cada sistema.

En conjunt, aquests contratemps han posat de manifest la importància de disposar, en primer lloc, d'una qualitat acurada de les dades, d'un suport i d'una infraestructura tecnològica adequada. Malgrat aquestes dificultats, el projecte ha suposat un aprenentatge en quant a la resolució autònoma de problemes tècnics que es desconeixien.

8 Treball futur

Aquest treball vol representar una línia prometedora en el desenvolupament de sistemes predictius, basats en intel·ligència artificial, més precisos i adaptats a la realitat territorial dels Països Catalans.

L'aplicació de les Xarxes Neuronals en Grafs a la predicción meteorològica presenta encara més reptes i oportunitats de millora, especialment en un àmbit més local que en els models meteorològics basats en IA de l'actualitat. A continuació es proposen diverses línies de treball futur per aprofundir i ampliar els resultats obtinguts:

- **Revisió de l'estratègia de predicción i de la incorporació de l'altitud a la definició de les arestes:** analitzar possibles millors en l'estratègia de predicción utilitzada pel model facilitant una major coherència entre prediccions futures. A més, revisar com es té en compte l'altitud en la creació de les arestes del graf, per tal que el model incorpori millor aquesta informació en la predicción de variables altament dependents d'aquest factor, com ara la temperatura.
- **Ampliació del conjunt de dades i millora en el processament i la qualitat d'aquestes:** incloure noves fonts de dades meteorològiques de qualitat, com ara dades de radar, atmosfèriques, imatges de satèl·lit o de reanàlisis, així com dades d'estacions de territoris veïns, i millorar-ne el processament. L'increment de la diversitat i la resolució de les dades pot contribuir a una millor detecció de patrons regionals i extrems.
- **Millora de l'arquitectura, els hiperparàmetres i l'entrenament:** investigar nous tipus d'arquitectures de GNNs, buscar una millor combinació de valors d'hiperparàmetres, fer millores en l'optimització, tècniques de regularització, així com l'ús de recursos computacionals més potents per abordar conjunts de dades encara més grans.
- **Gestió de la incertesa i mètriques probabilístiques:** implementar metodologies específiques per a la quantificació de la incertesa en les prediccions (per exemple, models bayesianos, ensembles o calibratge probabilístic), per tal d'obtenir no només prediccions puntuals sinó també intervals de confiança i una millor gestió del risc.
- **Anàlisi d'episodis extrems i fenòmens locals:** centrar esforços en l'estudi i predicción de situacions meteorològiques extremes, com ara temporals de pluja intensa, de vent o onades de calor o de fred. La millora de la detecció precoç d'aquests episodis pot tenir un gran impacte social i econòmic.
- **Integració de models híbrids:** explorar la combinació de MeteoGraphPC amb models físics tradicionals mitjançant enfocaments híbrids, ja sigui a través de la correcció de sortides, la generació de dades sintètiques o la incorporació de restriccions físiques (physics-informed neural networks). Aquest camí pot permetre aprofitar el millor de cada paradigma i millorar la robustesa i la interpretabilitat de les prediccions.

En resum, l'experiència i el coneixement adquirits en aquest treball obren la porta a múltiples línies de recerca i desenvolupament futures, amb l'objectiu de contribuir de manera efectiva a la millora dels sistemes de predicción meteorològica regional als Països Catalans i extrapolar-ho a altres zones. Així doncs, l'aplicació de GNNs pot oferir una eina complementària que ajudi a millorar la precisió i la capacitat predictiva dels models meteorològics tradicionals sempre i quan es disposi d'una gran quantitat de dades meteorològiques fiables i de qualitat amb el màxim període de temps possible, i se'n realitzi un bon processament i ànalisi.

9 Agraïments

En primer lloc, voldria agrair profundament a tot l'equip de la secció de meteorologia de 3Cat per haver-me cedit les dades meteorològiques utilitzades en aquest treball i per haver-me resolt qualsevol dubte quan ho he necessitat.

També voldria expressar el meu agraïment als meus tutors, en Jordi Casas Roma i en Josep Lladós Canet, pel seu suport durant el desenvolupament d'aquest Treball de Final de Grau. Les seves aportacions han estat fonamentals tant en l'aspecte acadèmic com en el professional.

Agraeixo també al Centre de Visió per Computador per proporcionar-me els recursos computacionals necessaris per a l'entrenament dels models de xarxes neuronals. L'accés a aquests recursos ha estat essencial per assolir els objectius d'aquest projecte.

Vull fer un agraïment especial i de tot cor a la meva família i amics, que han estat un pilar fonamental. El seu suport, la seva paciència i encoratjament són sempre imprescindibles.

Finalment, vull expressar la meva gratitud a la Universitat Autònoma de Barcelona i, en especial, al grau en Matemàtica Computacional i Analítica de Dades per les oportunitats i recursos proporcionats al llarg de tot el grau. Aquest treball és el resultat del suport i les eines que he rebut durant aquesta etapa de formació.

10 Bibliografia

- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The Graph Neural Network Model*. IEEE Transactions on Neural Networks, 20(1), 61–80. Recuperat de <https://ieeexplore.ieee.org/document/4700287>
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks*. arXiv preprint arXiv:1806.01261. Recuperat de <https://arxiv.org/abs/1806.01261>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. ISBN: 978-0262035613. Recuperat de <https://www.deeplearningbook.org/>
- Peña, J. C., & Monjo, R. (2019). *Climatologia i meteorologia dels Països Catalans*. Edicions Bromera. ISBN: 978-8490268650.
- Mercader-Carbó, J., Codina, B., Sairouni, A., & Cunillera, J. (2010). *Resultats del model meteorològic WRF-ARW sobre Catalunya, utilitzant diferents parametritzacions de la convecció i la microfísica de núvols*. Tethys, 7, 77–89. Recuperat de https://www.researchgate.net/publication/256745908_0007tethys-12-cat
- Esteban, P. (2014). *Els fenòmens meteorològics extrems a Catalunya*. Servei Meteorològic de Catalunya. Recuperat de <https://www.meteotecadecatalunya.cat/Meteoteca/fenomens-meteorologics-violents-a-catalunya/>
- Martín-Vide, J. (2003). *La climatologia de Catalunya: Aspectes generals i aplicats*. Edicions Universitat de Barcelona. ISBN: 978-8447527334.
- Cunillera, J., & Sairouni, A. (2007). *Verificació dels models de mesoescala operatius al Servei Meteorològic de Catalunya*. Notes d'Estudi del Servei Meteorològic de Catalunya, 71. Recuperat de https://www.gencat.cat/mediamb/publicaciones/monografies/verificacio_models_mesoescala.pdf
- Casellas, E., & Rius, A. (2017). *Meteorologia i canvi climàtic als Països Catalans*. Publicacions de l'Abadia de Montserrat. ISBN: 978-8498839644.
- DeepMind. (2024). *GraphCast: AI model for faster and more accurate global weather forecasting*. Recuperat de <https://deepmind.google/discover/blog/graphcast-ai-model-for-faster-and-more-accurate-global-weather-forecasting/>

- DeepMind. (2024). *GenCast predicts weather and the risks of extreme conditions with state-of-the-art accuracy*. Recuperat de <https://deepmind.google/discover/blog/gencast-predicts-weather-and-the-risks-of-extreme-conditions-with-sota-accuracy/>
- NVIDIA. (2022). *FourCastNet: Global data-driven weather forecasting model using adaptive Fourier neural operators*. NVIDIA Developer Documentation. Recuperat de https://docs.nvidia.com/deeplearning/modulus/modulus-v2209/user_guide/neural_operators/fourcastnet.html#&referrer=pdf&asset=solution-brief&id=ai-weather
- Microsoft Research. (2024). *Introducing Aurora: The first large-scale foundation model of the atmosphere*. Recuperat de <https://www.microsoft.com/en-us/research/blog/introducing-aurora-the-first-large-scale-foundation-model-of-the-atmosphere/>
- Bi, K., Xie, L., Zhang, H., Chen, X., & altres. (2023). *Accurate medium-range global weather forecasting with 3D neural networks*. *Nature*, 619, 533–538. doi: [doi disponible a la pàgina]. Recuperat de <https://www.nature.com/articles/s41586-023-06185-3>
- ECMWF charts. (n.d.). *charts.ecmwf.int* Recuperat de <https://charts.ecmwf.int/>
- Google Research. (2024). *Fast, accurate climate modeling with NeuralGCM*. Recuperat de <https://research.google/blog/fast-accurate-climate-modeling-with-neuralgcm/>
- PyTorch Geometric Documentation. (2023). *torch_geometric.nn – PyTorch Geometric*. Recuperat de <https://pytorch-geometric.readthedocs.io/en/2.5.1/modules/nn.html>
- Google Research. (2024). *Generative AI to quantify uncertainty in weather forecasting*. Recuperat de <https://research.google/blog/generative-ai-to-quantify-uncertainty-in-weather-forecasting/>
- Servei Meteorològic de Catalunya. (n.d.). *Meteo.cat*. Recuperat de <https://www.meteo.cat/>
- Agència Estatal de Meteorologia. (n.d.). *Agència Estatal de Meteorología – AEMET*. Recuperat de <https://www.aemet.es/ca/portada>
- Meteo Valls d’Aneu. (n.d.). *Meteo Valls d’Aneu*. Recuperat de <http://www.meteovallsdaneu.com/>
- Confederació Hidrogràfica del Júcar. (n.d.). *Sistema Automàtic d’Informació Hidrològica (SAIH)*. Recuperat de <https://www.chj.es/es-es/medioambiente/SAIH/Paginas/Inicio.aspx>
- AVAMET. (n.d.). *Associació Valenciana de Meteorologia*. Recuperat de <https://www.avamet.org/>
- Meteo Pirineus Catalans. (n.d.). *Meteo Pirineus Catalans*. Recuperat de <https://meteopirineuscatalans.com/>
- Meteoclimatic. (n.d.). *Meteoclimatic.net*. Recuperat de <https://www.meteoclimatic.net/>
- Meteoprades. (n.d.). *Meteoprades.net*. Recuperat de <https://www.meteoprades.net/>
- WeatherLink. (n.d.). *WeatherLink Home*. Recuperat de <https://www.weatherlink.com/>
- CuPy. (n.d.). *CuPy: NumPy & SciPy for GPU-accelerated computing with Python*. Recuperat de <https://cupy.dev/>
- Fankhauser, B., Bigler, V., & Riesen, K. (2023). *Graph-Based Deep Learning on the Swiss River Network*. GbR 2023. Recuperat de https://www.researchgate.net/publication/373355882_Graph-Based_Deep_Learning_on_the_Swiss_River_Network
- Fankhauser, B., Bigler, V., & Riesen, K. (2025). *Exploring a Graph Regression Problem in River Networks*. GbR 2025.
- Labonne, M. (2023). *Hands-On Graph Neural Networks Using Python*. Packt Publishing. ISBN: 978-1-80461-752-6. Recuperat de <https://github.com/PacktPublishing/Hands-On-Graph-Neural-Networks-Using-Python>

- Kipf, T. N., & Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks*. International Conference on Learning Representations (ICLR). Recuperat de <https://arxiv.org/abs/1609.02907>
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks*. International Conference on Learning Representations (ICLR). Recuperat de <https://arxiv.org/abs/1710.10903>
- Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, M., Lin, W., & Deng, H. (2019). *T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction*. IEEE Transactions on Intelligent Transportation Systems, 21(9), 3848–3858. Recuperat de <https://arxiv.org/abs/1811.05320>

A Annex: codi font i un dia complet de dades

El codi font emprat en el desenvolupament d'aquest Treball Final de Grau juntament amb un dia complet de les dades originals està disponible públicament al següent repositori de GitHub:
<https://github.com/nilfarres/MeteoGraphPC>