Wind speed prediction using the analog method over complex topography

Odak Plenković, Iris

University of Zagreb
FACULTY OF SCIENCE
Department of Geophysics

Iris Odak Plenković

WIND SPEED PREDICTION USING THE ANALOG METHOD OVER COMPLEX TOPOGRAPHY

DOCTORAL THESIS

Zagreb, 2020

University of Zagreb
FACULTY OF SCIENCE
Department of Geophysics

Iris Odak Plenković

WIND SPEED PREDICTION USING THE ANALOG METHOD OVER COMPLEX TOPOGRAPHY

DOCTORAL THESIS

Supervisors:
dr. sc. Kristian Horvath
Luca Delle Monache, PhD.

Zagreb, 2020

Sveučilište u Zagrebu
PRIRODOSLOVNO-MATEMATIČKI FAKULTET

Iris Odak Plenković

PROGNOZA BRZINE VJETRA UPOTREBOM METODE ANALOGONA NAD SLOŽENOM TOPOGRAFIJOM

DOKTORSKI RAD

Mentori:
dr. sc. Kristian Horvath
Luca Delle Monache, PhD.

Zagreb, 2020

The thesis was made under the supervision of dr. sc. Kristian Horvath and Luca Delle Monache, PhD.

Dr.sc. Kristian Horvath is Head of the Applied Meteorology and Modelling Department at Croatian Meteorological and Hydrological Service and senior research associate. He conducts and overseas applied meteorological research and works day-to-day with end-users in bringing research to applications for the benefit of society and different sectors of economy. Kristian earned his PhD at the Geophysical Department of the Faculty of Natural Sciences University of Zagreb in 2008., following his undergraduate degree at the same University in 2003. His interests include mesoscale meteorology, numerical weather prediction, multi-scale numerical modelling, post-processing of numerical weather prediction data and dynamical downscaling. Kristian has spent over two years abroad on research stays and collaborations at recognized institutions such as Nacional center for Atmospheric Research (USA), Desert Research Institute (USA), Risø; DTU National Laboratory for Sustainable Energy (Denmark) and University of Balearic Islands (Spain). He has conducted or participated on over a dozen of research projects funded by funding agencies from Croatia, EU and USA. He published more than 25 peer-review papers, edited one book and co-authored one book-chapter, as well as prepared close to a hundred of conference contributions. His scientific work was recognized through several international research awards and scholarships as well as two awards for research innovations designed to use numerical predictions to improve energy efficiency. Kristian regularly works as an evaluator for Croatian Science Foundation and European Commission, as well as reviewed papers for over a dozen of scientific journals. He is a permanent member of the external committee of the Hungarian Academy of Science, was an active member of Croatian Committee for Geodesy and Geophysics at Croatian Academy of Sciences and Arts and is a lasting member of Croatian Meteorological Society.

Luca Delle Monache, PhD, is the Deputy Director of the Center for Western Weather and Water Extremes (CW3E), Scripps Institute of Oceanography, University of California San Diego. Dr. Delle Monache oversees the development of the Center's modeling, data assimilation, postprocessing, and artificial intelligence capabilities, with the goal of maintaining state-of-the-art models and tools while actively exploring innovative algorithms and approaches. In close coordination with the Center Director and the management team, he develops new scientific and programmatic strategies to maintain and further expand CW3E leadership on understanding, observing, and predicting extreme events in Western North America.

He earned a Laurea (~M.S.) in Mathematics from the University of Rome, Italy (1997), an M.S. in Meteorology from the San Jose State University, U.S. (2002), and a Ph.D. in Atmospheric Sciences from the University of British Columbia, Canada (2005). His interests include the design of ensemble methods, probabilistic prediction and uncertainty quantification, numerical weather prediction, data assimilation, inverse modeling, post-processing methods including artificial intelligence algorithms, renewable energy, air quality and transport and dispersion modeling. Among his main scientific accomplishments, there is the development during his Ph.D. of the first ensemble for air quality prediction, and later in his career the design of the analog ensemble which has been applied successfully in several fields, and is based on a new paradigm for ensemble design. Luca Delle Monache has been the principal investigator of several multi-institution projects funded by the National Science Foundation, the National Oceanic and Atmospheric Administration, the National Aeronautics and Space Administration, the Department of Energy, the Department of Defense, and the private sector. Before joining CW3E, he was a postdoc and then a staff scientist at the Lawrence Livermore National Laboratory, Livermore, California (2006-2009), and a project scientist and then the Science Deputy Director of the National Security Applications Program at the National Center for Atmospheric Research, Boulder, Colorado (2009-2018).

The research was partially supported by the grant IPA2007/HR/16IPO/001-040507 through the WILL4WIND project (www.will4wind.hr) and by RC LACE.

Curriculum vitae

Iris Odak Plenković is born od 11th of February, 1984 in Metković, Croatia. She graduated in February 2013 at Geophysical Department (Faculty of Science, University of Zagreb), obtaining an M.S. degree (mag.phys.-geophys.) Right after, in April 2013, she started working at Croatian Meteorological and Hydrological Service in Zagreb. Currently, she is the Head of Model Data Processing and Applications Division in the Research and Development Division. The main topic of her research is the development of physically-consistent deterministic and probabilistic wind forecast post-processing methods, mostly the so-called analog-based method. She is selected to be the Croatian Meteorological and Hydrological Service representative in the EUMETNET post-processing module. In 2018, she recieved Young scientist award from Croatian Meteorological Society. Since she started her research in 2013, Iris presented her results at several scientific and professional conferences as oral presentations and/or posters, such as EMS meetings, ICAM conferences, ALADIN and HIRLAM joint meetings, etc. She also published several papers, as listed below.

- Odak Plenković, I., Schicker, I., Dabernig, M., Horvath, K., Keresturi, E., 2020: **Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain**. *Q J R Meteorol Soc.* 2020; 1-19 (https://doi.org/10.1002/qj.3769).
- Odak Plenković, I., Delle Monache, L., Horvath, K., Hrastinski, M., 2018: **Deterministic Wind Speed Predictions with Analog-Based Methods over Complex Topography. *J.*** *Appl. Meteor. Climatol.*, 57, 2047-2070 (https://doi.org/10.1175/JAMC-D-17-0151.1).
- Ivatek-Šahdan, S., Stanešić, A., Tudor, M., Odak Plenković, I., Janeković, I., 2018: **Impact of SST on heavy rainfall events on eastern Adriatic during SOP1 of HyMeX**. *Atmospheric Research*, Volume 200, 36-59 (https://doi.org/10.1016/j.atmosres.2017.09.019.).
- Odak Plenković, I., Delle Monache, L., Horvath, K., Hrastinski, M. i Bajić, A., 2015: **Post-processing of ALADIN wind speed predictions with an analog-based method.** *Hrvatski meteorološki časopis, 50* (50), 121-136 (https://hrcak.srce.hr/155407).

Finally, there is a paper currently in the review:

- Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z.B., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenković, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., Ylhaisi, J., 2020: **Statistical Postprocessing for Weather Forecasts – Review, Challenges and Avenues in a Big Data World**. *Bulletin of the American Meteorological Society (BAMS)* (preprint: https://arxiv.org/abs/2004.06582).

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to the Croatian Meteorological and Hydrological Service that supported my research.

Also, I would like to acknowledge my mentors Dr. Kristian Horvath and Luca Delle Monache for recognizing the potential of the fast-growing world of post-processing and letting me in. Moreover, I thank them for having the confidence in my work and respectfully allowing me to grow and choose the directions where the research takes me while helping me on the way.

Special thanks to Dr. Maja Telišman Prtenjak, Dr. Željko Zečenaj and Dr. Ivan Güettler for reading and improving this dissertation.

I also thank NCAR and ZAMG for accomodating me in such a nice environment for learning. I met some great colleagues there, felt welcomed and motivated to do as much as I can. I thank them all, but especially Irene Schicker and Markus Dabernig for their help.

I thank my colleagues in Zagreb for their help. And, also, for listening to me wining. Especially my former office roommates Mario Hrastinski and Suzana Panežić. Believe it or not, that helped me through the hard days.

I thank my family and friends the most for keeping me somewhat sane. They are the brick I stand on and this would not be possible without them. Even if it would, it would mean nothing.

Finally, I will thank my husband Paulo, who already heard too much regarding this thesis material, for encouraging and supporting me all these years.

Table of Contents

University of Zagreb
Faculty of Science
Department of Geophysics
Doctoral Thesis

ABSTRACT

WIND SPEED PREDICTION USING THE ANALOG METHOD OVER COMPLEX TOPOGRAPHY

Iris Odak Plenković
Croatian Meteorological and Hydrological Service

The performance of the analog-based post-processing method is tested in climatologically and topographically different regions, for point-based wind speed predictions at 10 m above the ground, and compared to the baseline Kalman filter (KF) model. This research shows that the deterministic analog-based predictions produced using deterministic numerical weather prediction (NWP) model output improve the correlation between predictions and measurements while reducing the forecast error compared to the starting model predictions regardless of the terrain complexity. While the KF based approach generally outperforms the analog-based predictions in the bias reduction, the combination of the KF and analog approach can be similarly successful.

In the coastal complex area, characterized by a larger frequency of high wind speed, the analog-based predictions are more successful in reducing the dispersion error than the KF. The application of the KF algorithm to the analogs in the so-called analog space (KFAS) is the least prone to the standard deviation underestimation among the analog-based predictions. All analog-based predictions improve prediction of larger than diurnal motions while the KFAS is superior among all analog-based predictions in predicting alternating wind regimes on the time scales shorter than a day. The analog-based predictions better distinguish different wind speed categories in the coastal complex topography by using a higher-resolution model input.

The analog method is also applied to the ensemble NWP. Evaluation of several configurations using various predictor variables is conducted through a set of sensitivity experiments. The results are compared to the ensemble model output statistic (EMOS) baseline model. Results show that both analog-based and EMOS experiments considerably improve the raw model forecast. The analog-based predictions are overall comparable to or even outperform the EMOS. Assessing the post-processing performance for high wind speeds, it is shown that the analog experiments can improve the raw forecast, exhibiting significantly higher skill than the EMOS. The processes at lower altitude stations seem to be better represented by the raw model, which leads to better input forecast to the post-processing and better overall result than for the mountain stations. Generally, the difference between several analog-based experiments is less pronounced. Furthermore, it is demonstrated that the usage of summarized ensemble measures is an optimal way to improve the forecast skill, compared to the other analog-based experiments.

Keywords: *analog-ensemble forecast, complex topography, ensemble model output statistics, Kalman-filter, mesoscale model, statistical post-processing, wind ensemble forecast*

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Geofizički odsjek
Doktorska disertacija

SAŽETAK

PROGNOZA BRZINE VJETRA UPOTREBOM METODE ANALOGONA NAD SLOŽENOM TOPOGRAFIJOM

Iris Odak Plenković
Državni hidrometeorološki zavod

Metoda analogona, koja se koristi za naknadnu obradu produkata numeričkog modela, testirana je za prognoze vjetra na 10 m iznad tla na lokacijama koje pripadaju topografski i klimatološki različitim područjima te uspoređena s metodom koja koristi Kalmanov filtar (KF). Deterministički produkt metode analogona ima veću koreliranost prognoze i mjerenja te manju pogrešku u odnosu na numerički model koji metoda koristi kao ulazni podatak, neovisno o složenosti topografije. Metoda naknadne obrade KF iznimno je uspješna u uklanjanju pristranosti prognoze. Kombinacija metode analogona i KF gotovo je jednako uspješna u uklanjanju pristranosti, pri čemu pokazuje i dodatne prednosti svojstvene metodi analogona.

U obalnom području, karakteriziranom kompleksnom topografijom i učestalim jakim vjetrom, metoda analogona uspješnija je od KF u uklanjanju pogreške disperzije. Dodatno, primjena Kalmanovog filtra u takozvanom prostoru analogona (KFAS) je eksperiment koji je najmanje podložan podcjenjivanju prirodne varijabilnosti vjetra, mjereno standardnom devijacijom. Svi eksperimenti koji koriste analogije poboljšavaju prognoze na vremenskim skalama duljima od jednog dana. Međutim, na skalama kraćima od jednog dana je KFAS najuspješniji eksperiment. Korištenje modela veće rezolucije kao ulazni podatak za metodu analogona doprinosi da prognoza lakše razlikuje kategorije vjetra.

Metoda analogona primijenjena je i na ansambl prognozu numeričkog modela. Pritom je testirano nekoliko različitih konfiguracija metode kroz testove osjetljivosti. Eksperimenti se prvenstveno razlikuju po ulaznim parametrima, tj. po načinu korištenja informacija iz početne ansambl prognoze modela. Rezultati metode analogona uspoređeni su s metodom naknadne obrade koja je bazirana na statistici simuliranih podataka za ansambl prognoze (EMOS). Obje testirane metode naknadne obrade vidno poboljšavaju prognozu ulaznog modela. Pritom je metoda analogona usporediva s metodom EMOS, ili čak i bolja. Dodatno, metoda analogona ostvaruje signifikantno bolji rezultat za prognozu jakog vjetra od početnog modela te metode EMOS. U numeričkom modelu procesi su bolje razlučeni za lokacije smještene na nižoj nadmorskoj visini nego za planinske lokacije. Posljedično, to znači i bolji rezultat nakon naknadne obrade produkata modela te bolji ukupan rezultat za lokacije nižih nadmorskih visina. Općenito, razlika među eksperimentima s različitim konfiguracijama metode analogona manje je izražena. Štoviše, pokazano je da je upravo korištenje sažetih informacija o prognozi ulaznog modela optimalan način da se poboljša točnost prognoze.

Ključne riječi: *EMOS, Kalmanov filtar, kompleksna topografija, mezoskalni model, metoda analogona, statističke metode naknadne obrade, ansambl prognoza vjetra*

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Geofizički odsjek
Doktorska disertacija

PROŠIRENI SAŽETAK

i. UVOD

Čak i najsuvremeniji prognostički modeli proizvode lokalne pogreške koje se ne mogu zanemariti, posebno pri prognoziranju nad kompleksnom topografijom [Horvath et al., 2012]. Zato je, uz razvoj prognostičkih modela, od izrazite važnosti razviti i dodatne alate koji korištenjem raspoloživih mjerenja smanjuju pogrešku modela, poput metoda naknadne obrade. Jedna od takvih metoda, tzv. metoda analogona, temelji se na desetljećima staroj ideji da se u prognozi koristi analogija s prethodnim situacijama (npr. Lorenz [1969]). Naime, pretpostavka je da će dva inicijalno slična stanja atmosfere neko vrijeme ostati slična. U prošlosti su se u metodi analogona koristile razne formulacije te uspoređivale točkaste prognoze, prognoze polja, mjerenja, analize i dr. U nedavnoj prošlosti razvijena je formulacija koja koristi numeričku prognozu za određenu lokaciju, uspoređuje je s povijesnim prognozama i odabire najsličnije (tzv. analogone) te je pokazala zavidne rezultate [Delle Monache et al., 2011, 2013]. Nakon što se odaberu analogoni, vrijednosti koje su izmjerene u tom terminu u prošlosti formiraju članove ansambla analogona (AnEn) (shema na Slici 1 na str. 7). Ako je model konzistentan u smislu da u sličnim situacijama proizvodi slične pogreške ili propušta predvidjeti procese fine lokalne skale, korištenjem mjerenja u rezultate prognostičkog sustava se uključuju učinci koje model nije u mogućnosti dinamički razlučiti.

ii. NAKNADNA OBRADA DETERMINISTIČKE PROGNOZE

U prvom dijelu ispitana je metoda analogona koja koristi deterministčku prognozu operativnog numeričkog modela Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) [ALADIN International Team, 1997], koji se koristi na Državnom hidrometeorološkom zavodu u Hrvatskoj (Slika 4, str. 14). Pritom je ispitana deterministička prognoza srednjaka (***AN***) i medijana (***ANM***) ansambla analogona. Pošto rezultati prognoze

ANM nisu uspješni kao ***AN*** (npr. Slika 6 na str. 22) te nisu pokazali specifične prednosti u odnosu na ostale prognoze, nisu detaljnije prikazani. Rezultati prognoze ***AN*** uspoređeni su s linearnim, rekurzivnim i prilagodljivim pristupom koji se temelji na primjeni Kalmanovog filtra (KF) [Kalman, 1960; Delle Monache et al., 2011]. Ovaj pristup koristi identične podatke (početnog) numeričkog modela i dostupnih mjerenja kao metoda analogona, producirajući prognozu ***KF***. Dodatno, testirana su dva eksperimenta koji sjedinjuju metode analogona i KF. Prvi se temelji na primjeni KF na vremenskom nizu prognoza ***AN***, rezultirajući prognozom ***KFAN***. Drugi eksperiment primjenjuje KF, no umjesto da koristi vremenski niz prognoza početnog modela, koristi prognoze sortirane po sličnosti s posljednjom prognozom (onom koja se pokušava poboljšati). Tako se formira takozvani prostor analogona te se metoda zove Kalmanov filtar u prostoru analogona (***KFAS***). Shema prognoza ***KFAN*** i ***KFAS*** prikazana je na Slici 5 (str. 17), a ograničenje prognoze ***KF*** kod izrazite varijabilnosti pogreške objašnjeno na Slici 9 (str. 27). Konačno, determinističke prognoze metodom analogona uključuju ***AN,*** ***KFAN*** i ***KFAS***.

U radu se ispituje primjena metode analogona na području karakteriziranom kompleksnom topografijom. U fokusu je obalno područje Hrvatske, gdje se značajan udio mezoskalne energije prenosi strujanjima niz padine prema moru te termički induciranom obalnom cirkulacijom [Grisogono and Belušić, 2009]. Ispitana je primjena metode i nad planinsko-kompleksnom topografijom te ravnicom kontinentalne Hrvatske (Slika 2, str. 9; Slika 3, str. 11).

Evaluacija prognoze brzine vjetra kao kontinuiranog prediktanda

Analizirajući korijen srednje kvadratne pogreške (*RMSE*), koeficijenta korelacije ranga (*RCC*) te pristranosti srednjaka, pokazano je da sve testirane metode naknadne obrade poboljšavaju rezultat operativnog modela ALADIN (Slika 6, str. 22). Pritom su najbolji rezultati postignuti pri korištenju 15 članova AnEn. Korištenjem više od 15 članova uočen je porast pogreške, što je vjerojatno posljedica klimatološke razlike između razdoblja koje se koristilo za učenje metode u odnosu na razdoblje koje se koristilo za verifikaciju.

U radu je pokazano da su eksperimenti ***KF*** i ***KFAN*** najuspješniji testirani pristupi za uklanjanje pristranosti srednjaka (Slika 7, str. 24). Očekivan je to rezultat, jer je KF konstruiran u svrhu uklanjanja sustavne pogreške ako se ona naglo ne mijenja (kod naglih i velikih dnevnih varijacija KF nije jednako uspješan). Uz to, prognoza ***KF*** povećava koeficijent korelacije između prognoze i mjerenja u odnosu na početni model nad relativno

ravnom topografijom u kontinentalnoj Hrvatskoj, gdje postoje indikacije da u prognozama numeričkog modela postoje sustavne pogreške koje utječu na gibanja velike skale (npr. za periode dulje od 10 dana). Međutim, prognoza ***KF*** nije jednako uspješna u uklanjanju nesustavne (disperzijske) pogreške na obalnom području. Za razliku od prognoze ***KF,*** ostale metode naknadne obrade pokazale su se uspješnima i na kompleksnoj topografiji poput obalnog područja. Iako svi pristupi koji koriste analogije pritom pokazuju veliku sposobnost prilagodbe području, u smanjenju nesustavne pogreške, najuspješnija je prognoza ***AN***.

Model ALADIN s horizontalnom razlučivošću od 8 km (***A8***) podcjenjuje prirodnu varijabilnost vjetra nad kompleksnom topografijom (Slika 8, str. 26). Standardna devijacija (σ) prognoze ***KF*** bliža je σ izmjerenih vrijednosti nego je to slučaj kod ***A8***. Podcjenjivanje σ izmjerenih vrijednosti manje je izraženo kod metode analogona na obalnom području. Pritom je prognoza ***AN*** najsklonija podcjenjivanju σ. Razlog je najvjerojatnije razlika u varijabilnosti između razdoblja učenja metode analogona, ali i usrednjavanje koje se koristi pri prognoziranju srednjaka ansambla i djelomično smanjuje prirodnu varijabilnost vjetra. Eksperimenti koji kombiniraju metodu analogona i KF uspješniji su u uklanjanju sustavne pogreške pristranosti standardne devijacije σ od prognoze ***AN***, pri čemu je najuspješnija prognoza ***KFAS***. Različiti eksperimenti prognoze metodom analogona djeluju na različite aspekte početnog numeričkog modela, no u konačnici rezultiraju sličnim smanjenjem pogreške mjerene s *RMSE*. Prednost primjene metode analogona nad primjenom (isključivo) KF posebno se ističe u obalnom području.

Utjecaj početnog numeričkog modela na rezultat nakon naknadne obrade njegovih produkata ispitan je koristeći tri različite konfiguracije operativnog modela ALADIN [Tudor et al., 2013]: dvije verzije s punim paketom fizike i horizontalnom razlučivosti od 8 km (***A8***), odnosno 2 km (***A2***), te model dinamičke adaptacije (***DA***) s horizontalnom razlučivosti od 2 km. U svim ispitanim slučajevima dolazi do poboljšanja rezultata ulaznog modela nakon primjene metoda naknadne obrade (Slika 10, str. 32). Testirana je hipoteza da se korištenjem modela veće razlučivosti, koji je tako u mogućnosti simulirati više fizikalnih procesa, mogu izabrati i kvalitetniji analogoni. Međutim, za rezultate nakon naknadne obrade nije moguće donijeti jednoznačan zaključak. Osim utjecaja samog početnog modela, ovakav rezultat može biti posljedica nesavršenosti postupka pri ocjenjivanju rezultata prognoze. Takve nesavršenosti pri evaluaciji točkaste prognoze, poput velike osjetljivosti verifikacijske metrike na male prostorne i fazne pogreške, posebno se ističu kod modela velike razlučivosti (npr. od

oko 1 km). Analiza zato sadrži i evaluaciju prognoze za različite kategorije brzine vjetra te spektralnu analizu. Korištenje prostornih polja u ocjeni prognoze olakšalo bi identificiranje dodatnih prednosti korištenja modela velike razlučivosti. No, analize adekvatne razlučivosti i točnosti, kojima bi se takve prednosti kvantificirale, nisu dostupne.

Evaluacija prognoze brzine vjetra kao kategoričkog prediktanda

Kategorička verifikacija prognoza brzine vjetra provedena je koristeći vrijednost 50.-og i 90.-og percentila za identifikaciju tri kategorije vjetra: slab, umjeren i jak. Polihorički koeficijent korelacije (*PCC*; Slika 11, str. str 35) mjeri asocijaciju koristeći tablicu kontingencije (Tablica 3, str. 34) [Juras i Pasarić, 2006]. Mjera *PCC* pokazuje da modeli veće razlučivosti (***A2*** i ***DA***) bilježe i veću asocijaciju s mjerenjima u obalnom području, no to nije slučaj za ostale tipove topografije (Slika 12, str. 38). Osim prognoze ***KF*** nad obalno-kompleksnom topografijom, sve metode naknadne obrade povećavaju asocijaciju prognoze i mjerenja. U prosjeku, metoda analogona ostvaruje bolji rezultat od prognoze ***KF***, pri čemu najbolji rezultat ostvaruje ***AN***.

Nad obalno-kompleksnom topografijom prognoza ***A2*** je nepristrana za sve kategorije vjetra (Slika 13, str. 40). Ostala dva modela podcjenjuju učestalost pojave jakog vjetra (model ***DA*** jakog i slabog vjetra), dok precjenjuju učestalost umjerenog vjetra. Nad ostalim tipovima topografije svi modeli podcjenjuju učestalost slabog, a precjenjuju učestalost umjerenog i jakog vjetra. Nakon primjene bilo koje metode naknadne obrade, u prosjeku se smanjuje pristranost pri prognoziranju klimatološki učestalih kategorija (slab i umjeren vjetar). Međutim, podcjenjivanje učestalosti kategorije jakog vjetra predstavlja najveći izazov za metodu analogona. Prognoza ***KFAS*** čini se pritom najmanje pristranom među eksperimentima metode analogona u kategoriji jakog vjetra, dok je za ostale kategorije gotovo jednako nepristrana kao ***AN***. Konačno, rezultati za prognozu ***KF*** pokazuju manju pristranost u ovoj kategoriji vjetra. Ovi rezultati samo su indikacija određenih karakteristika, jer su zbog veličine uzorka intervali pouzdanosti veliki.

Iako pristranost daje informaciju o (ne)adekvatnoj razdiobi, ne podrazumijeva i točnost prognoze. Zato je korišten kritični indeks uspjeha (*CSI*), mjera relativne točnosti za prognoze kategoričkog tipa [Wilks, 2011; Jolliffe and Stephenson, 2011]. Rezultati pokazuju da prognoza ***KF*** ima vidno veću relativnu točnost od početnih modela u gotovo svim testiranim slučajevima nad relativno ravnom kontinentalnom i planinsko-kompleksnom topografijom, no to nije slučaj i na obalnom području (Slika 14, str. 43). Korištenje analogona rezultira još višim vrijednostima, pokazujući veću relativnu točnost i od prognoze ***KF***. Iznimka je

prognoza jakog vjetra u kontinentalnoj Hrvatskoj. U tom slučaju najbolji je rezultat prognoze ***KF,*** što sugerira sustavnu pogrešku modela pri prognozi jakog vjetra. Od eksperimenata metodom analogona, ***AN*** pokazuje najbolji rezultat u kategoriji slabog vjetra, dok su ***KFAN*** i ***KFAS*** uspješniji u ostale dvije kategorije. Može se primijetiti da korištenje veće rezolucije početnog modela dovodi do povećanja relativne točnosti kod prognoziranja jakog vjetra u obalnom području. Međutim, efekt nad ostalim vrstama topografije nije jednoznačan. Iako su razlike među eksperimentima koji koriste različit početni model manje nego za (neobrađene) prognoze modela, ipak se međusobne razlike među modelima zadržavaju i nakon primjene metoda naknadne obrade na njihovim produktima. Rezultati, posebno u kategoriji jakog vjetra, temeljeni su na relativno malom uzorku, a i mjera relativne točnosti *CSI* je osjetljiva na klimatološku učestalost pojave određene kategorije pa ih treba razmatrati kroz okvir određene nepouzdanosti.

Jedan od načina kako poboljšati pouzdanost rezultata je povećati veličinu uzorka. Međutim, to je u suprotnosti s osnovnom idejom pri korištenju metoda naknadne obrade – da je metoda brza i efikasna, ali i jednostavna za implementaciju. Korištenje duljih nizova zahtjeva više vremena za izračun. Dodatno, pri svakoj promjeni modela potrebno je reproducirati povijesne prognoze, što je računalno zahtjevan postupak koji se u praksi rijetko radi za dulje razdoblje. Postupak treba ponoviti kod sljedećeg ažuriranja modela, što u praksi najčešće ne traje dugo (do par godina, no često kraće). Alternativno, može se koristiti verifikacijska mjera koja je posebno razvijena za evaluaciju rijetkih i ekstremnih događaja – indeks koji ovisi o ekstremima (*EDI*). Ovaj indeks nije, poput mjre relativne točnosti *CSI*, osjetljiv na klimatološku učestalost pojave određene kategorije (npr. jakog vjetra). Rezultati indeksa *EDI* u skladu su s prethodnim rezultatima, pri čemu su intervali pouzdanosti manji (Slika 15, str. 44). Metoda analogona, u prosjeku, postiže bolji rezultat od prognoze ***KF*** te prognoze numeričkog modela, pri čemu je najbolji rezultat prognoze ***KFAN***. Rezultat je bolji ako se koristi model sa svim potrebnim parametrizacijama i većom razlučivošću (***A2***), što u skladu s prethodnim rezultatima.

Spektralna analiza prognoze brzine vjetra

Spektralnom analizom jasno je potvrđena pretpostavka da je primjena (isključivo) metode KF ograničena na gibanja velike skale (npr. periode dulje od 10 dana) kad postoji pristranost u spektru snage prognoze početnog modela (Slika 16, str. 48). Drugim riječima, prognoza ***KF*** povećava energiju gibanja velike skale u obalnom te ju smanjuje u kontinentalnom području,

no samo za periode veće od 10 dana. Zato je spektar snage prognoze ***KF*** gotovo identičan spektru početnog modela, a spektar ***KFAN*** gotovo identičan spektru ***AN***. Moguće je bolje parametrizirati KF, no i u tom slučaju za očekivati je da nema efekta na kraće vremenske skale (npr. na period od 1 dan ili manje). Selektivno uključivanje mjerenih vrijednosti u metodi analogona vodi do boljeg prognoziranja na skalama duljim od 1 dan (LTD, od "longer-than-diurnal" [Horvath et al., 2012]) u odnosu na početni model (Slika 17, str. 49). Skala LTD je bitnija od velike (tj. veće od 10 dana) za prognoze do 72 prognostička sata. Na skali LTD metoda analogona smanjuje podcjenjivanje energije u odnosu na početni model u obalnom i precjenjivanje energije u kontinentalnom području. Ako se uzme u obzir i utjecaj metode naknadne obrade na skale kraće od 1 dan (STD, od "shorter-than-diurnal" [Horvath et al., 2012]), prognoza ***KFAS*** superiorna je ostalim eksperimentima. Razlog je što ***KFAS*** na skali LTD smanjuje pristranost spektra snage početnog modela jednako učinkovito kao prognoza ***AN***, ili čak bolje. Uz to, ***KFAS*** za skale STD zadržava energiju simuliranih gibanja početnog modela. Zbog toga je manje sklona podcjenjivanju energije male skale od, primjerice, prognoza ***AN*** i ***KFAN***. Sve metode naknadne obrade adekvatno prognoziraju amplitudu harmonika dnevnog hoda (24 h, 12 h, 8 h periodi), slično kao i početni model.

Korištenje veće horizontalne rezolucije u početnom modelu općenito generira više energije u spektru (Slika 18, str. 52). Posljedično, manje je situacija u kojima početni model podcjenjuje gibanja na skalama LTD. Kad je takvo podcjenjivanje ipak prisutno, metoda analogona ponaša se u skladu s prethodno pokazanim rezultatima (kod korištenja modela manje rezolucije). Kad model precjenjuje energiju skale LTD, spektar prognoze metodom analogona je vrlo sličan spektru mjerenja (***KFAS***) ili ga blago podcjenjuje (***AN***). Na skali STD postoji podcjenjivanje energije metodom analogona, pri čemu najbolji rezultat ostvaruje prognoza ***KFAS***.

iii. NAKNADNA OBRADA ANSAMBL PROGNOZE

Dostupnost kvalitetnih izmjerenih podataka u planinskom području Hrvatske je ograničena. U prvom dijelu ovog istraživanja samo tri lokacije nakon kontrole kvalitete odgovaraju potrebnim zahtjevima (npr. dovoljna količina raspoloživih podataka u traženom razdoblju) za uspješno testiranje i implementaciju metode analogona. Da bi se bolje istražila primjena metode nad kompleksnom topografijom planinskog tipa, drugi dio ovog istraživanja obuhvaća 29 mjernih postaja u Austriji (Slika 19, str. 53) tijekom zimskog (siječanj) i ljetnog mjeseca (srpanj) u 2018. godini. Nakon što je u prvom dijelu potvrđena uspješnost primjene ove

metode u svrhu poboljšanja rezultata determinističke prognoze numeričkog modela, ispitana je njena sposobnost da se primjeni na ansambl prognozu modela. U prvom dijelu je, pritom, u fokusu deterministička prognoza metodom analogona (kao prognoza kontinuiranog ili kategoričkog prediktanda), dok je u drugom dijelu fokus na ansambl i probabilističkoj prognozi. Drugim riječima, ocijenjen je njen potencijal za kalibraciju ansambl prognoze. U tu svrhu temeljito je analizirana primjena metode analogona na prognozu austrijskog numeričkog modela ALADIN-LAEF (Limited-Area Ensemble Forecasting) (Slika 20, str. 56; Wang et al. [2019]). Cilj drugog dijela istraživanja je poboljšati prognozu brzine vjetra (***LAEFws***) te pritom zadržati računalnu efikasnost izvršavanja. Provedeno je zato nekoliko eksperimenata, koji koriste različite informacije iz prognoze ALADIN-LAEF kao ulazne podatke (tzv. prediktor varijable ili prediktori). Prethodno provođenju eksperimenata provedeni su testovi osjetljivosti. Testovi optimiziraju utjecaj određenog meteorološkog parametra kao prediktora na postupak izdvajanja najkvalitetnijih analogona, neovisno za svaku lokaciju [Junk et al., 2015; Alessandrini et al., 2015a]. Osim pretpostavljenog utjecaja informacije o prognoziranoj brzini vjetra, najbitnija je informacija o smjeru vjetra, zatim temperaturi i relativnoj vlažnosti (Slika 21, str. 59). Pritom je prednost korištenja većeg broja prediktora istaknutija nad topografski planinsko-kompleksnim nego nad pretežno ravnom topografijom (Slika 22, str. 60). Osim izbora meteoroloških parametara, ansambl prognoza početnog numeričkog modela nudi više načina kako koristiti njene prognostičke informacije kao ulazne podatke za metodu analogona. Primjerice, može se koristiti svaka pojedina vrijednost članova ansambla (za jedan ili više meteoroloških parametara) ili sumirati informacije pa, primjerice, koristiti samo informaciju o srednjoj vrijednosti i raspršenju ansambla. Provedeni testovi u potonjem slučaju pokazuju da optimalan doprinos informacije o raspršenju ansambla (mjereno standardnom devijacijom σ) iznosi oko 40 % vrijednosti doprinosa informacije o srednjaku ansambla (Slika 23, str. 61).

Provedeno je ukupno šest eksperimenata metodom analogona, koji se prvenstveno razlikuju po izboru prediktor varijabli iz modela ALADIN-LAEF (Tablica 5, str. 62). Prediktor varijable uključuju:

- Kontrolni (prvi) član ansambla za 6 dostupnih meteoroloških parametara (***AnEnCtrl***)
- Sve članove ansambla prognoze brzine vjetra (***AnEnWs***)
- Srednjake ansambla za 6 dostupnih meteoroloških parametara (***AnEnMu***)

- Srednjake i raspršenja (mjereno sa σ) ansambl prognoza za 6 dostupnih meteoroloških parametara (***AnEnStd***)
- Sve članove ansambla za 6 dostupnih meteoroloških parametara (***AnEnAll***)
- Prognozu za 6 meteoroloških parametara, pri čemu svaka prognoza odgovara (jednom) određenom članu ansambla (***AnEnMem***).

Kratice pripadnih eksperimenata navedene su u zagradama. Dostupni meteorološki parametri uključuju 10-m brzinu i smjer vjetra, 2-m temperaturu, 2-m relativnu vlažnost, prizemni tlak i količinu oborine. Svi eksperimenti produciraju ansambl prognozu brzine vjetra sastavljenu od 17 članova. Rezultati metode analogona uspoređeni su s metodom koja je bazirana na statistici simuliranih podataka za ansambl prognoze (EMOS) [Messner et al.; 2014]. Provedena su dva EMOS eksperimenta: ***EMOSws***, koji koristi zadnjih 30 dana za učenje metode te samo informacije o prognozi brzine vjetra kao ulazni podatak, i ***EMOSstd***, koji koristi cijelo raspoloživo razdoblje za učenje te sve raspoložive meteorološke parametre. Analiza pokazuje da je ***EMOSws*** nešto uspješniji u uklanjanju sustavne pogreške, a ***EMOSstd*** disperzijske pogreške prognoze početnog modela.

Evaluacija ansambl i probabilističke prognoze brzine vjetra

Rezultati pokazuju da su svi AnEn eksperimenti uspješni u poboljšanju prognoze početnog modela (Tablica 6, str. 71; Tablica 7, str. 73). Pritom je računalno najzahtjevniji eksperiment ***AnEnMem*** najmanje uspješan (Slika 29, str. 75). Nepovoljna svojstva početnog modela, poput nedovoljne raspršenosti te loše rezolucije (u smislu da se distribucije prognoza uvjetovanih mjerenim vrijednostima ne razlikuju dovoljno za različite mjerene vrijednosti) ostaju nakon primjene metode analogona u ovom eksperimentu više prisutna nego kod ostalih eksperimenata. Činjenica da je prostor za traženje analogona manji nego u eksperimentima u kojim se članovi razmatraju neovisno, što vjerojatno utječe na ovaj rezultat. Eksperiment ***AnEnWs,*** koji koristi isključivo informacije o brzini vjetra, uspješniji je ili usporediv s eksperimentom ***AnEnMem*** u poboljšanju uspješnosti prognoze te u uklanjanju sustavne pogreške pristranosti srednjaka (ansambla). Dakle, ako je iz nekog razloga dostupna samo prognoza jednog meteorološkog parametra, eksperiment ***AnEnWs*** pokazuje da metoda analogona može poboljšati rezultate. Još bolji rezultati postignuti su u eksperimentima koji koriste informacije o prognozi više od jednog meteorološkog parametra. Primjerice, sličan ili bolji rezultat je postignut pri korištenju prognoza kontrolnog člana ansambla u metodi analogona (***AnEnCtrl***), a korištenjem više od jednog člana ansambla rezultat se dalje

poboljšava. Pritom je pokazano da često nema potrebe koristiti sve raspoložive informacije iz početne prognoze. Naime, korištenje sažetih informacija (u obliku srednjaka i raspršenja ansambla) poboljšava početnu prognozu u gotovo jednakoj mjeri kao kad se koriste sve informacije, s vrlo malo statistički značajnih razlika. Uz to, potonje je računalno manje zahtjevan postupak. Uzevši sve navedene argumente u obzir, može se zaključiti da je upravo ovaj pristup optimalan za primjenu u operativno prognostičkom sustavu. Uz informaciju o pogrešci prognoze, na sažet i efikasan način uključuje se tako informacija o razvoju pogreške koja je dinamički simulirana numeričkim modelom.

Svi eksperimenti poboljšali su rezultate prognoze početnog modela, povećavajući (pretjerano malu) raspršenost ansambla te povećavajući svojstva prognoze poput pouzdanosti i diskriminacije, posebno u siječnju (karakteristični oblici krivulja i načini tumačenja dijagrama korištenih za evaluaciju pobliže su opisani na Slikama 24-28, str. 65 - 70). Općenito su bolji rezultati primjene metoda naknadne obrade postignuti za ljetni mjesec, kada je i rezultat početnog modela nešto bolji nego za zimski mjesec. Iznimka je prognoza ***EMOSws***, koja pokazuje manju raspršenost od očekivane, što je vjerojatno posljedica korištenja manjeg razdoblja učenja nego kod ostalih eksperimenata te samo jednog meteorološkog parametra.

Općenito, točnost ansambl prognoza može se detaljno analizirati pomoću *RMSE* te mjerom neprekidno rangiranog ishoda vjerojatnosti (*CRPS*), koji se može razmatrati kao poopćenje srednje apsolutne pogreške na probabilističke prognoze [Wilks, 2011]. Eksperimenti temeljeni na metodi analogona signifikantno poboljšavaju prognozu početnog modela ALADIN-LAEF (***LAEFws***) za sve prognostičke sate u oba testirana mjeseca (siječanj i srpanj; Slika 30, str. 76). Rezultati su bolji noću nego tijekom dana. Pritom su rezultati metode analogona usporedivi s ili nadmašuju rezultate metode EMOS. Bolji rezultati metode analogona od metode EMOS mogu se uočiti za kratko nastupno vrijeme prognoze, općenito više u siječnju nego u srpnju.

Svi eksperimenti zadovoljavaju zahtjev statističke konzistentnosti da je histogram ranga uniforman (Slika 34, str. 83). Eksperiment ***EMOSws*** pokazuje pretjeranu pouzdanost kod prognoziranja velike vjerojatnosti za ostvarenje događaja, dok eksperiment ***EMOSstd*** premalo pouzdan kod prognoziranja male vjerojatnosti za ostvarenje događaja (Slika 33, str. 82). Eksperimenti koji koriste analogone gotovo su savršeno pouzdani. Dodatno, svojstvo diskriminacije (između situacija koje jesu i onih koje nisu rezultirale ostvarenjem događaja) veće (bolje) je kod prognoza metodom analogona zbog većeg udjela točnih prognoza u

ukupnom broju ostvarenih događaja. Razlike među pojedinim eksperimentima metode analogona manje su istaknute nego kad se usporede s metodom EMOS ili rezultatom početnog modela. Rezultati ***AnEnStd*** i ***AnEnAll*** gotovo su identični, potvrđujući da je korištenje sažetih informacija o prognozi početnog modela najčešće sasvim dovoljno.

Prostorno gledajući, pogreška prognoze ***LAEFws*** prati klimatološku razdiobu prosječne brzine vjetra, ispoljavajući veću pogrešku u područjima sklonim pojavi jačeg vjetra (Slika 31, str. 78). Primjenom metoda naknadne obrade, prognoza se dodatno poboljšava slijedeći sličnu prostornu razdiobu. Pri prostornoj evaluaciji primijećeno je da u području izrazito kompleksne topografije za prostorno bliske lokacije postoje velike razlike u uspješnosti prognoze numeričkog modela. Prognoza ***LAEFws*** uspješnija je za lokacije koje su smještene u kotlini od onih koje su na višoj nadmorskoj visini (na planini). Koristeći (bolju) ***LAEFws*** prognozu i konačan rezultat nakon primjene metode analogona je bolji za postaje u kotlini (Slika 32, str. 80). Međutim relativno poboljšanje u odnosu na prognozu početnog modela je zapravo mnogo više izraženo kod korištenja (lošije) ***LAEFws*** prognoze na višim nadmorskim visinama. Takav efekt posljedica je uklanjanja sustavnih izvora pogreške (pristranost srednjaka i σ), koji su u većoj mjeri prisutni u prognozi numeričkog modela za lokacije na planini.

Iako je pojava slabog i umjerenog vjetra mnogo češća, bitno je razmotriti i kvalitetu prognoze za jak vjetar zbog njegovog utjecaje na ljude i imovinu. Koristeći više pragova za brzinu vjetra (u rasponu 0.5 – 20 ms^{-1}), testirana je uspješnost prognoze za različite brzine vjetra (Slika 35, str. 85). Pokazano je da ***LAEFws*** prognoza pokazuje uspješnost isključivo za malu brzinu vjetra (npr. do 3 ms^{-1}). Sve testirane metode naknadne obrade poboljšale su uspješnost prognoze i za veću brzinu vjetra. Pritom je metoda analogona značajno uspješnija od metode EMOS za brzinu vjetra do 10 ms^{-1}, neovisno o dobu godine. Štoviše, eksperimenti ***AnEnStd*** i ***AnEnAll*** značajno poboljšavaju rezultate početnog modela za sve testirane pragove brzine u siječnju.

iv. ZAKLJUČAK

Rezultati pokazuju da deterministički produkt metode analogona ima veću koreliranost prognoze i mjerenja te manju pogrešku u odnosu na početni numerički model koji metoda koristi kao ulazni podatak. Dok prognoziranje srednjaka ansambla analogona rezultira najvećom korelacijom, primjena Kalmanovog filtra u takozvanom prostoru analogona

(KFAS) je eksperiment koji je najmanje sklon podcijeniti prirodnu varijabilnost vjetra, čak i na kratkim vremenskim skalama.

Metoda analogona primijenjena je i na ansambl prognozu numeričkog modela, pri čemu je pokazano da je upravo korištenje sažetih informacija o prognozi ulaznog modela optimalan način da se poboljša točnost prognoze, čak i za prognozu jakog vjetra. U numeričkom modelu procesi su bolje reprezentirani za lokacije smještene na nižoj nadmorskoj visini nego za planinske lokacije, što znači i bolji ukupan rezultat nakon naknadne obrade produkata modela. Međutim, relativno poboljšanje u odnosu na početni model istaknutije je na višim nadmorskim visinama.

§ 1. INTRODUCTION

1.1. Motivation

The skill of short and medium-range numerical weather prediction models has improved at both global and regional scales. Their ability to simulate and forecast winds in complex topography and coastal areas is, however, still largely affected by insufficient resolution, imperfect boundary and initial conditions, simplification of physical processes and numerical approximations. It is often considered that the higher the model resolution the more accurate the forecast, due to better resolved lower boundary conditions and flow adaptation when decreasing the grid spacing. These benefits are not always evident [e.g. Mass et al., 2002; Rife and Davies, 2005]. Even at the sub-kilometer grid spacing, state-of-the-art mesoscale models still exhibit considerable errors, especially in complex topography [Horvath et al., 2012]. This is particularly relevant for operational weather prediction systems that are constrained by the available computing resources. It is thereby useful to develop suitable post-processing methods that reduce starting model errors at locations where measurements are available, besides improving the model itself (e.g., using a higher resolution or improved parametrization package).

1.2. Using the analogies to predict the weather

The idea that analogies (i.e., similar past forecast, measurements, or analysis) can be used for forecasting future weather has been explored for decades. It is based on an assumption that if two atmospheric states are initially very close, they will remain somewhat close for some time in the future. For instance, Lorenz [1969] claims that it is hard to identify any state in the past that can be considered a good match to the present large-scale flow pattern, except for mediocre analogues. Furthermore, Rousteenoja [1988] and Lorenz [1969] state that one needs to wait an astronomically large number of years until the likelihood of finding two atmospheric states that differ less than the present-day observational error is sufficiently high

enough to be considered as usable. Back then, the applicability of analogues for short-range weather forecasting is practically discarded. Van den Dool [1989], however, shows that it is possible to find useful analogies if the number of degrees of freedom in the matching procedure is reduced. The author uses analyses over a localized area (i.e., not entire Northern Hemisphere as in Lorenz [1969]) and then uses the 12-h subsequent analysis to each analogue as a plausible 500 hPa height forecast. Various procedures are formulated afterward, including different predictors and analogue selection criteria. This is done mainly because the use of analogues for forecasting of meteorological fields is limited due to excessive degrees of freedom of the problem at stake. Applications including long-range weather predictions using National Oceanic and Atmospheric Administration (NOAA) outgoing long-wave radiation fields [Xavier and Goswami, 2007] and very short-term orographic precipitation predictions using radar observations [Panziera et al., 2011] are proved to be skillful. The Southern Oscillation Index (SOI) forecasts using SOI measurements [Drosdowski, 1994] and point wind speed forecasts using wind speed measurements [Klausner et al., 2009] exhibit satisfactory results as well. Besides single fields, also the use of spatially correlated observational variables [Wu et al., 2012] also proofed to be suitable.

Besides predicting the weather using past measurements or analyses, analogies can be employed to reduce the errors in the numerical weather prediction (NWP) model simulations. This approach utilizes the achievements of numerical modeling in predicting future state of the atmosphere. Additionally, it can reasonably absorb the information of the analogues in historical data (statistical model) in order to improve forecast skill as shown for idealized cases with low-order models [Ren and Chou, 2006] and general circulation modeling [Gao et al., 2006; Ren and Chou, 2007].

Van den Dool [1989] reveals that analogues can be used to predict the forecast skill of a NWP model. Hamill et al. [2006] and Hopson [2005] extend the idea and apply the analogues to ensemble forecasts. Hamill and Whitaker [2006] state that, when comparing the pattern match of the historical local ensemble-mean forecast to the current ensemble-mean forecast in the same region, it is possible to find many similar and useful analogs within a few decades of re-forecasts. Their study focuses on probabilistic forecasts of 24-h precipitation. All the aforementioned analog-techniques are able to improve the Brier skill score, resulting in a skill comparable to a logistic regression technique. The authors, while comparing different analog-techniques, also conclude that selecting analogs for each member rather than for the ensemble

mean generally decrease the forecast skill. Another successful example of a calibrating ensemble forecast can be found in Hopson and Webster [2010]. The authors seek analogs in order to generate the final set of discharge ensembles accounting for all aspects of discharge forecast uncertainty (meteorological and hydrological). This part of the fully automated operational 1-10-day multi-model ensemble forecasting scheme for the major river basins of Bangladesh helped to evacuate many thousands of people and livestock during flood events in 2007.

As a very successful continuation of the aforementioned studies, Delle Monache et al. [2011] propose two variations of analog-based post-processing method to improve deterministic NWP forecasts of 10-m wind speed, based on a historical data set including NWP data and observations at a single site. The weighted mean (***AN***) of the analog ensemble (AnEn) is tested and compared to a linear, adaptive and recursive Kalman filter (KF) post-processing approach [Delle Monache et al., 2006, 2008, 2011]. Another approach is to apply Kalman filter to the historical set of (starting) model forecasts in the analog space, ordered from the worst to the best analog (Kalman Filter in Analog Space – ***KFAS***; Delle Monache et al. [2011]). With that approach, the correction of the current forecast is based on a higher weight to the analog forecasts closer to it. The authors demonstrate that both approaches increase correlation and reduce random and systematic errors. Similar approaches are used for predicting other variables as well. Djalalova et al. [2015] show similar results predicting PM2.5 concentrations, while Nagarajan et al. [2015] test the techniques across several models and meteorological variables. Additionally, Djalalova et al. [2015] apply the KF to the time series of the ***AN***, resulting in a new deterministic forecast called the ***KFAN***.

Delle Monache et al. [2013] explore benefits from using the analogs to produce probabilistic 10-m wind speed and 2-m temperature AnEn forecasts from a deterministic NWP. The authors show that the AnEn exhibits high statistical consistency, reliability and the ability to capture the flow-dependent behavior of errors. The use of an analog-based method to produce probabilistic output is not limited to short- or medium-range forecasts. Vanvyve et al. [2015] provide high-quality long-term wind resource estimates, characterized by an accurate wind time series and frequency distribution. In addition to using probabilistic analog-based predictions to gain wind resource estimates [Vanvyve et al., 2015; Zhang et al., 2015], they are also used to downscale precipitation [Keller et al., 2017], to predict solar irradiance

[Alessandrini et al., 2015a], 10-m wind speed [Sperati et al., 2017] and wind power [Alessandrini et al., 2015b; Junk et al., 2015].

Additional to using a deterministic NWP to create AnEn [Delle Monache et al., 2011; 2013], the same approach can also be applied using an NWP ensemble. The AnEn ability to capture the flow-dependent error growth is complemented with the aspects of error growth that can be represented dynamically by the multiple model runs of an NWP ensemble. Following that idea, Eckel and Delle Monache [2016] produce m analogs for each member of the n-member NWP ensemble, resulting in an m×n "hybrid" AnEn. The approach yields mixed results for the 10-m wind speed forecasts, while the application for the 2-m temperature forecast is more successful. Mugume et al. [2017], who uses the analog-based method to post-process ensemble members with different convection parameterization schemes, also explore the same idea. The authors demonstrate a root-mean-square error (*RMSE*) and bias reduction in rainfall prediction when using corresponding predictions of the (starting) ensemble mean analog as a forecast. Slightly better results (e.g. significant reduction of negative bias error) are achieved when seeking the analog for every (starting) ensemble member and then average the analogs. Finally, since the AnEn can be affected by a conditional negative bias, especially when predicting events in the right tail of the forecast distribution, the novel bias correction method is proposed by Alessandrini et al. [2019].

1.3. Research objectives

In this research, we propose an in-depth analysis of analog-based method over complex topography. The target area of this research is located in Croatia, where different mesoscale wind regimes include strong bora downslope windstorms (which may reach hurricane scale strength, e.g., see review by Grisogono and Belusic [2009]), mountain valley and slope winds, and thermally-induced land-sea breeze (e.g., Telišman Prtenjak and Grisogono [2007]; Horvath et al. [2011]). Due to the importance of model resolution necessary to represent wind processes in the target area, we study whether the post-processing improves results when using a higher-resolution starting model. We thus test the role of 8- and 2-km grid spacing full-physics Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) model. In addition, we use a model that dynamically adapts the 8-km ALADIN output to the

2-km grid spacing. The latter is a configuration (e.g., Žagar and Rakovec [1999]; Ivatek-Šahdan and Tudor [2004]) used for operational wind forecasting in the ALADIN consortium and Croatian Meteorological and Hydrological Service.

We study the performance of different post-processing methods using metrics that consider wind speed as both continuous and categorical predictand. These include ***AN***, ***KF***, ***KFAS***, and ***KFAN***, as described above. We analyze the results across three regions with distinct wind regimes:

i. coastal complex topography where the most significant portion of mesoscale energy is governed by strong downslope windstorms as well as thermally induced land-sea circulations,
ii. mountain complex topography where the most significant portion of mesoscale energy is governed by the weak-to-moderate valley and slope mountain winds, and
iii. continental nearly flat topography where the motions are predominantly of synoptic-scale variability and origin [Zaninović et al., 2008; Horvath et al., 2011].

The focus is set on the complex topography, primarily coastal region. Therefore, we study the importance of the starting model resolution and formulation by using three versions of ALADIN focusing on coastal complex topography characterized by a plethora of mesoscale wind processes.

In contrast to coastal complex topography, the availability of the quality data over mountain complex topography in Croatia is limited. Only three mountain locations satisfy the necessary quality demands for the analog method testing and implementation in the first part of this research (i.e. having a similar amount of data after basic quality control as for other sites). For that reason, the research is extended using 29 meteorological observation sites (TAWES) in Austria for winter (January) and summer (July) month of 2018. After investigating wind speed as continuous and categorical predictand, the focus is now extended to the ensemble and probabilistic wind speed forecasting. In addition to using deterministic NWP input to analog-based method, the ability to calibrate the ensemble NWP is also investigated. Therefore, an in-depth analysis of the analog-based method applied to the Austrian ALADIN-LAEF (Aire Limitée Adaptation dynamique Développement InterNational – Limited-Area Ensemble Forecasting) ensemble forecasts is provided in the second part of this research. Following the work of Eckel and Delle Monache [2016] and Mugume et al. [2017], the main goal is to significantly improve the ALADIN-LAEF ensemble 10-m wind

speed forecast while maintaining low computational cost for the analog search. To test the performance of the analog-based method and determine the optimal configuration, several experiments using different sources of information available of the ALADIN-LAEF ensemble forecasts are performed. The experiments include using one or more ALADIN-LAEF meteorological variables as predictors. The experiment using only ALADIN-LAEF control member for several meteorological variables as predictors is included to represent the analog-based method performance using the deterministic input, similarly as the ALADIN model is used within the first part of this thesis.

Through performed analysis, the experiments including only information about the ALADIN-LAEF ensemble mean (as suggested by Hamill and Whitaker [2006]) or every ensemble member (similar as in Mugume et al. [2017]) are also tested. A novelty in this research is the usage of the starting model ensemble uncertainty through its standard deviation (σ) in addition to ensemble mean (μ). The hypothesis additionally explored in this thesis is that using a summarized measure, like standard deviation σ, is the optimal way to dynamically represent the aspects of error growth of the input ensemble model to the flow-dependent error growth, which is already captured by the analog approach [Odak Plenković et al., 2020]. The ensemble model output statistic post-processing approach (EMOS*;* [Gneiting et al., 2005]) is used as a reference model in order to better understand the analog-search impact on the raw forecasts. All experiments provide 17 members wind speed AnEn forecast, as well as the ALADIN-LAEF forecast.

§ 2. ANALOG-BASED METHOD

The AnEn can be used to estimate the probability distribution $f(y|x^f)$ of the observed future value of the variable y at a given time and location. The x^f represents k variables (predictors) from the deterministic (starting) model $x^f = \left(x_f^1, x_f^2, \ldots, x_f^k\right)$. To generate y samples, the analog-based method uses historical data within a specified analog training period for which both the deterministic NWP (starting model) and the verifying observation are available, as schematically shown in Figure 1.

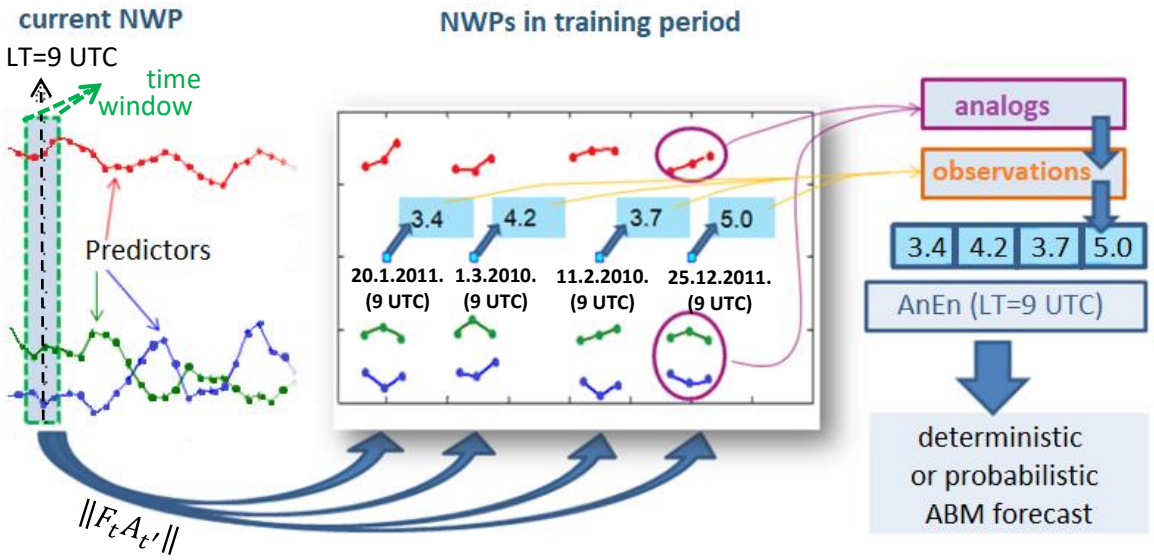

Figure 1. The analog-based method scheme for 4-member AnEn forecast at 09 UTC lead time. In this example, 3 predictor variables (i.e. wind speed, wind direction, and temperature) from the current NWP are used in the analog search procedure. For each variable, the values within a 3-lead-time-steps-wide time window (centered around 09 UTC) are compared to the historical forecast within the time window of the same width (also centered around 09 UTC). The predefined metric $\|F_t A_{t'}\|$ is used to determine the quality of the match. Once the most similar historical forecasts are found, the AnEn is formed out of verifying observations. The deterministic forecast can then be issued as, for example, the mean of the AnEn. On the other hand, the probability of a pre-defined event (probabilistic forecast) can be calculated by counting the AnEn members predicting the event will happen.

The best-matching historical forecasts to the current prediction, so-called analogs, may originate in any past date in the training period. The quality of the analog is evaluated by the following metric:

$$\|F_t A_{t'}\| = \sum_{i=1}^{N_A} \frac{w_i}{\sigma_{fi}} \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} \left(F_{i,t+j} - A_{i,t'+j}\right)^2}, \quad (1)$$

where F_t is the current NWP deterministic forecast at a given location, valid at the future time t, whereas $A_{t'}$ is an analog at given location with the same forecast lead time, but valid at a past time t'. The N_A is the number of predictors used in the search for analogs, w_i are the weights corresponding to the particular predictor. The absolute value of the metric is not important as such since it is only used for the inter-comparison of analogs when used for sorting by the quality. Therefore, the weights are not constrained (i.e. their sum does not need to be fixed). For the fair comparison between different meteorological parameters, however, the weights are normalized using the standard deviation (σ_{fi}) of past forecasts of a given variable at the same location. The $\tilde{t}$ is equal to half the number of additional times over which the metric is computed (the half of the time window of any specified width). Therefore, $F_{i,t+j}$ and $A_{i,t'+j}$ are the values of the forecast and the analog in the time window for a given variable, respectively. The time window is used to account for shifts and/or trends in the starting model forecast. Analogs are found independently for every forecast time and location, narrowing the search around a particular time of a day by a time window. In other words, the number of degrees of freedom in analog finding procedure is reduced (as proposed in Van den Dool [1989]). The $\tilde{t}$ value used in this research is equal to 1 lead time step, as proposed by Delle Monache et al. [2013]. The verifying observations of the best-matching analogs are the members of AnEn.

The assumption is that the errors of the good (quality) analog forecasts are likely to be similar to the error of the current forecast [Delle Monache et al., 2011] and hence reduced by the historical observation used. Several authors state that the AnEn rank histograms are uniform (e.g., Delle Monache et.al. [2013]). Therefore, every member of the AnEn is an equally probable outcome, even though, measured by previously defined metrics, some analogs are closer to the current forecast than the others are. Once the AnEn is formed, it can be used to produce the deterministic analog-based prediction, as well as the probabilistic forecast (e.g., to estimate the probability of a predefined event).

§ 3. POST-PROCESSING THE DETERMINISTIC NWP

3.1. Observations and climatology

The post-processed forecasting methods are tested at 14 locations in Croatia, covering different climatological regions (Figure 2). The locations are selected based on the availability of wind speed measurements (10-minute average value) at 10 m above the ground in the 2010-2012 period. The list of locations with the geographical features is given in Table 1.

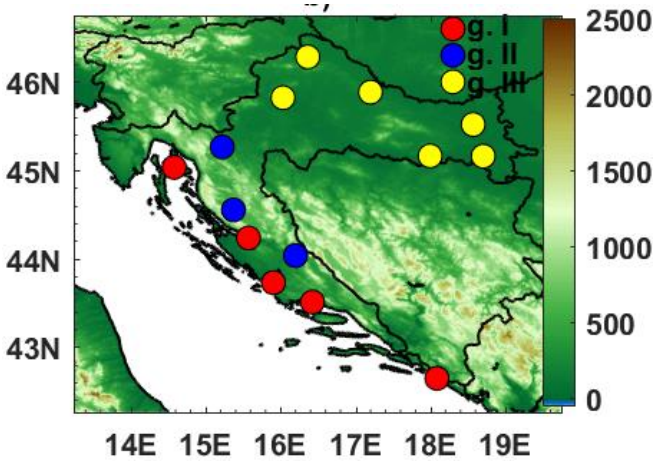

Figure 2. Topography and spatial distribution of the 14 stations providing the 10-m wind speed observations used in the section 3. The stations are divided in three groups: coastal complex (group I; red markers), mountain complex (group II; blue markers) and nearly flat continental topography (group III; yellow markers).

Our goal is to compare and contrast the performance of the different methods, generated from different NWP models, and at different complex topography and coastline sites. The locations are thereby divided in three groups:

I. Group I is a coastal complex topography region that includes the locations near the coastline and near the western slopes of Dinaric Alps. The prominent wind in this area is bora, a strong and gusty downslope windstorm (e.g., see review by Grisogono and Belušić [2009]). The bora wind is more frequent in the northern than in the southern Adriatic. Nevertheless, its maximal strength is similar in both regions [Horvath et al., 2009]. Other mesoscale wind circulations are also notable and are governed by the

surface inhomogeneity (e.g. land-sea breeze) and vicinity of the mountains (e.g., mountain-plain circulation, gap flows, weak downslope flows). Therefore, the diurnal cycle is shaped by the proximity of the sea and terrain elevation. The highest wind speeds analyzed in this section are recorded in this area (Figure 3a) and the mean wind speed is 4.0 ms^{-1}.

II. Group II is a mountain complex topography region with highly-complex topographical features. Locations in this area are farther from the coastline and at higher elevation than the locations in any other group, with mountain tops reaching 1500 m above sea level. Because of terrain complexity and low population density the measurements are coarse in space in this area. The measurements may also be prone to longer data gaps due to remoteness of locations and generally more severe winter climate. After our analysis, we therefore choose three locations that satisfy the basic quality requirements within this area (e.g. that there are no gaps longer than a few weeks). This area is characterized by a significant portion of energy variance due to mountain slope and valley winds. Wind speeds in the mountain complex topography are lower than in the coastal complex topography (Figure 3b) and the mean wind speed is 2.0 ms^{-1}.

Table 1. The list of the 14 stations providing the 10-m wind speed observations used in section 3. The stations are divided in three groups: coastal complex (group I; red), mountain complex (group II; blue) and nearly flat continental topography (group III; yellow).

Location name	Latitude	Longitude	Altitude [m]
Dubrovnik	42.6	18.1	52
Jasenice	44.2	15.6	170
Krk	45.2	14.6	57
Split	43.5	16.4	122
Šibenik	43.7	15.9	77
Gospić	44.6	15.4	564
Knin	44.0	16.2	255
Ogulin	45.3	15.2	328
Bilogora	45.9	17.2	262
Gradište	45.2	18.7	97
Osijek	45.5	18.6	89
Slavonski Brod	45.2	18.0	88
Varaždin	46.3	16.4	167
Zagreb	45.8	16.0	123

III. Group III stations are located in the nearly-flat inland continental climatological region of Croatia. The terrain elevation is up to 100 m above sea level. The diurnal cycle is shaped mainly by the gentle microscale variations of the topography. The region is still influenced by non-local effects of the Dinarides mountain system to the west and southwest, since these mountains affect predominant westerly flow through channeling, blocking and other mesoscale processes. A strong wind is very rare in the continental area, and it occurs during the cold air outbreaks from polar or Siberian areas in winter or during rough weather in summer [Zaninović et al., 2008]. The wind speeds are relatively low (Figure 3c) and the mean wind speed is 2.0 ms^{-1}.

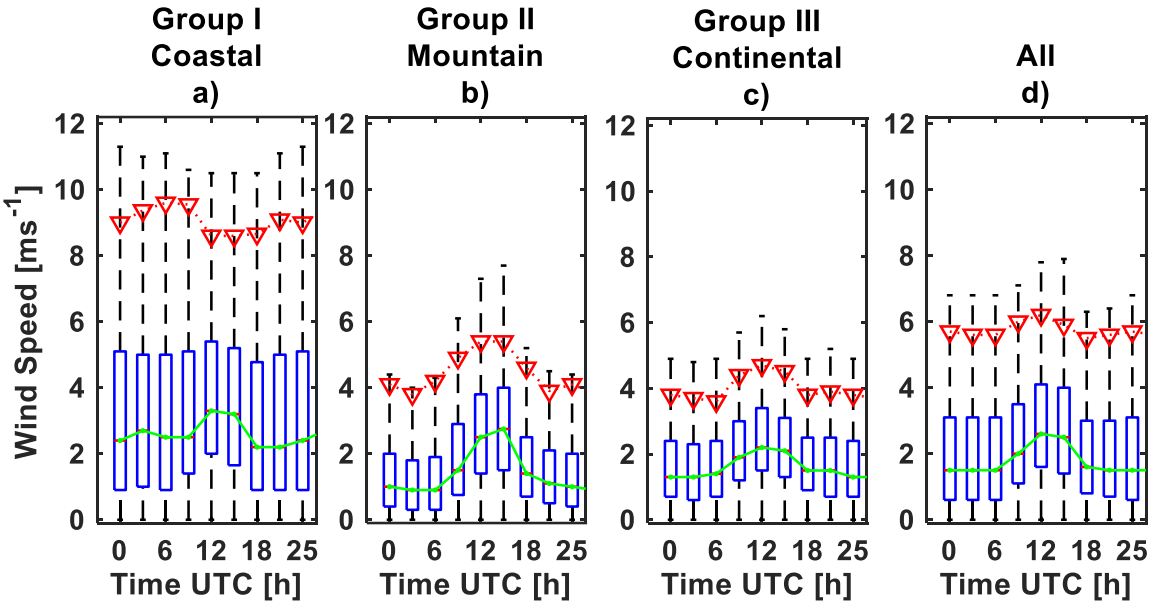

Figure 3. The boxplots of the observed data (outliers are not shown), depending on time of the day. The data are measured during the 2010-2012 period at 14 stations in Croatia. In addition to the boxplot for all the data available (d), the data are sorted into groups (a-c) based on topography type and basic climatological features. The green lines represent the 50^{th} and red triangle markers the 90^{th} percentile, respectively. Those values are used as thresholds between categories in the verification procedure. The exact values are listed at Table 2.

Mean wind speed for all 14 stations is 2.7 ms^{-1}. The maximum of the diurnal cycle occurs around 12 UTC on average for all stations (Figure 3d). However, different processes contribute to the average daily cycle at different locations.

Table 2. The exact values of the 50th and 90th percentile of the observed data at 14 stations in Croatia during the 2010-2012 period, depending on time of the day (as shown in Figure 3). The data is sorted into groups based on topography type and basic climatological features.

Time UTC [h]	Group I		Group II		Group III		All	
Percentile	50th	90th	50th	90th	50th	90th	50th	90th
0	2.4	9.0	1.0	4.1	1.3	3.8	1.5	5.7
3	2.7	9.4	0.9	3.8	1.3	3.7	1.5	5.6
6	2.5	9.6	0.9	4.2	1.4	3.6	1.5	5.6
9	2.5	9.6	1.5	4.9	1.9	4.4	2.0	6.0
12	3.3	8.6	2.5	5.4	2.2	4.7	2.6	6.2
15	3.2	8.6	2.8	5.4	2.1	4.5	2.5	5.9
18	2.2	8.7	1.4	4.6	1.5	3.8	1.6	5.5
21	2.2	9.1	1.1	3.9	1.5	3.9	1.5	5.6

Finally, the values of 50th and 90th percentile are shown in Figure 3 and listed in Table 2. Those values are used as thresholds between categories in the verification procedure.

3.2. NWP model data

Three operational configurations of the limited-area mesoscale NWP model ALADIN (Aire Limitée Adaptation dynamique Développement InterNational model) [ALADIN International Team, 1997], that were issued at the Croatian Meteorological and Hydrological Service in the 2010-2012 period, are used to generate 10-m wind speed forecasts in this thesis:

I. The operational limited-area mesoscale ALADIN model was launched twice a day (00 UTC and 12 UTC) at 8-km horizontal grid spacing (***A8***). The ***A8*** model used the hydrostatic dynamics with spectral solver on 37 hybrid sigma-pressure vertical levels [Tudor et al., 2013; Ivatek-Šahdan et al, 2018]. The initial conditions were based on a variational data assimilation scheme for the upper-air fields and optimal interpolation for surface variables [Stanešić, 2011]. The lateral boundary conditions were given by the Action de Recherche Petite Echelle Grande Echelle (ARPEGE) global model, which was run operationally at Meteo France. Vertical transfer of momentum, heat, and moisture were based on a scheme that used prognostic turbulence kinetic energy

[Geleyn et al., 2006] combined with modified Louis [1982] stability dependency in the surface layer [Redelsperger et al., 2001]. Contribution of shallow convection to the evolution of prognostic fields was calculated within the turbulence parametrization according to Geleyn et al. [1987]. Deep convection is described by a modified diagnostic Kuo scheme [Geleyn et al., 1994]. Microphysics parametrization [Catry et al., 2007] included prognostic treatment of cloud water/ice, rain, and snow, as well as a statistical approach for sedimentation of precipitation [Geleyn et al., 2008]. Radiation effects were described according to Geleyn and Hollingsworth [1979], and Ritter and Geleyn [1992]. The impact of soil processes on prognostic model fields was accounted for by a two-layer Interaction Soil Biosphere Atmosphere (ISBA) scheme [Noilhan and Planton, 1989], which was also used for the surface data assimilation [Giard and Bazile, 2000]. Physics contribution was coupled to the dynamics via interface based on a flux-conservative set of equations [Catry et al., 2007].

II. An operational ALADIN high-resolution dynamical adaptation (***DA***) model. The ***DA*** procedure [Žagar and Rakovec, 1999] was taking the output fields from the ***A8***. The ***DA*** dynamically adapted wind fields to the higher resolution horizontal terrain (2-km grid spacing) by adopting the model field to reach a quasi-stationary state forced by time-invariant lateral boundary conditions [Ivatek-Šahdan and Tudor, 2004]. Vertical levels in the planetary boundary layer were approximately at the same heights as in the ***A8*** model (the lowest level is about 17 m above ground). The vertical levels in the upper troposphere and stratosphere were reduced, i.e., the ***DA*** was run on 15 levels in the vertical. The wind field was interpolated to the height of measurements using the stability functions and the Monin-Obukhov similarity theory [Geleyn, 1988]. Turbulence was the only parametrization scheme used in the ***DA***, while contributions of moist and radiation processes were neglected. This cost-effective forecast refinement was run operationally twice a day (00 and 12 UTC run) for 72 h ahead with a 3-h model output frequency. In the complex topography, the ***DA*** improved near-surface wind predictions, as described in a number of studies such as Tudor and Ivatek-Šahdan [2002], Ivatek-Šahdan and Tudor [2004], Ivatek-Šahdan and Ivančan-Picek [2006], Bajić et al. [2007, 2008], Horvath et al. [2011], etc. The ***DA*** was used for operational wind forecasting in several countries that are members of the ALADIN consortia.

III. ALADIN at 2-km horizontal grid spacing (***A2***) was configured similar to the ***A8***, but with non-hydrostatic dynamics [Ivatek-Šahdan et al, 2018]. Physics parametrizations included a full parametrization set as in the ***A8***, with an upgrade of a deep convection parametrization. Unlike the ***A8***, the deep convection in the ***A2*** was a prognostic mass-flux type scheme [Gerard and Geleyn, 2005; Gerard, 2007]. The convective processes in the ***A2*** were accounted for the use of prognostic variables for updraft and downdraft vertical velocities and mesh fractions [Gerard et al., 2009]. The ***A2*** was initialized from the 06-h forecasts of the operational ***A8*** 00 UTC run, and it was run with the Scale-Selective Digital Filter Initialization [Teremonia, 2008]. This high-resolution forecast was run once daily for 24 hours in advance (until 06 UTC of the following day), with 1-h model output frequency on 37 vertical levels [Tudor et al., 2013].

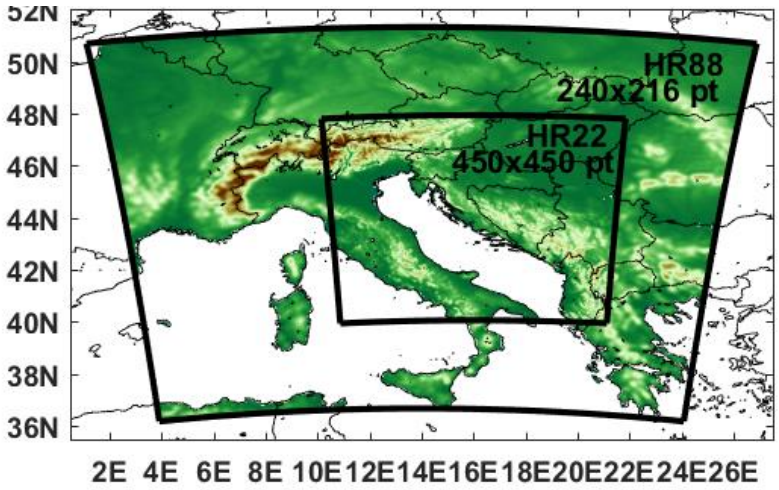

*Figure 4: The ALADIN model domains and topography – larger with 8 km horizontal grid spacing (**A8**) and smaller with 2 km horizontal grid spacing (**A2**, **DA**).*

All three ALADIN configurations (***A8*** 00 UTC, ***DA*** 00 UTC and ***A2*** 06 UTC) were used to prepare forecasts for the period 2010-2012. The domains for all configurations are shown in Figure 4. For every location of the analyzed measurement stations, the closest model grid point (on land) is chosen from the four grid points surrounding the observation location.

3.3. Reference method: Kalman filter

Generally, the Kalman filter (KF) approach is a recursive post-processing method used to estimate a signal from noisy measurements. It has been mainly used in data assimilation schemes to improve the accuracy of the initial conditions for the NWP [e.g., Burgers et al., 1998; Houtekamer et al., 2005]. The KF has also been used for NWP model forecasts as a predictor bias correction method during post-processing of short-term weather forecasts [Homleid, 1995; Roeger et al., 2003]. In a post-processing predictor bias correction method, the information (i.e., recent past forecasts and observations) is used to revise the estimate of the current raw forecast. Previous bias values are used as input to KF. The bias here is defined as the "difference of the central location of the forecasts and the observations" [Jolliffe and Stephenson 2003]. The filter estimates the systematic component of the forecast errors (i.e. bias). Once the future bias has been estimated, it can be removed from the forecast to produce an improved forecast. Such a corrected forecast should be statistically more accurate in a least-squares sense. Further details on the KF predictor bias correction post-processing method are given below.

The optimal recursive predictor of forecast bias x_t at time t is derived by minimizing the expected mean square error. Kalman [1960] shows that x_t at time t can be written as a combination of the previous bias estimate and the previous forecast error y_t (the hat (^) indicates the estimate):

$$\hat{x}_{t+\Delta t|t} = \hat{x}_{t|t-\Delta t} + K_t\left(y_t - \hat{x}_{t|t-\Delta t}\right). \tag{2}$$

The K_t is a weighting factor called Kalman gain and can be calculated from:

$$K_t = \frac{p_{t-\Delta t}+\sigma^2_{\eta,t}}{\left(p_{t-\Delta t}+\sigma^2_{\eta,t}+\sigma^2_{\varepsilon,t}\right)}. \tag{3}$$

The expected mean-square error p can be computed as:

$$p_t = \left(p_{t-\Delta t} + \sigma^2_{\eta,t}\right)(1 - K_t). \tag{4}$$

The $\sigma^2_{\eta,t}$ and $\sigma^2_{\varepsilon,t}$ are variances of the noise term and the unsystematic error term, respectively. Their so-called error ratio is set to 0.01 value, following the other authors (i.e. Delle Monache et al. [2006; 2011]). However, it needs to be noted that the KF performance is sensitive to the error ratio. If the ratio is too high, the filter will put excessive confidence in the previous forecast, and the predicted bias will respond very quickly to previous forecast errors. On the other hand, if the ratio is too low, the predicted bias will change too slowly

over time. More details on the sensitivity of the error ratio can be found in Delle Monache et al. [2008].

For any plausible estimate of p_0 and K_0 the KF algorithm converges promptly, producing the Kalman filter forecast (***KF***). Additional details of the procedure and algorithm applied in this research can be found in Delle Monache et al. [2006].

The KF is easy to implement and computationally inexpensive. Since the KF approach adapts its coefficients during each timestep there is no need for a long training period. The advantages of the KF approach also include the ability to adapt to changing seasons, and even changing models. However, a disadvantage of this method is that it is not likely to predict sudden changes in the forecast error caused by rapid transitions from one weather regime to another [Delle Monache et al., 2011]. Overall, these advantages and disadvantages make the KF a valuable reference to assess the performance of the proposed analog-based method.

3.4. Description of experiments

The ***AN*** forecast for the future time t at a given location is an average (weighted, if $\gamma \neq 1/N$) of the observations O_i corresponding to N most similar analogs $A_{t'}$ (measured by metrics previously defined in equation 1):

$$AN_t = \frac{1}{N}\sum_{i=1}^{N} \gamma O_i(A_{t',i}). \quad (5)$$

In other words, the $\boldsymbol{AN}_t$ is a (weighted) mean of N-sized AnEn for a (future) time *t*. Several authors, such as Delle Monache et al. [2013], state that the AnEn rank histograms are uniform. Every member of the AnEn is thus an equally probable outcome, even though some analogs are closer to the current forecast than the others (measured by previously defined metrics). Hence, the value assigned to the weight γ is 1.

Forecasting the median of the AnEn (***ANM***) is additionally used as an alternative to the AN that is less sensitive to the assumptions about the overall nature of the data (e.g. robust) and to the small number of outliers (e.g. resistant) [Wilks, 2011]. The analogs are searched in forecast space only, for both ***AN*** and ***ANM.*** Therefore, no observations are used to select the best analogs and some sort of correction in real-time is desired.

The KF approach uses all the available information to estimate the error of the current forecast, recursively giving higher weights to the most recent data. However, the KF alone is not able to predict large day-to-day changes in the prediction error, as discussed thoroughly in Delle Monache et al. [2011]. Benefits and shortcomings of the methods using analogs and KF complement one another, hence combining them seems like a reasonable choice. In this research two different ways to combine these methods are tested and schematically presented in Figure 5.

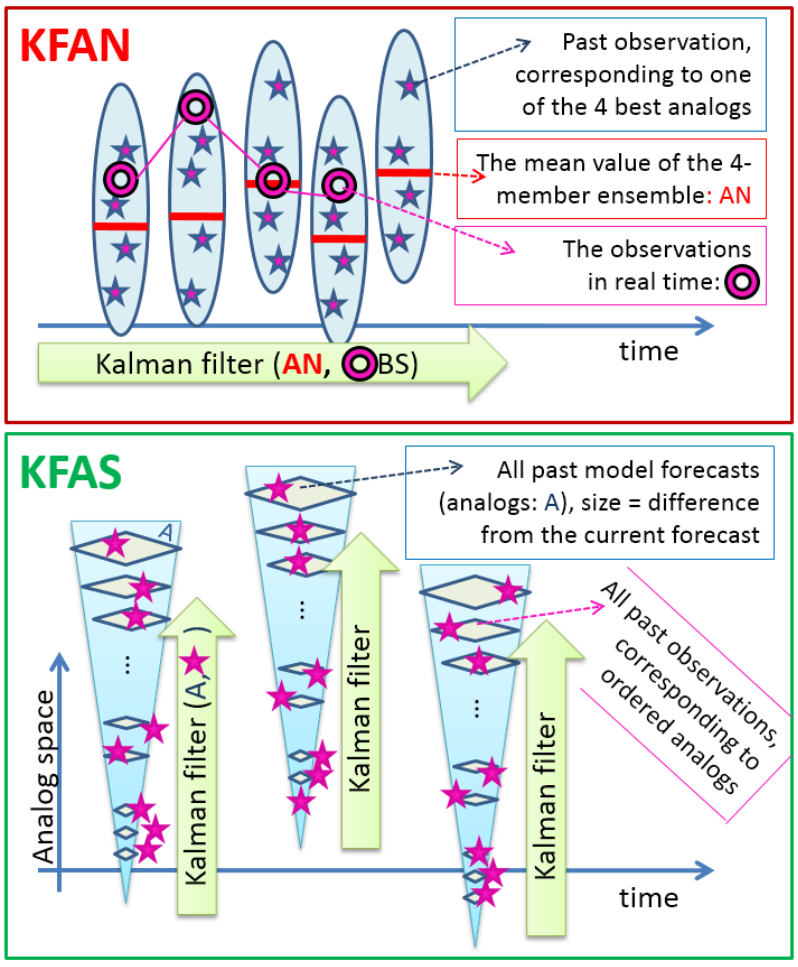

*Figure 5. The schemes for the **KFAN** and the **KFAS** forecasts in real-time. For the **KFAN** forecasting, the last member of the **AN** time series is created, while previously issued **AN** forecasts are saved. The **AN** is hereby the mean of the N-member ensemble (N=4 in this example). The KF is then applied to the time series of **AN** values and real-time measurements, recursively giving the highest weight to the most recent **AN** (i.e. closest in time). For the **KFAS** forecasting, the entire time series of previously issued model forecasts (analogs) are sorted by their similarity to the current model forecast, thus forming an analog space. Then, the KF is applied to the analogs and corresponding measurements in the analog space, giving the most weight to the most similar forecast.*

The first combination of analog- and KF-based approaches includes running algorithms independently. First, the ***AN*** forecasts are issued (or already saved), completing the time series of the ***AN*** forecasts. The last member of the ***AN*** time series is valid at the future time *t*. Then, the KF algorithm is applied (in time) to the time series of the ***AN*** forecasts. The Kalman filter of the ***AN*** forecast is created – therefore the ***KFAN*** forecast. In other words, the KF is applied to the time series of the mean AnEn values. Hereby, every ensemble consists of observations corresponding only to the *N* best analogs. The KF algorithm gives more weight to the recent ***AN*** forecast than the ***AN*** forecasts issued at some time in the past. The hypothesis is that the ***KFAN*** forecast is as adaptable as the ***AN*** forecast (e.g. when large day-to-day changes in the prediction error are present), but unbiased as the ***KF*** forecast.

Another possibility is to run the KF algorithm through an ordered set of (all) analog forecasts, rather than in time. The entire time series of analogs is ordered from the least similar (worst analog) to the most similar (best analog) model forecast to the current one, forming an analog space for every future time t. Then, the KF is applied to the ordered set of analogs in analog space (the KF in Analog Space - ***KFAS***). The ***KFAS*** algorithm weights closeness in analog space, and not proximity in time (as the ***KFAN*** forecast). Therefore, the starting model forecast (issued in the past) that is the most similar to the current starting model forecast is given the most weight. This procedure should be able to cope even with drastic changes in both the starting model and the ***AN*** forecast error.

Model and observation datasets over the 2010-2012 period are divided into training and verification periods. The training period is from 2010 to 2011, and 2012 is used as the verification period. The training period increases gradually after every forecast. As the newer observations might be available in some real-time operational settings, they are added to the training database, together with the corresponding NWP model forecast. Therefore, the training period is initially 24 months long (for the first verified forecast initialized January 1st, 2012) and then prolonged on a daily basis up to 36 months (for the last forecast, initialized December 31st, 2012). Delle Monache et al. [2006] show that there is an improvement in skill for longer training datasets. The improvement is intense with increasing the training period, especially for training periods up to 6 months. The improvement in skill becomes less notable at around a yearlong dataset. Thus, a dataset ranging from 2 to 3 years should be long enough for this method in our opinion. Furthermore, the analog-based predictions work best with a consistent model setup. Since (operational) model setup changes

every once in a while, in our opinion it would be better to develop a methodology that can easily adapt to those changes. It is, however, possible that by using longer training dataset the prediction of rare events such as extremely strong wind would be even better.

When using the ***A8*** or the ***A2*** as the starting model, five predictors are used: wind speed and direction logarithmically interpolated to 10-m height, air temperature and relative humidity logarithmically interpolated to 2-m height, and air pressure reduced to the mean sea level. The ***DA*** does not include moist and radiation physics. Hence, only physical variables related to wind fields are included in the search for the best analogs: wind speed and direction logarithmically interpolated to 10-m height, and vorticity and divergence at the lowest vertical level (~17 m). The weight assigned to wind speed and direction is 1, and it is 0.8 for all other variables. The time window used to find the most similar analogs is defined by one time step before and after the lead time of interest. For instance, in eq. (5) $\tilde{t}$ is equal to 1, hence forming a 6-h time window for the ***A8*** and the ***DA*** models, or 2-h time window for the ***A2*** model. The time window, the predictors and the corresponding weights used to find the most similar analogs are the same for the ***KFAN*** and the ***KFAS*** as for the ***AN*** and the ***ANM***. The same recursive algorithm is used for generating the ***KFAN*** and the ***KFAS*** as for the ***KF***.

To determine if the difference in scores between the experiments is statistically significant, the bootstrap technique is applied. The Matlab function „*bootci*“, with default bias corrected and accelerated percentile method using 1000 re-samples at a confidence level of 95%, is used.

3.5. Evaluation of the wind speed as a continuous predictand

To evaluate the performance of the different deterministic post-processing methods, wind speed can be considered as a continuous or categorical predictand. Considered as a continuous variable, wind speed forecasts error is quantified by root-mean-square-error (*RMSE*), which penalizes a larger discrepancy more than a smaller one. The source of error of a model can be specified when decomposing the *RMSE* to the bias of the mean (or simply bias), the bias of the standard deviation (σ bias), and the dispersion (phase) error (e.g., Murphy [1988]; Horvath et al. [2012]):

$$RMSE^2 = (\bar{F} - \bar{O})^2 + (\sigma_F - \sigma_O)^2 + 2\sigma_F\sigma_O(1 - r_{FO}), \quad (6)$$

where F represents forecast and O observations, σ is the standard deviation, and r is the correlation coefficient between the forecast and observed data. Since the sum of the three terms in (6) is exactly the square of the *RMSE* value, it is enough to provide information about two out of these three terms to describe the dominant source of the error (the third term is the squared *RMSE* value reduced by the value of the other two terms). The term describing the dispersion error involves the Pearson correlation coefficient, weighted with the standard deviation σ of both forecasts and measurements. Correlation coefficient and dispersion error are thus closely related: the smaller the correlation coefficient, the larger the dispersion error term in RMSE decomposition. In this section, the rank correlation coefficient (*RCC*) is used as a robust and resistant alternative to Pearson correlation, appropriate if dealing with non-Gaussian distributed variables such as wind speed. Unlike the Pearson correlation coefficient, the *RCC* is a nonparametric statistic. The *RCC,* therefore, allows a nonlinear relationship between predictions and observations [Wilks, 2011; Jolliffe and Stephenson, 2011].

3.5.1. The impact of the ensemble size to the deterministic forecasting

The first step in testing an ensemble-based method is to select a number of ensemble members (*N*). For that purpose, we analyze the *RMSE* averaged over all locations and all lead times (Figure 6a). The optimal ensemble size is presented and determined for the ***A8*** starting model. The mean confidence intervals shown here are estimated with bootstrapping, as previously described.

Generally, the results are determined by the wind climate, complexity of topography, and the low resolution of the driving mesoscale model. The starting model forecasts (***A8***) yield *RMSE* of 2.35 ms^{-1}, correlation coefficient *RCC* of 0.58, and almost non-existing bias of -0.01 ms^{-1}. However, it needs to be noted that this is aggregated (averaged) bias value, therefore not necessarily implying that the ***A8*** forecast bias is small everywhere or during any time of a day. For that reason, more detailed insight is provided in the following subsections.

All tested post-processing methods, if averaged over the three studied regions, improve the results of the ***A8*** model. The ***KF*** forecast significantly reduces *RMSE* (Figure 6a), improves correlation (Figure 6b), while bias remains small (Figure 6c) when compared to the ***A8***. Using analogs improves results even further than just the ***KF***, as it can be seen for the ***KFAS***. The ***KFAS*** uses the entire analog space and therefore does not depend on the ensemble size. The other analog-based predictions (***AN***, ***ANM,*** and ***KFAN***) produce similar results as the ***KFAS*** for about 10 or more ensemble members. Furthermore, the ***AN***, the ***ANM***, and the ***KFAN*** show similar behavior – the *RMSE* is reduced at first by increasing the ensemble size, but then it increases again for more than 15 ensemble members. The correlation also improves by increasing the ensemble size, while bias slightly worsens. The mean of the observed wind speed during the verification period differs from the mean during the training period for approximately 0.2 ms^{-1}. The bias is likely converging to that value when increasing the ensemble size. Even though the biases after post-processing are significantly different from bias for the ***A8***, one should take into consideration that the bias under 0.5 ms^{-1} can be considered relatively small. It is an order of magnitude smaller than the other two terms in *RMSE* decomposition and comparable to observational error (up to 0.5 ms^{-1} or even higher; WMO, 2008). Additional uncertainty comes from the fact that some of the observation stations are subject to urban effects (heat islands, some larger-scale sheltering), while these urban effects are not represented in tested ALADIN model configurations. Given the *RMSE* and bias growth with the ensemble size, the optimal number of ensemble members is set to 15, which is used hereinafter (in section 3).

It can be noticed that the ***ANM*** experiment has the highest *RMSE* and the highest bias if different analog-based predictions are compared. Since the other analog-based predictions produce better results than the ***ANM***, and specific benefits are not achieved in tested cases presented in this work, results for the ***ANM*** are discarded hereinafter.

Both ***AN*** and ***KFAN*** considerably reduce the *RMSE* (as evident from Figure 6a), better than any other technique tested here. At the same time, they improve the correlation (Figure 6b). Both ***AN*** and ***KFAN*** have a very small negative bias, mostly between -0.1 and -0.2 ms^{-1}. The ***AN*** has slightly better correlation and worse bias results than the ***KFAN***, resulting in indistinguishable *RMSE*.

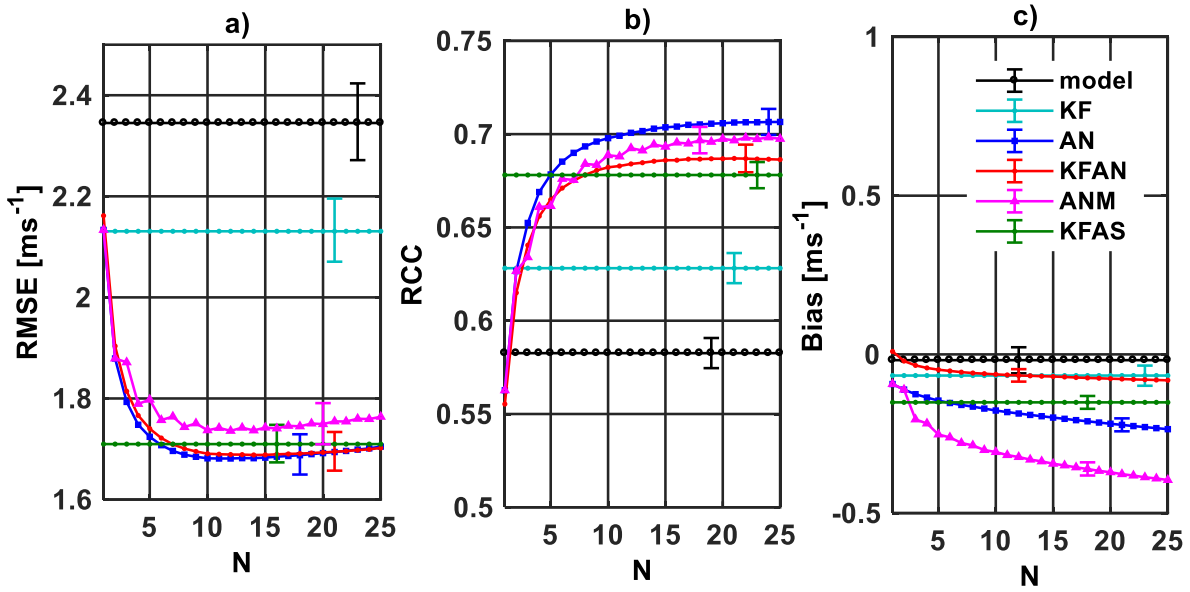

*Figure 6. a) Root-mean-square-error (RMSE), b) rank correlation coefficient (RCC), and c) bias dependency on number of analog ensemble members (N) for the **AN**, the **ANM** and the **KFAN** forecasts. The results are generated with the **A8** and averaged over all of lead times and 14 locations during 2012. The **AN**, the **ANM** and the **KFAN** results are then compared to the **A8** model, the **KF** and the **KFAS** forecasts, which do not depend on N. The mean values of the 95% bootstrap confidence intervals are indicated by the error bars.*

Since the ***KFAN*** forecast is created by applying the KF to the ***AN*** forecast, the differences between the ***KFAN*** and the ***AN*** in the correlation and bias results may be expected. The KF algorithm updates its estimate of the future bias by using the old bias plus uncertainty. The estimate is corrected by a linear function of the difference between the previous prediction and the verifying bias. It is, therefore, very successful in removing the systematic errors (such as a bias of the mean), if the bias does not change rapidly (i.e. large hour-to-hour variations). However, the application of the KF algorithm can also lead to the decrease of the correlation coefficient (i.e. an increase of the dispersion error), especially if there are large hour-to-hour bias variations [Delle Monache et al., 2006; 2008].

3.5.2. Lead time performance for different topography types

A more detailed insight into the performance of the post-processing methods can be gained by analyzing the metrics in topographically different regions and at different lead times.

The first step is to analyze the ***A8*** performance in the coastal complex topography. The ***A8*** model has the highest *RMSE* for the coastal complex topography among all groups of stations (Figure 7a). Besides the increasing trend for longer lead times, the ***A8*** *RMSE* error is typically the largest during nighttime and peaks at 06 UTC in the coastal area. While during nighttime the ***A8*** exhibits maximum correlation (Figure 7e), it underestimates the mean (Figure 7i) and underestimates the standard deviation σ (Figure 8a) more than during the daytime. While observed wind speed shows the highest variability at 06 UTC (Figure 8a), the ***A8*** forecast almost does not show the standard deviation σ diurnal cycle. That result suggests a systematic source of the errors for the diurnal shape of ***A8*** *RMSE* (Figure 7a). It is possible that the ***A8*** model underestimates land breeze, the combination of land breeze and downslope wind called burin [Poje, 1995] or underestimates both mean speed and variability of the strong bora wind, which can be determined with analysis by season (e.g., bora occurs mostly during wintertime and it is variable and intense, while land breeze can be dominant during summertime stable conditions) or by examining case studies.

It is crucial to determine which post-processing method is the most successful in the error reduction, especially in this particular group of stations where the error is the largest. Additionally, it is important to demonstrate which term of the *RMSE* decomposition is reduced by which post-processing method. For that reason, the performance of different post-processing methods in the coastal complex topography will be presented in the next paragraph. The results are presented in such a manner that one can thus decide which post-processing method is the most applicable for a specific situation, after a simple statistical analysis of the potential starting model.

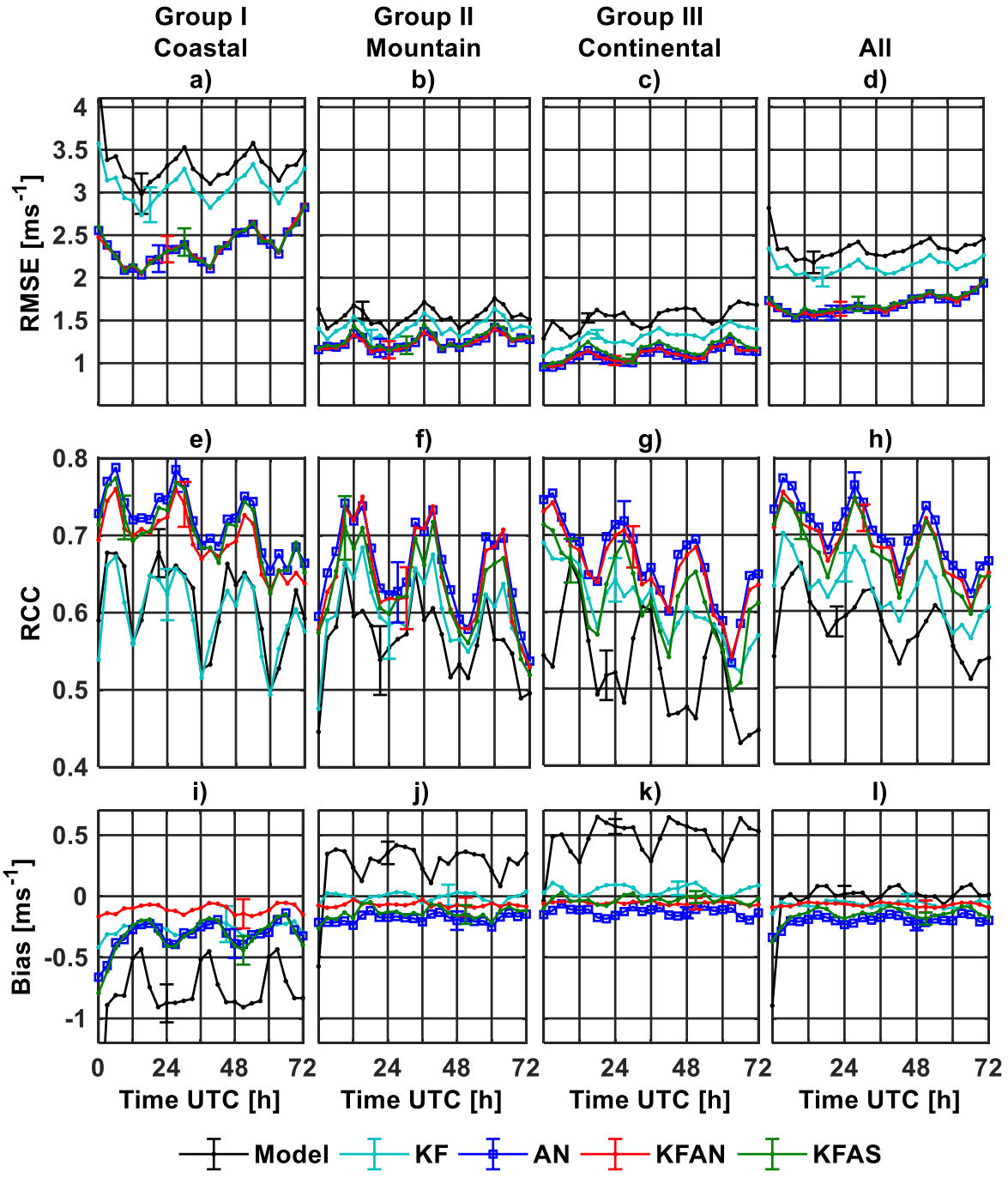

*Figure 7. Root-mean-square-error (RMSE)(a-d), rank correlation coefficient (RCC) (e-h) and bias (i-l) dependency on forecast lead time for the **A8** starting model and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**). The results are averaged over the corresponding groups and for 14 locations in Croatia during 2012. The mean values of the 95% bootstrap confidence intervals are indicated by the error bars.*

Secondly, we aim to answer how well does the ***KF*** reference method perform against the ***A8*** model and against other analog-based experiments in the coastal complex topography. The ***KF*** reduces *RMSE* and bias (Figure 7a and Figure 7i) while increases standard deviation σ (Figure 8a), maintaining very similar dependency on lead time as the ***A8***. The other analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) improve the ***A8*** results even further – reducing *RMSE* and bias while standard deviation σ is even closer to the standard deviation of the measurements. Moreover, even though the standard deviation is still a bit underestimated, the diurnal cycle of the standard deviation is more similar to the diurnal cycle of the measurements than for the ***A8***. Previously mentioned systematic ***A8*** error (possibly unresolved land breeze, underestimation of burin wind, etc.) is thus reduced or removed completely. The standard deviation of the analog-based predictions is very close to the standard deviation of the measurements available over the training period. The analog-based predictions underestimation of the standard deviation is, therefore, partially explained by the fact that there is a standard deviation difference between training and testing period. Also, in the coastal complex area, the ***KF*** has a smaller correlation coefficient (*RCC*) than the ***A8***, unlike all the analog-based predictions which have a higher correlation coefficient than the ***A8***. Improving the correlation shows that by using analogs and measurements to build a prediction the random error is reduced, suggesting that additional information on physical processes is included in the analog-based predictions.

After the general comparison of the analog-based predictions against the reference method ***KF*** in the coastal complex topography, we will take a more detailed look into the differences among analog-based predictions for this group of stations in the next three paragraphs. We will focus on the underestimation of the standard deviation and the ability of the analog-based predictions to reduce random error (i.e. increase the correlation).

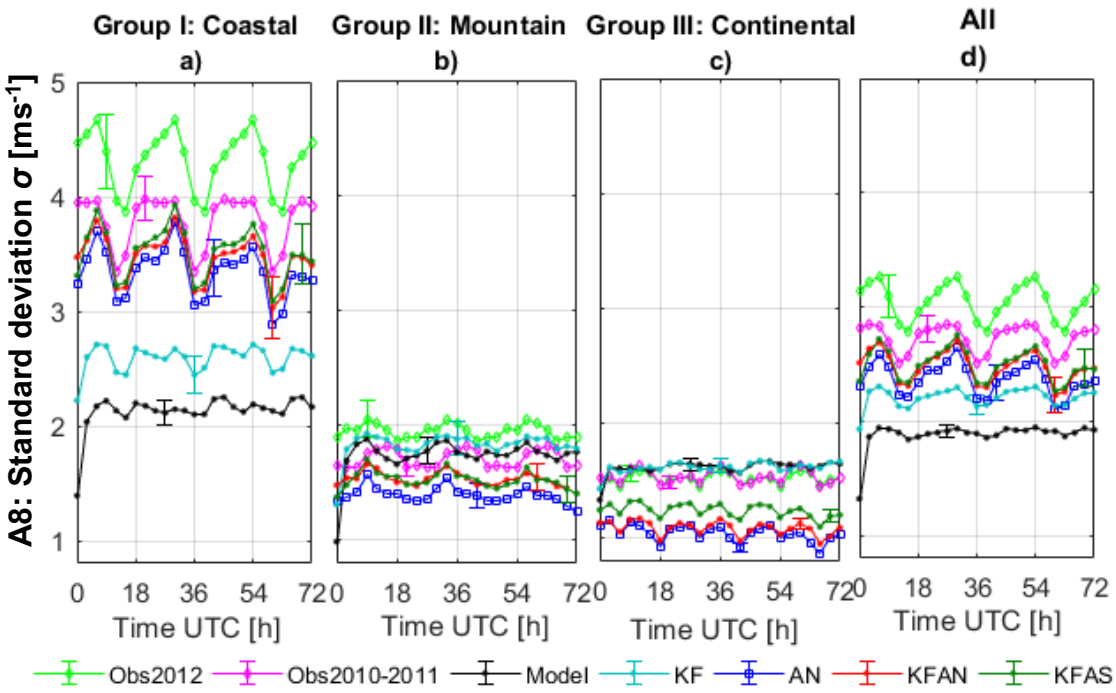

*Figure 8. The dependency of the standard deviation on forecast lead time for the observations during the training (2010-2011) and the verification period (2012), the **A8** starting model and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**). The results refer to the corresponding groups (a-c), and to 14 locations in Croatia (d) during 2012. The mean values of the 95% bootstrap confidence intervals are indicated by the error bars.*

Among the analog-based predictions, the ***AN*** forecast is the most prone to systematic underestimation of the standard deviation (Figure 8) in the coastal complex topography (but also in general). This reduction of the forecast variability is due to averaging of AnEn members while predicting the mean of the ensemble. This averaging naturally reduces the variability and might partially be improved by using the lower number of ensemble members. This systematic error is partially removed by the application of the KF algorithm in the ***KFAN*** forecast.

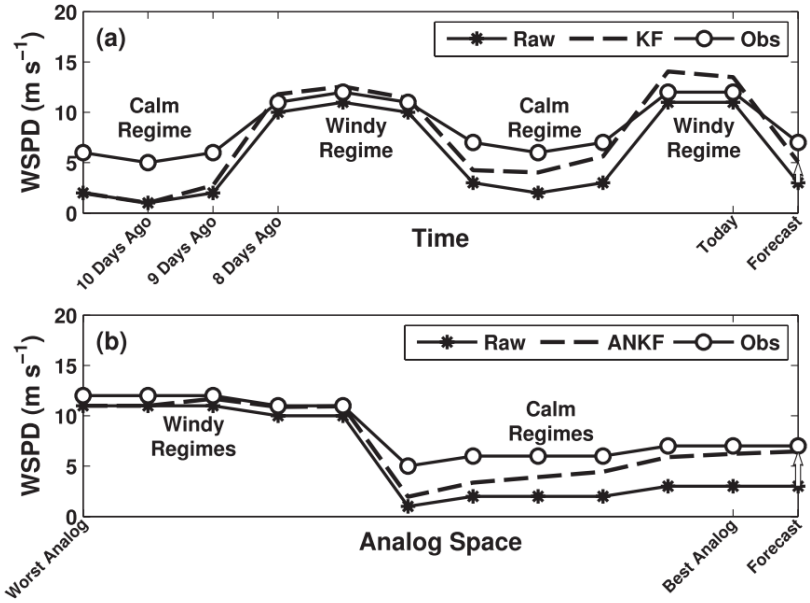

*Figure 9. Schematic representation of a Kalman filter correction for wind speed prediction (WSPD) (a) run in time (**KF**) or (b) through an ordered set of analog forecasts (ANKF; equivalent to the abbreviation **KFAS** used in this thesis). White arrows at forecast time (far right) indicate the post-processing methods estimate of the forecast error. Circles indicate observations, asterisks refer to the raw prediction, and the dashed line represents the corrected predictions (from Delle Monache et al. [2011], page 3557).*

The ***KFAS*** forecast, on the other hand, exhibits the highest standard deviation among the analog-based predictions in the coastal complex topography and in general. This is worth additional discussion. The simplified schematic example for improving the adaptability of the ***KFAS*** forecast is provided in Delle Monache et al. [2011], as shown in Figure 9. The hypothesis is that applying the KF algorithm in analog space (rather than in time), results in higher forecast variability during alternating wind regimes. The higher ***KFAS*** standard deviation than the ***KFAN*** standard deviation in the coastal area supports this hypothesis. The difference in the standard deviation between the ***KFAN*** and the ***KFAS*** does not necessarily mean that the higher variability for the ***KFAS*** is occurring during alternating wind regimes (i.e. on the time scales shorter than a day). The remaining underestimation of standard deviation depends on other aspects such as the variability of starting model forecasts and fine-tuning of the analog search setup (e.g., choice of predictors, corresponding weights, as shown by Junk et al. [2015]). The variability in the training period might be enlarged by prolonging

the period itself (i.e. including El Niño/Southern variations). Finally, the variability of the post-processed forecasts, in general, might be further improved by additional calibration. For example, applying the ensemble model output statistic post-processing approach (EMOS; [Gneiting et al., 2005]) on the analog forecasts or directly combining the two methods might be a possible future research avenue.

Among different analog-based predictions, the ***AN*** seems to have the highest correlation, while the ***KFAN*** reduces the bias the most, as previously described in the more general case. The ***KFAS*** exhibits the highest standard deviation among the analog-based predictions, supporting the hypothesis that using the analog space improves variability during alternating wind regimes. After all, there are no significant differences in the reduction of *RMSE* for the ***AN,*** the ***KFAN***, and the ***KFAS***.

After analyzing the forecasts in the coastal complex area, we will shift our focus to the other topography types. We will also start by examining the starting model ***A8*** performance. The ***A8*** exhibits considerably smaller *RMSE* for the mountain complex (Figure 7b) and nearly flat topography (Figure 7c) than it is the case for the coastal complex area (Figure 7a). The smaller ***A8*** *RMSE* is predominantly due to lower, less underestimated standard deviation of measured wind speed for these groups (Figure 8b-c) than for the coastal complex topography. Even though the ***A8*** error is smaller than in the coastal complex topography, it is still very important to determine which term in the *RMSE* decomposition is dominant and how it can be reduced by post-processing. Unlike underestimation of (on average) higher wind speed in the coastal topography, the ***A8*** overestimates (on average) lower wind speed in the mountain complex (Figure 7j) and the nearly flat topography (Figure 7k), exhibiting the similar absolute value of the bias. The ***A8*** standard deviation is much closer to measured wind speed standard deviation for the mountain complex (Figure 8b) and the nearly flat (Figure 8c) than the coastal complex topography. The ***A8*** correlation coefficient (*RCC*) is lower for the mountain (Figure 7f) and for the nearly flat (Figure 7g) than for the coastal complex topography, therefore decreasing with measured mean wind speed and corresponding standard deviation. It seems that the lower the average wind speed for a certain group, the lower the correlation of measurements and predictions, implying that weak wind is less predictable than a strong one. This especially makes sense for wind speeds that are comparable to observational error (up to 0.5 ms^{-1} or even higher; WMO [2008]). In other words, models are more successful in

simulating winds which are due to stronger forcings i.e. pressure gradients, than weak winds in non-gradient situations.

Even though some statistical properties of the ***A8*** predictions are similar for the mountain and nearly flat topography, the physical processes influencing the flows are different. This is due to different dominant topographic characteristics, as explained in section 3.1. For this reason, it is interesting to compare the effect of post-processing in a certain group of stations. We will start by examining the ***KF*** performance for different topography types. The ***KF*** forecast exhibits significantly lower *RMSE* than the ***A8*** in the mountain and nearly flat topography. The ***A8*** bias is almost completely removed by the ***KF***, regardless of the topography type and if the ***A8*** is underestimating (Figure 7i) or overestimating (Figure 7j-k) wind speed. The ***KF*** standard deviation σ in the mountain and the nearly flat topography is almost the same as the ***A8,*** and very close to measured σ as well. In addition to reducing the ***A8*** bias of the mean and maintaining bias of the standard deviation almost non-existent, the ***KF*** also improves the correlation for all of the lead times in the mountain and the nearly flat topography. Unlike for the coastal complex, dispersion error is therefore reduced by the ***KF***, especially for the nearly flat topography. Furthermore, the ***KF*** forecast dependency on lead time is different than for the ***A8*** in the nearly flat topography. The ***KF*** forecast exhibits a local correlation coefficient maximum around 00 UTC, while the ***A8*** exhibits a local minimum (Figure 7g).

After examining the ***KF*** performance in different topography types, we will compare those results against the analog-based predictions. The analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) in the mountain complex and the nearly flat topography reduce the ***A8*** *RMSE* even more than the ***KF*** forecast, further improving correlation and reducing bias. The *RMSE*, correlation and bias dependencies on a lead time are similar as for the ***KF***. This is especially interesting in the nearly flat topography, where previously mentioned improvement of the ***A8*** correlation coefficient *RCC* is even more indicated when using analogs than for the ***KF***. The analog approach selects similar numerical predictions (not necessarily recent) for assessment of the starting model error, unlike non-selectively using previously predicted (recent) values in the KF algorithm. The KF would be capable of improving persistent error in predicting stable boundary layer flow once it is started, as previously mentioned for the application of the KF algorithm. The analog-based method would be more adaptable and capable of

predicting the beginning of the flow, thus resulting in an even higher correlation coefficient than for the KF.

We will now take a detailed look into the analog-based predictions performance in different topography types. Similarly to the coastal complex, in the mountain complex and the nearly flat topography the ***AN*** seems to be the most highly correlated with measurements. The ***KFAN*** has a slightly lower correlation coefficient *RCC* but is almost unbiased. Unlike the ***A8*** and the ***KF***, the analog-based predictions exhibit a slight underestimation of σ in the mountain complex (Figure 8b) and nearly flat topography (Figure 8c). The underestimation of the standard deviation is the smallest for the ***KFAS*** and the largest for the ***AN***, for the same reasons as previously mentioned. The results for the ***KFAN*** are mostly in between these two (***AN*** and ***KFAS***), which may be explained by the fact the ***KFAN*** shares important features with both methods.

Finally, we will try to summarize the previous analysis by aggregating results for all available stations, regardless of the topography type. Overall, the ***A8*** *RMSE* is significantly reduced by every post-processing method tested for all of the lead times, more by the analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) than for the ***KF*** (Figure 7d). All post-processing methods reduced the ***A8*** bias, which is evident for a specific group and lead time (Figure 7i-k), even though it seems non-existent on average for the ***A8*** (Figure 7l). The ***KFAN*** predictions seem to be the most successful in removing bias, while the ***AN*** appears to exhibit the highest correlation (Figure 7h). Measured wind speed standard deviation σ is underestimated on average by the ***A8*** model and all post-processing methods (Figure 8), mostly due to the underestimation of standard deviation in the coastal area (group I). Overall, the standard deviation of ***KFAS*** is the closest to the observed value.

3.5.3. The influence of the starting model

To investigate the influence of the starting model used to generate analogs, results are averaged over all lead times for every group of stations. A reasonable hypothesis could be that the more physical processes that are directly simulated in the starting model (e.g., with higher resolution), the better the forecast will be. The *RMSE* (Figure 10a) and bias (Figure 10i) are lower for the ***A2*** and the ***DA*** models than for the ***A8*** in the coastal complex topography,

empirically supporting this hypothesis. The correlation coefficient *RCC* does not differ significantly among different models (Figure 10e). It must be noted that it is difficult to quantify the improvement of more detailed forecasts over coarser ones using point-based verification metrics [Rossa et al., 2008; Jolliffe and Stephenson, 2006]. Point-based verification metrics tend to penalize spatial and phase errors, contaminating finer resolution simulations more than coarser ones. Hence, it might be challenging to easily demonstrate the true benefits of using a higher-resolution forecast. To determine if that is the case, it would be advisable to do case studies and some sort of spatial verification (for gridded forecasts). The selection of bora and sirocco case studies might provide an interesting insight into post-processing performance benefits of using high resolution (i.e. prediction of extremely high wind speed). This is especially the case if the experiments are provided using (even the simple) NWP model but with a more similar setup, preferably changing the resolution and making only the necessary adjustments. Furthermore, using the gridded forecasts and analysis in the analog search, as well as the spatial verification tool, is an inevitable future development. Since the computational efficiency needs to be adequate, the analog approach might also include other methods (such as clustering, using empirical orthogonal functions, etc.).

All post-processing methods tested in this section improve model predictions. It is to be expected that the analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) also achieve better results when using the ***A2*** or the ***DA*** than when using the ***A8***. The quality of an analog should increase the better the representation of physical processes simulated in the starting model (i.e. with higher resolution, non-hydrostatic dynamics in the ***A2***, etc.). This type of improvement is clearly evident, for example, for the ***AN*** results in the coastal complex topography. The results show that the differences in using different starting model configurations are much smaller after post-processing than for three starting models. However, the *RMSE*, correlation, and bias scores are similar for the post-processing methods applied to all three starting models. For some scores, such as the *RMSE*, the analog-based predictions have the best results when applied to the ***A8*** model.

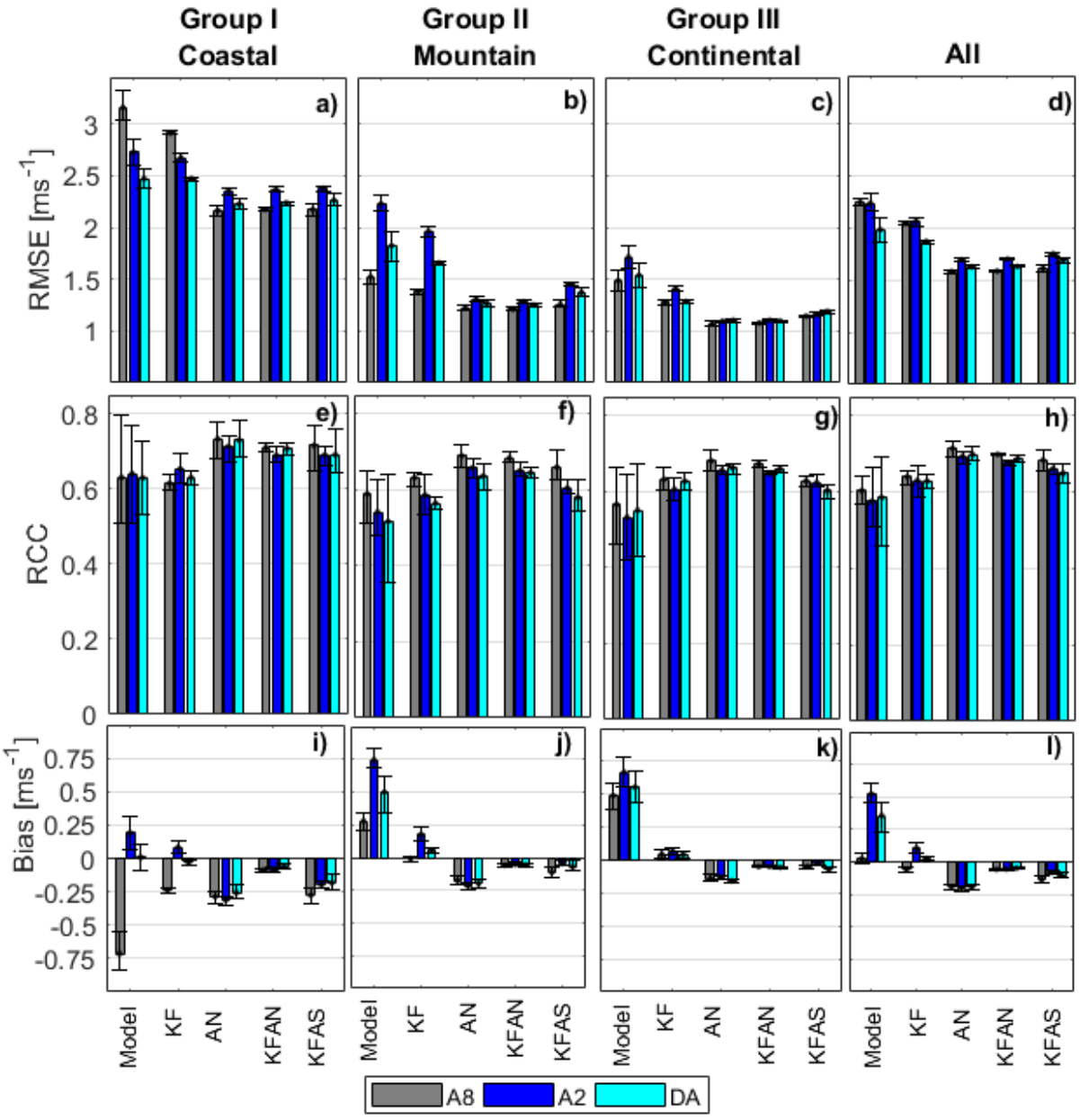

*Figure 10. The average root-mean-square-error (RMSE) (a-d), rank correlation coefficient (RCC) (e-h) and bias (i-l) for three different starting models and the corresponding post-processing methods (**KF**, **AN**, **KFAN**, and **KFAS**). The results are averaged over the corresponding groups and over all stations in Croatia during 2012. The colors represent the starting model used (**A8**, **A2**, and **DA**), while the x-tick labels stand for model and corresponding post-processing methods. The values of the 95% bootstrap confidence intervals are indicated by the error bars.*

Finally, it seems that even though the higher resolution ***A2*** and ***DA*** models achieve better results if the results are averaged over all available stations. However, the analog-based predictions based on the ***A2*** and the ***DA*** do not statistically outperform the analog-based predictions based on the ***A8*** (Figure 10d, 10h and 10l). This does not necessarily mean that improvement is not made at all. The benefits might be partially hidden because of the imperfections of the verification metrics used. To investigate the benefits of using higher resolution further, one can analyze the forecasts categorically (i.e. to examine the forecasts of the rare events such as strong wind), perform a spectral analysis or look at the case study. The categorical verification results and spectral analysis are presented in the next two sections, while it is previously discussed how the case studies are a possible future research avenue.

3.6. Evaluation of the wind speed as a categorical predictand

To verify a categorical predictand the event or events need to be pre-defined. Wind speed is therefore divided into 3 categories: weak (or no wind at all), moderate and strong wind, depending on the climatology of the corresponding group of stations. Thresholds are determined as the 50^{th} and 90^{th} percentile of the entire group. This is done independently for each lead time, so the thresholds vary due to the diurnal cycle (Figure 3). After defining categories (events), the next step is the calculation of a so-called contingency table (Table 3). The forecast-observation pairs corresponding to the same (real) time populate the contingency table, representing the joint distribution (i.e. fields A-I in Table 3). At the right side and the bottom, the marginal distributions are also shown (Fields J-P in Table 3).

The categorical verification procedure includes frequency bias (*FBias*), critical success index (*CSI*) and polychoric correlation coefficient (*PCC*). The choice of these measures is consistent with the continuous case.

Table 3. The example of a contingency table

Wind speed predictor		Observed			
		Below 50th percentile	Between the 50th and 90th percentile	Above 90th percentile	Total
Forecast	Below 50th percentile	A	B	C	J
	Between the 50th and 90th percentile	D	E	F	K
	Above 90th percentile	G	H	I	L
	Total	M	N	O	P

The polychoric correlation coefficient *PCC* measures the association of forecasts and observations in the contingency table. The idea behind the *PCC* is to assign a density function to the contingency table and then cut the domain into rectangles corresponding to the cells of the contingency table (Figure 11). The *PCC* is the parameter value of the standard bivariate normal density function for which the volumes of the discretized distribution are equal to the corresponding joint probabilities of the contingency table [Juras and Pasarić, 2006]. The standard bivariate normal density function is completely determined by one parameter (*PCC*), while the mean value is set to 0 and the standard deviation parameter is set to 1. However, it is not applied to the latent (i.e. underlying continuous) variables directly, but to corresponding standard normal deviates Z_X using the following transformation for the continuous variable X:

$$Z_X = \Phi^{-1}\left(\Phi_X(X)\right), \tag{7}$$

where the Φ_X represents the cumulative distribution function of X, while the Φ is the cumulative distribution function of standard normal distribution. Having the contingency table, it is implicitly accepted that we are dealing with categorical variables, which in our case are observation (O) and forecast (F). It is assumed that the random vector (Z_O, Z_F) follows the bivariate normal density function. Similarly, the thresholds between different categories are also transformed.

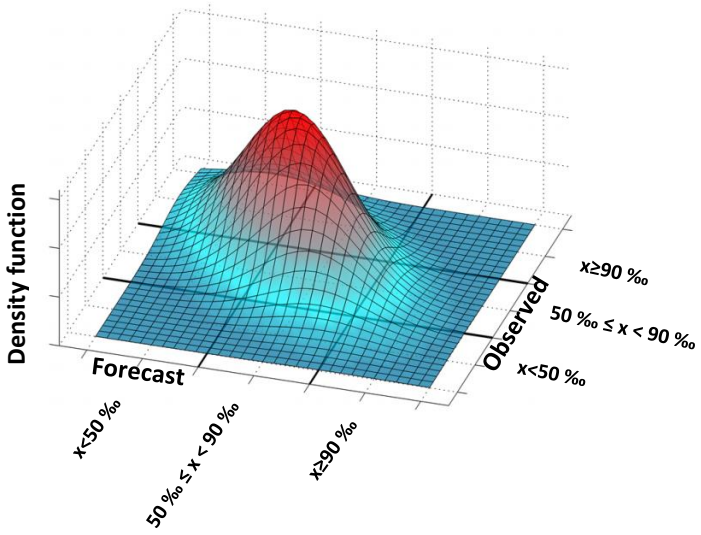

Figure 11. Standard bivariate normal density function for which the volumes of the discretized distribution are equal to the corresponding joint probabilities of the 3×3 contingency table.

For dichotomous forecasts, the *PCC* is called the tetrahoric correlation coefficient. If the z_o and z_F are standard normal deviates of the marginal probabilities, the relation between the tetrahoric correlation coefficient and the A field of the 2×2 contingency table is uniquely determined:

$$A = \frac{1}{2\pi} \int_{\arccos(TCC)}^{\pi} \exp\left[-\frac{1}{2}(z_O^2 + z_F^2 - 2z_O z_F cos\omega) cosec^2\omega\right] d\omega \,. \tag{8}$$

For the higher-order (i.e. 3×3) contingency tables, the relation between *PCC* and *A* is not unique. Nevertheless, it can be approximated. The conditional maximum likelihood method is used in this research. Additional details of the procedure applied in this research can be found in Juras and Pasarić [2006].

The range for the polychoric correlation coefficient *PCC* is between -1 and 1. The *PCC* value for the random forecast is defined as 0, while it is undefined for the constant forecast. The measure does not depend on the underlying climatology for the pre-defined events. For this reason, it is suitable for comparison among climatologically different regions.

Ekström [2011] shows the (asymptotical) equivalence of the rank correlation coefficient *RCC* and the polychoric correlation coefficient *PCC* under several conditions including that the number of categories is as large as the number of measurement-forecast pairs, the

underlying joint distribution is binormal, etc. Even though a "simplified" rank correlation coefficient *RCC* can be re-calculated if the ordinal variables arise from discretization such as groupings of values into categories (as in this section), it has some undesirable properties. For instance, it can achieve a value of 1 even if non-discretized empirical variables are not perfectly dependent. The polychoric correlation coefficient *PCC* is therefore considered to be more conservative and better suited for statistical inference about the association of the underlying, non-discretized variables than the rank correlation coefficient *RCC*.

The frequency bias *Fbias*, similarly to bias, measures the tendency to forecast too often (*FBias* greater than 1) or too rarely (*FBias* less than 1) a particular category [Wilks, 2011; Jolliffe and Stephenson, 2011]. In other words, it is the ratio of the number of forecasted events and the number of occurred events, calculated separately for each category, as follows:

$$FBias_1 = \frac{J}{M};\ \ FBias_2 = \frac{K}{N};\ \ FBias_3 = \frac{L}{O}. \tag{9}$$

The *Fbias* provide the information about the forecast distribution (i.e. whether the event is under- or over-forecasted) and not the forecast accuracy. For example, the persistence forecasting (forecasting the last measured value) is almost completely unbiased. However, it is often not accurate and it lacks skill.

The critical success index *CSI* measures the fraction of observed forecast events that are correctly predicted. It can be thought of as the relative accuracy when correct negatives are removed from consideration. It is computed from the contingency table, separately for each category, as follows:

$$CSI_1 = \frac{A}{J + M - A};\ \ CSI_2 = \frac{E}{K + N - E};\ \ CSI_3 = \frac{I}{L + O - I}. \tag{10}$$

The *CSI,* therefore, measures the error (similar to the RMSE in continuous case). Sensitive to hits, the *CSI* penalizes both misses and false alarms. It does not distinguish the source of forecast errors and hence additional verification measures need to be examined [Wilks, 2011; Jolliffe and Stephenson, 2011]. The *CSI* value ranges from 0 to 1. Ideally, it is equal to 1, which means there is not a single false forecast.

Assessing the quality of predictions of extreme weather events is complicated by the fact that measures of forecast quality typically degenerate to trivial values as the rarity of the predicted event increases. The extremal dependence index *EDI* is a measure developed for the extreme weather events verification independent on underlying climatology [Ferro and

Stephenson, 2011]. It is a function of the false alarm rate *F* and hit rate *H* and is calculated as follows:

$$H = \frac{I}{O};\ F = \frac{L - I}{P - O};\ EDI_3 = \frac{\log F - \log H}{\log F + \log H}. \tag{11}$$

The EDI_3 is of use when the aim is to assess the quality of the forecast for discriminating the antecedent conditions leading to the occurrence of extreme weather from those which do not (i.e. discrimination property). It is a regular, asymptotically equitable measure that is difficult to hedge and always has range [−1, 1]. The value for the perfect forecast is 1.

3.6.1. The association of forecasts and observations in the contingency table

The polychoric correlation coefficient *PCC* results for different forecasts (Figure 12a-d) do resemble the rank correlation coefficient *RCC* results (Figure 7e-h) when results are averaged for all of the lead times in a certain group. The ***DA*** and the ***A2*** exhibit higher association in the coastal complex but not in the other topography types. Association is significantly improved by almost all post-processing methods in all groups of stations and overall, as already presented. The exception is the ***KF*** forecast in the coastal complex topography. The analog-based predictions achieve better both rank and polychoric correlation coefficient results than the ***KF*** in general, particularly the ***AN***. There are some differences between the rank correlation coefficient *RCC* and polychoric correlation coefficient *PCC* results that need to be highlighted in order to determine the origin; whether it is due to statistical properties of the verification measure used or it is a direct consequence of discretization (i.e., the grouping of wind speed into 3 categories). If both coefficients are calculated for the same (ordered) data and grouped into identical categories, the rank correlation coefficient *RCC* would have a slightly higher value [Ekström, 2011]. The polychoric correlation coefficient *PCC* shows higher values than the rank correlation coefficient *RCC* calculated for the continuous variable, hence confirming the assumption that it is easier to predict the category than the exact (continuous) value of wind speed.

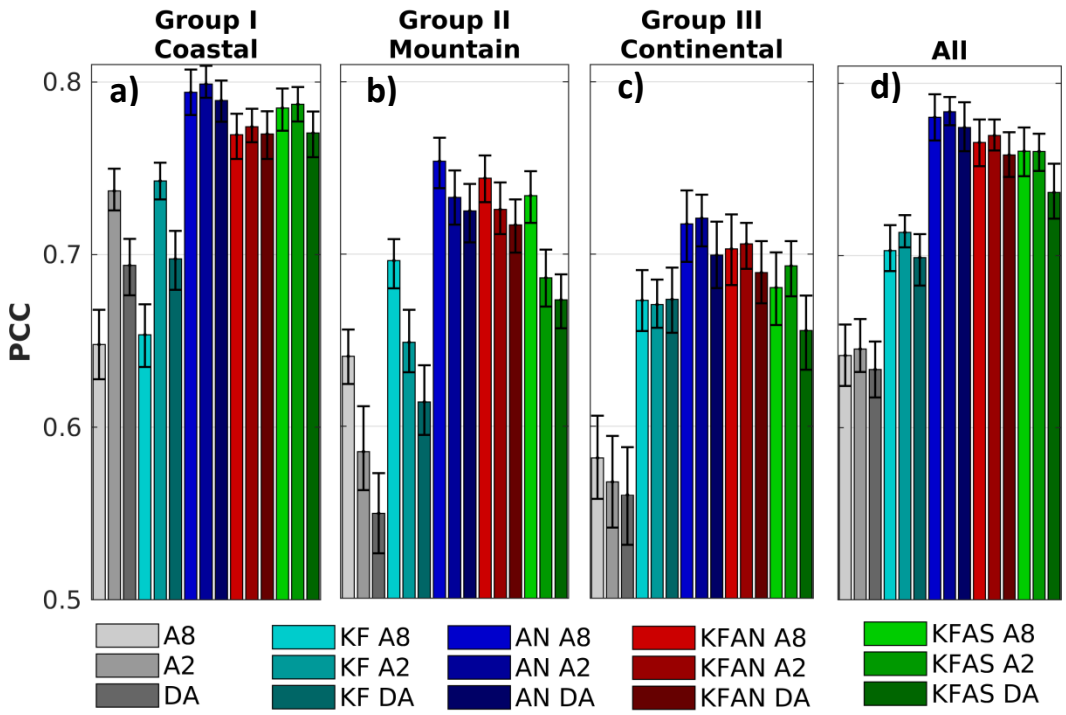

*Figure 12. The polychoric correlation coefficient (PCC) for three different starting models (**A8**, **A2**, and **DA**) and the corresponding post-processing methods (**KF**, **AN**, **KFAN**, and **KFAS**). The results are averaged for the corresponding groups (a-c) and for all (d) of the locations in Croatia during the year 2012. The PCC is calculated using three different categories, divided by the 50th and 90th percentile. The values of the 95% bootstrap confidence intervals are indicated by the error bars.*

3.6.2. Frequency bias

There is a variety of frequency bias (*Fbias*) results depending on the exact model, group of stations and wind category (Figure 13). For instance, the ***DA*** predicts category 2 too often (Figure 13e), while predicting the other two categories (Figure 13a and 13i) too rarely in the coastal area. The frequency bias results for the ***A8*** model are somewhat similar, while the ***A2*** is almost unbiased in this case. All starting models under-forecast weak wind category while over-forecast moderate and strong wind categories in the mountain complex and the nearly flat topography (Figure 13f, 13g, 13j and 13k). The exact values differ for different models and categories yielding mixed results in terms of determining the best-performing starting model.

The ***KF*** only slightly impacts the ***A8*** frequency bias by decreasing the value for the weak wind category (Figure 13a), while only indicating the increased value for the moderate and the strong wind categories (Figure 13e and Figure 13i) in the coastal area. More generally, besides the frequency bias reduction for the weak wind category (Figure 13a-Figure 13d), the ***KF*** does not have a noticeable impact on the starting model results. Unlike the coastal area, in the mountain complex and the nearly flat topography, the ***KF*** seems to be less biased than the corresponding (starting) model for all cases tested. This is indicated by the significantly smaller bias for the weak wind category, and smaller confidence intervals near the zero value for the moderate and strong wind categories. The smaller confidence interval referring to the same sample size means smaller variability within the results.

The frequency bias results for the analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) seem to exhibit much less variety depending on a different group of stations. The results are indistinguishable among different starting models, especially for the moderate and strong wind categories (Figure 13e-Figure 13l). For any given group, the analog-based predictions consistently over-predict moderate wind speeds (Category 2), while under-predict rarer and stronger wind (Category 3). These analog-based predictions sometimes even under-predict the occurrence of weak wind. The ***KFAS*** seems to be the least biased analog-based prediction, showing the highest values for strong wind category while being as unbiased as the ***AN*** in the other two categories. However, it needs to be mentioned that these differences are not statistically significant, partially due to the small sampling size.

Overall, the post-processed forecasts, in general, reduce bias for the climatologically most common wind speed category (weak wind). The analog-based predictions frequency bias results are not as variable as for the starting model and the ***KF***, inheriting only a slight difference from the corresponding model for an exact technique (***AN***, ***KFAN*** or ***KFAS***). The main deficiency of the post-processing methods seems to be under-forecasting the occurrence of strong wind, with the ***KFAS*** being the most successful (Figure 13l).

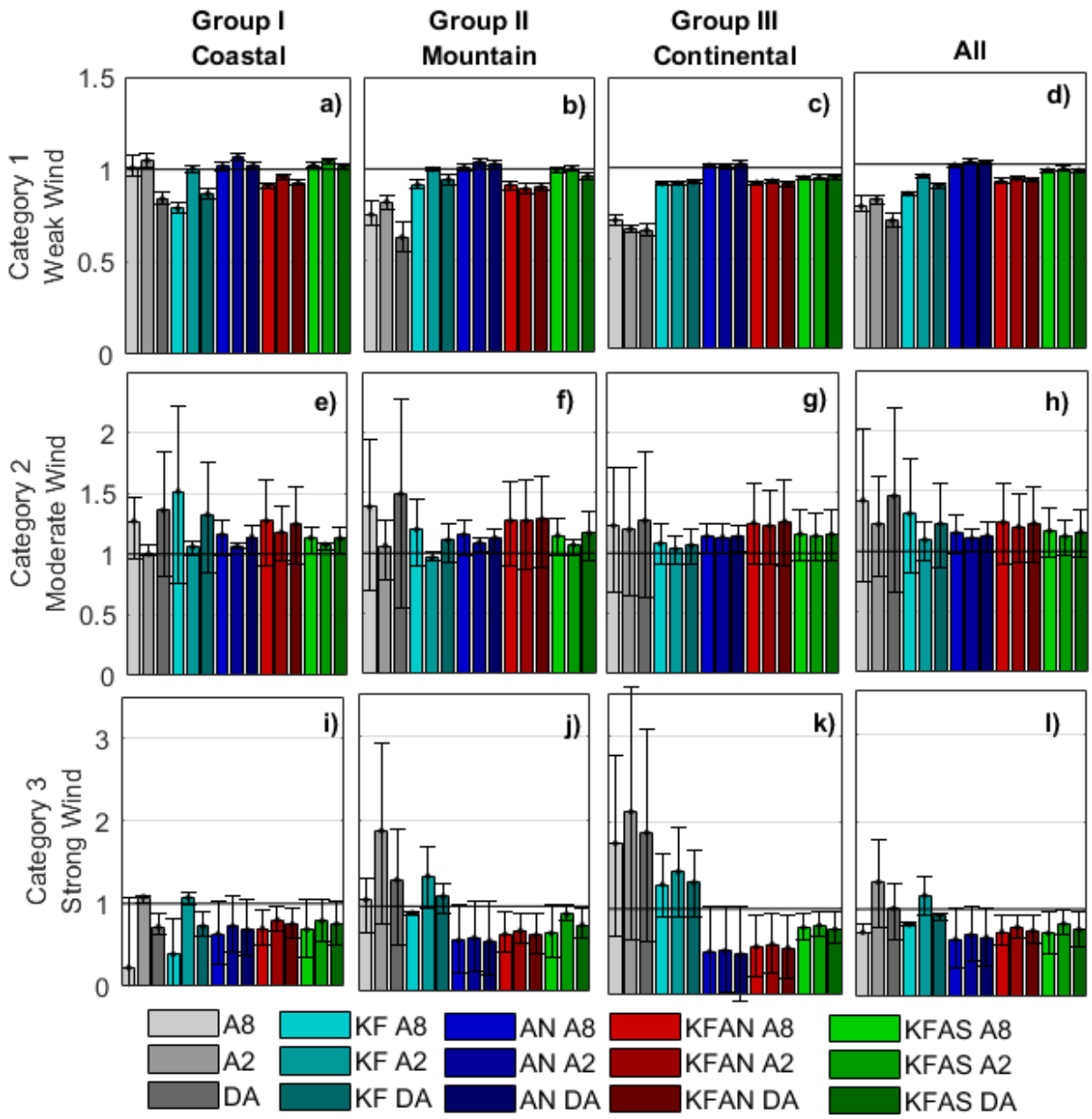

*Figure 13. The frequency bias (Fbias) for three different starting models (**A8**, **A2**, and **DA**) and the corresponding post-processing methods (**KF**, **AN, KFAN**, and **KFAS**). The results are averaged for the corresponding groups and all of the locations in Croatia during the year 2012. The Fbias is calculated for three different categories, divided by the 50th and 90th percentile. The values of the 95% bootstrap confidence intervals are indicated by the error bars.*

3.6.3. Evaluation of the forecast quality

If results among different starting models are compared, it can be seen that for the weak wind category the ***A2*** produces higher critical success index *CSI* than the ***A8*** and the ***DA*** (Figure 14a). Furthermore, finer horizontal resolution slightly improves relative accuracy for the strong wind category in the coastal complex topography (Figure 14i). The results for the moderate wind category are similar across the different starting models (Figure 14e). Increasing the horizontal resolution does not necessarily improve the critical success index in other groups of stations. Due to the small sample size, the results rarely differ significantly.

The critical success index results are considerably higher for the ***KF*** than for the starting models (***A8***, ***A2***, and ***DA***) for the weak wind category in the mountain complex and the nearly flat topography, but not as much in the coastal area. The indication ***KF*** being the most successful in predicting the strong winds (Category 3) in nearly flat continental topography, even though not statistically significant, might still suggest a dominant systematic error in the models' predictions of the strong wind. The frequency bias is lower for the ***KF*** than for any starting model, which combined with a higher critical success index indicates that the number of false alarms is reduced.

Analysis suggests that analog-based predictions outperform starting models and corresponding ***KF*** forecasts for all of the categories and all groups of stations except the strong winds in the nearly flat continental topography (Figure 14k). The improvement of the critical success index value is the most evident, and statistically significant, for the most common weak wind category (Figure 14a-d). However, the larger sample is needed to provide a more rigorous proof of that statement for the moderate and strong wind.

Overall, all post-processing methods improve the critical success index value. The ***AN*** forecasts achieve the best result for predicting weak wind (Figure 14d), while the ***KFAN*** and the ***KFAS*** produce slightly better results than the ***KF*** and the ***AN*** for the other two categories (Figure 14h and 14l).

It needs to be noted that the results for the moderate and strong wind speed categories are rarely statistically significant, partially due to the small sample size. However, analysis suggests that the best results are achieved when using the ***A2*** as the starting model, mostly due to the higher critical success index in the coastal complex topography than when using a coarser resolution starting model. It is possible that additional improvements may be

generated by increasing the resolution (1 km or less) in the complex topography. The necessity to use even 2-km grid spacing is, however, questionable and might be reexamined for nearly flat continental topography (i.e. by spectral analysis). In addition to improving the relative accuracy in coastal complex topography, the categorization suggests the higher association for the full-physics ***A2*** model and corresponding post-processing methods in the coastal complex and the nearly flat continental topography, as shown before. These results combined might suggest that the higher resolution full-physics ***A2*** model is better capable to distinguish low from moderate or unusually strong wind, especially in the coastal complex topography. This capability is then mostly inherited by the different post-processing methods, including the analog-based predictions.

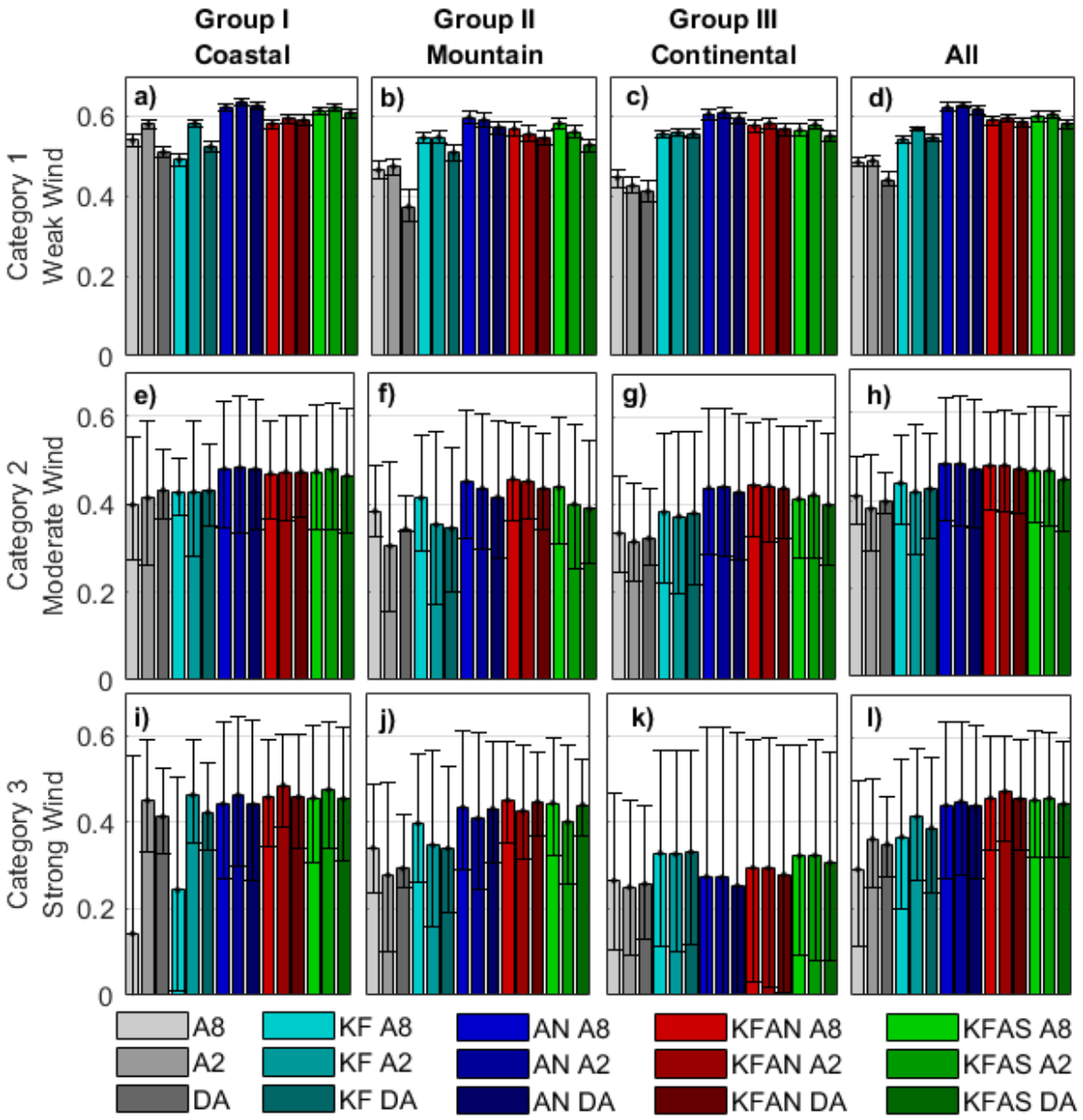

*Figure 14. Critical success index (CSI) for three different starting models (**A8**, **A2**, and **DA**) and the corresponding post-processing methods (**KF**, **AN**, **KFAN**, and **KFAS**). The results are averaged for the corresponding groups and for all of the locations in Croatia during the year 2012. The CSI is calculated for three different categories, divided by the 50th and 90th percentile. The values of the 95% bootstrap confidence intervals are indicated by the error bars.*

There is a decrease in the critical success index values for moderate (Category 2) and in particular strong wind (Category 3), regardless of the exact group of the stations or the

forecast. It should be mentioned that that decrease is partially the direct consequence of sensitivity of the critical success index metrics to the climatological probability of the predefined category that is being evaluated, and therefore it should be analyzed with caution. The sensitivity to climatology is due to counting the portion of correct forecasts that can be accurately predicted by random chance. Also, the different values across different groups for the same category (e.g., strong winds at Figure 14i-k) might suggest that unusually strong and rare wind is predicted more easily in coastal than in continental area, regardless of the exact forecast.

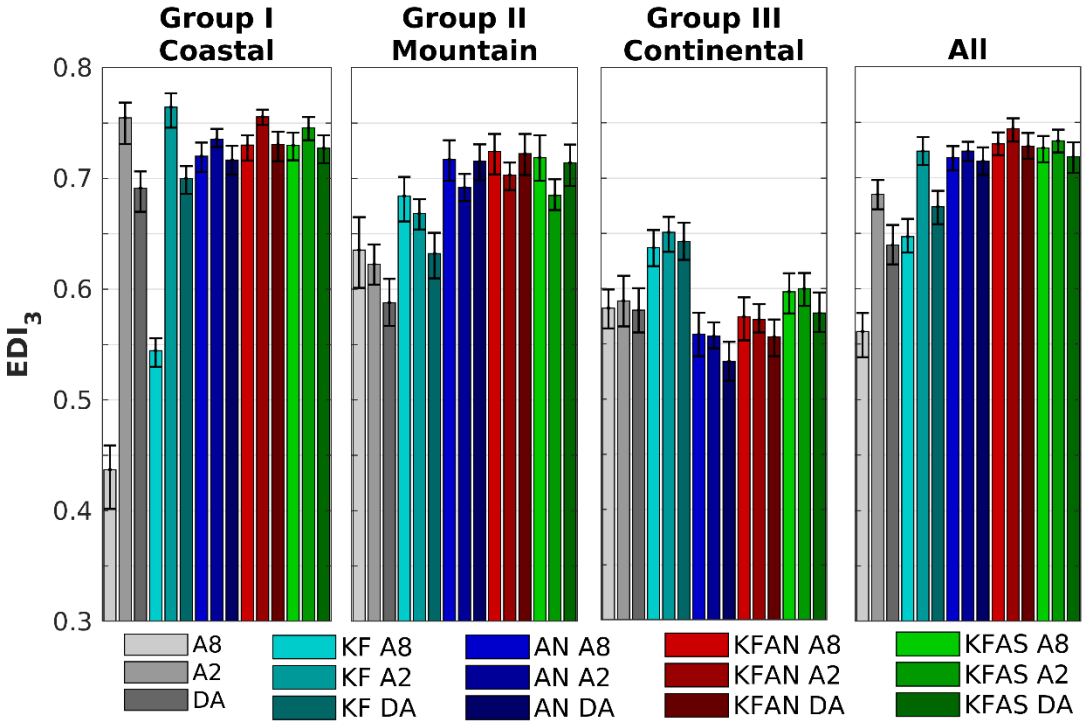

*Figure 15. Extremal dependence index (EDI) for three different starting models (**A8**, **A2**, and **DA**) and the corresponding post-processing methods (**KF**, **AN**, **KFAN**, and **KFAS**). The results are averaged for the corresponding groups and for all of the locations in Croatia during the year 2012. The EDI is calculated for the Category 3 (strong wind; above 90th percentile). The values of the 95% bootstrap confidence intervals are indicated by the error bars.*

Since the critical success index value degenerates as the rarity of the predicted event increases, it is hard to produce a statistically significant result, especially when dealing with only a year-long dataset. For that reason, the measure extremal dependence index EDI_3, which is independent of underlying climatology, is also used to evaluate the forecast of rare events

(i.e. strong wind). The results (Figure 15) are generally consistent with the previously shown critical success index analysis (Figure 14), with smaller confidence intervals. If results among different starting models are compared, it can be seen that for the coastal complex topography the ***A2*** produces significantly higher EDI_3 than the ***A8*** and the ***DA***. This is not the case for other types of topography. The KF approach performs better in a flat continental, while the analog-based method performs better in the mountain complex topography. In the coastal complex topography, the ***KF*** is the best post-processing method for the ***A2*** post-processing, while the analog-based method is more successful than ***KF*** for post-processing ***A8*** and ***DA*** forecasts. Overall, the analog-based method performs better than ***KF***. Among different analog-based experiments, the best result is achieved for the ***KFAN*** forecast. The analog-based method is more successful if it started with ***A2*** than if it started with ***A8*** or ***DA*** models, which is consistent with the previous results.

3.7. Spectral analysis of wind speed forecast

The small spatial and temporal errors of (generally) well-simulated phenomena can profoundly change the verification results [Mass et al. 2002; Rife et al. 2004]. For that reason, spectral analysis in the frequency domain is utilized to provide a scale-dependent measure of different post-processing methods performance which is essentially insensitive to temporal errors. Spectral analysis allows quantification of power distribution among different temporal scales. It is relevant to determine the exposure of a particular station to longer-than-diurnal (LTD), diurnal (DIU) and shorter-than-diurnal (STD) motions and the forecast ability to simulate these motions [Horvath et al., 2012].

Spectral decomposition of the detrended time series is performed using the Welch periodogram-based method [Welch, 1967] with 50% overlapping segments. The data time series is divided into smaller segments. The periodogram is calculated for each segment, and the estimations are then averaged. In other words, by introducing so-called data-, lag- and spectral-window, the variance of the estimator is reduced for longer time series (otherwise it is independent on time series length), making the spectrum smoother. The length of the Hamming spectral window (chosen length is 256; approximately a month-long) is adjusted to optimally emphasize the difference among tested post-processing methods. Here, for a year-

long time series, there are approximately 24 estimations. The distribution of the spectral estimator is often approximated as χ^2 distribution to provide the information on typical variability and confidence intervals [Papoulis, 1984; Koopmans, 1974]. The confidence interval for the power spectra $S(\upsilon)$ is calculated as:

$$\frac{\gamma\overline{S_T}(\nu)}{\chi^2_\gamma(1-\alpha/2)} < S(\upsilon) < \frac{\gamma\overline{S_T}(\nu)}{\chi^2_\gamma(\alpha/2)}, \tag{12}$$

where $\overline{S_T}$ represents the averaged periodograms (estimations) in frequency domain ν, γ represents the number of degrees of freedom (depending on the exact spectral window, overlapping, time series, and interval length), α represents the significance level and the distribution used is χ^2_γ. This interval is usually shown as a small cross sign that is independent on the logarithmic scale. Since not changing the size, it can easily be moved up and down providing a simple visual comparison with the spectrum.

It should be noted that power spectral density (*PSD*) analysis performed contains the effect of aliasing, necessarily contaminating all scales by oscillations with periods shorter than 6 hours (here corresponding to the Nyquist frequency). Testing this effect on measured data suggests that it is rather small on longer-than-diurnal scales. Significant effects, however, may be found on shorter-than-diurnal scales, especially near the periods corresponding to the Nyquist frequency [Žagar et al., 2006; Hrastinski et al., 2015]. Since the ***A8*** and the ***DA*** forecasts are archived every 3 hours (and the ***A2*** and the measurements are adjusted by simply using the same output frequency), it is not possible to circumvent this effect. However, it may be noted that all the forecasts tested (and measurements) are aliased similarly; therefore, the effect is not crucial for the inter-comparison.

The forecast output frequency is 3 h for all forecasts, and only the 24-hour forecast period is considered in the analysis (making a continuous time-series). Missing data are provided by using linear interpolation. It should be noted that the typical diurnal rotation of winds in the Adriatic partially hides the diurnal spectral peak if the analysis is performed using wind speed values [Telišman Prtenjak and Grisogono, 2007]. However, the preferred spectral analysis of wind components is not possible as the analog-based method in our analysis predicts only the wind speed (not the direction).

The spectral analysis is performed for all forecasts and locations included in this part of the thesis (section 3) for the entire year of 2012 (shown in Appendix A). However, it is

decided that it is more comprehensive to show the results for several representative locations, instead of any sort of averaging or summarizing the results. The particularities that could not be easily seen on figures are pointed out and explained in the text. Two locations (Dubrovnik and Jasenice stations) correspond to the coastal group of stations, covering the northern and the southern part of the coastline. The reason for including these two stations is that the governing processes somewhat differ (e.g. processes that lead to bora windstorm as explained in Horvath et al. [2009]). Osijek is chosen as a representative station for the nearly flat continental topography, while Ogulin is chosen to represent the mountain complex topography.

3.7.1. The Kalman filter approach influence

The KF influences the motions on the time scales longer than 10 days if the model's power spectral density *PSD* function is biased. The ***KF*** forecast, therefore, enlarges the energy of these large-scale motions in the coastal area, as shown in Figure 16a. Similarly, KF reduces the energy that is overestimated by the NWP model at the nearly flat continental topography (Figure 16b). Besides the large scale motions, the KF does not significantly influence the shorter time scale. Similarly, the ***KFAN*** is almost the same as the ***AN*** spectrum, except rarely significant differences for large scale motions. The same effect can be noticed, regardless of the starting model (as shown in Odak Plenković et al. [2018]). The very small difference among spectra before and after the application of the KF algorithm might mean that the ratio of the variances (error ratio) used in the algorithm is not optimal. If the error ratio is set too high, the filter puts excessive confidence in the past forecasts, and therefore failing to remove any error. On the other hand, if the ratio is too low, the filter will be unable to respond to changes in bias [Delle Monache, 2006]. The increase of the error ratio might lead to KF algorithm affecting somewhat shorter scales (e.g. synoptic), and possibly even increase the correlation with the observations (as in Delle Monache et al. [2008]). The sensitivity of these results to changing the ratio of the variances used in the algorithm, therefore, might be tested in future work. However, the qualitative effect of affecting only large scale motions would presumably remain the same. Finally, the ***KF*** spectra are the same as the model spectra and the ***KFAN*** spectra are the same as the ***AN*** spectra for the scales

shorter than 10 days and therefore would not be shown or discussed further in this section. However, it needs to be mentioned that these forecasts are not the same, even if their spectra approximately are. Other verification measures shown in previous sections exhibit substantial differences. For example, one can compare a forecast time series to a forecast time series that is exactly the same but time-lagged and bias of the mean is added (e.g. a persistency forecast with a 3-hourly time lag with an added fixed value of 5 ms^{-1}). In comparison, the spectra for these two forecasts will differ only in frequency of 0 Hz. However, the accuracy might differ substantially, leading to very different accuracy measures (i.e. *RMSE* values). This is precisely the reason why the verification procedure needs to include various aspects.

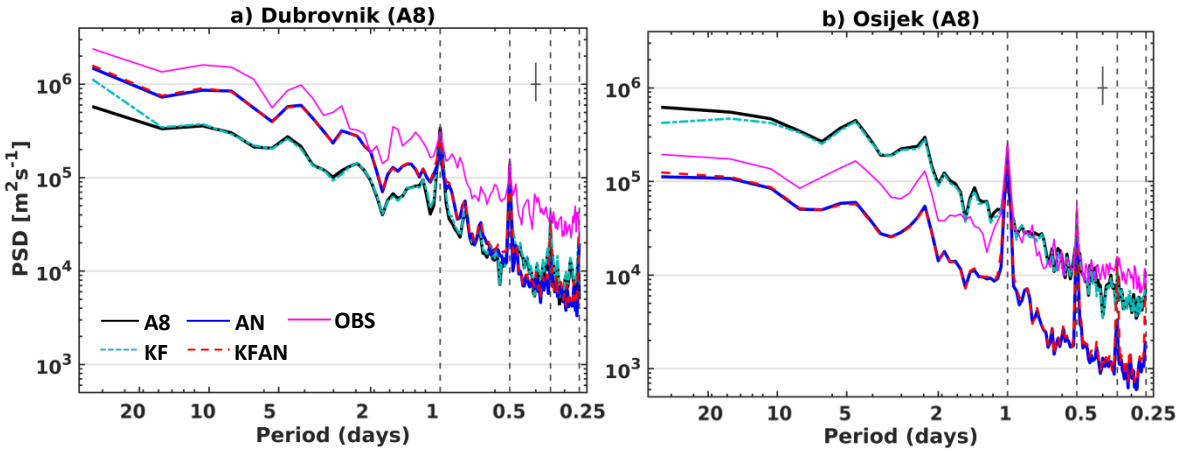

*Figure 16. The power spectral density (PSD) of the observed 10-m wind speed, starting model forecast **A8**, the corresponding **AN**. The effect of the **KF** on the spectra is shown via **KF** (KF applied on the NWP model data) and **KFAN** (KF applied on the AN forecasts). The spectra are shown for coastal Dubrovnik and continental Osijek stations in 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

3.7.2. How the analog-based method affects the A8 NWP spectra

It can easily be seen that the largest portion of measured power at all stations is associated with the longer-than-diurnal motions. These longer-than-diurnal motions are more energetic for the coastal area (Jesenice and Dubrovnik stations – Figure 17a-b) than for the mountain complex (Figure 17c) and the nearly flat continental topography (Figure 17d). As

shown by several other authors, this is related to the strong and gusty bora wind [Horvath et al., 2009; Horvath et al., 2011; Hrastinski et al. 2015, etc.].

The longer-than-diurnal motions are severely underestimated with the ***A8*** model in the coastal area (Figure 17a-b). The longer-than-diurnal motions in the ***AN*** and the ***KFAS*** data contain more energy compared to the model power spectral density *PSD*, therefore, improving the model. This shows great potential for the analog-based predictions to improve the model forecast when there is a model underestimation of longer-than-diurnal motions, even in the complex topography.

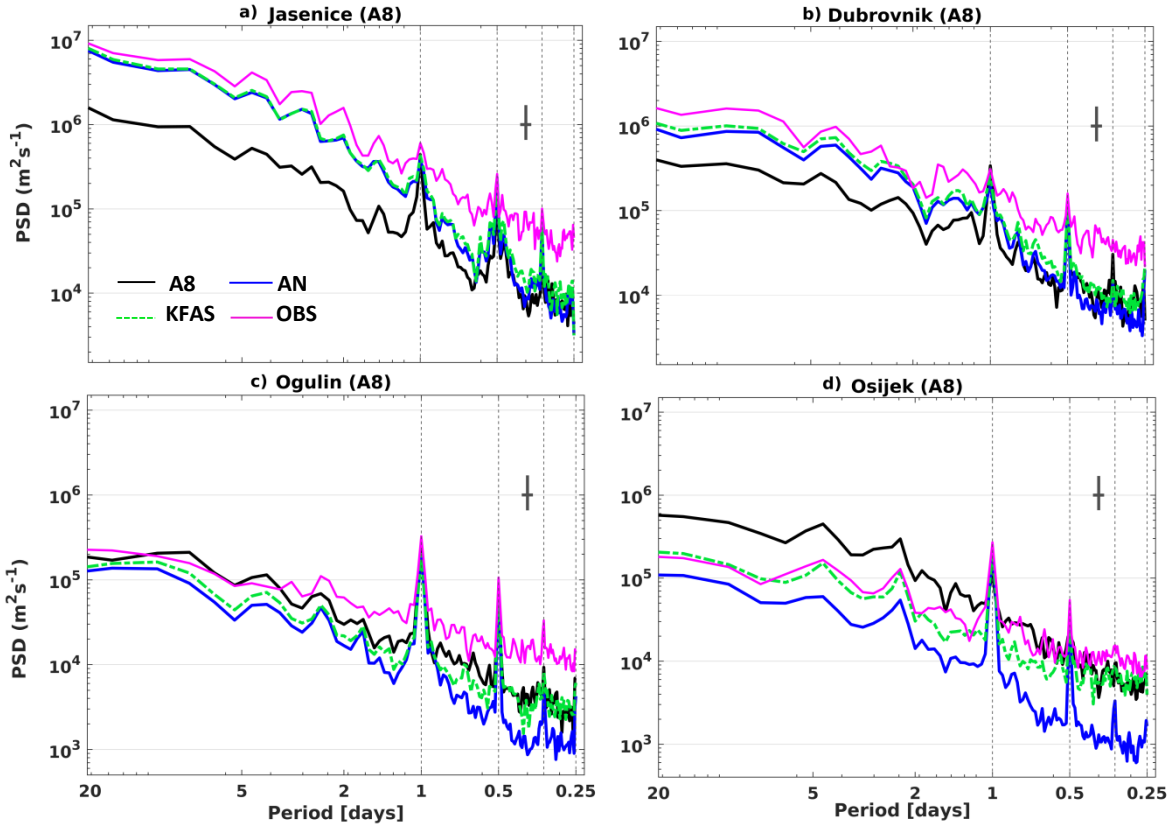

*Figure 17. The power spectral density (PSD) of the observed 10-m wind speed, starting model forecast **A8**, and the corresponding post-processing methods (**AN** and **KFAS**) for stations Jasenice, Dubrovnik, Ogulin and Osijek during the year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

In the nearly flat topography, the ***A8*** model simulates well or overestimates the energy of longer-than-diurnal motions (Figure 17d). The analog-based predictions (***AN***, ***KFAN***, and ***KFAS***) lower the energy of longer-than-diurnal motions if it is well simulated or

overestimated by the model. This sometimes leads to an underestimation of longer-than-diurnal motions, especially for the ***AN***. The ***KFAS*** exhibits the longer-than-diurnal power spectral density spectrum very similar to measurements. Thus, the ***KFAS*** shows the greatest potential for the forecast improvement if the model overestimates the energy of longer-than-diurnal motions in the nearly flat topography.

The ***A8*** results for longer-than-diurnal motions in the mountain complex consist of all previously mentioned scenarios, depending on the location and the exact time scale. For instance, it is well simulated for periods longer than 3 days and underestimated for shorter time scales at Ogulin station (Figure 17c). The analog-based predictions act similarly as in previous types of topography; exhibiting more energy if it is underestimated by the ***A8*** model, or less if it is not.

The shorter-than-diurnal motions are severely underestimated by the ***A8*** model for the majority of locations, regardless of the topography (e.g. Horvath et al. [2011]). Only at a few stations (e.g. Osijek in Figure 17d) is the amount of energy at these scales comparable to measured values. The ***AN*** forecast is, once again, the most prone to energy underestimation. The shorter-than-diurnal ***KFAS*** spectra, on the other hand, seems very similar to model spectra. Moreover, it seems that the ***KFAS*** exhibits power spectral density *PSD* values similar to the ***AN*** and observations for longer time scales, but it is similar to model values at shorter time scales. However, it must be noted again that aliasing of scales shorter than 6 hours adds a considerable share of the energy of shorter-than-diurnal motions in spectral analysis, which is why these results should be interpreted with care. Finally, it is interesting to note that even though the energy of the shorter-than-diurnal motions is underestimated, the harmonics of the diurnal cycle (24 h, 12 h and 8 h period) are very well simulated by the ***A8*** model and all of the post-processing methods.

3.7.3. The influence of the starting model on the analog-based predictions

Introducing the higher-resolution orography affects the dynamical processes and increases the amount of energy at all temporal scales (e.g. Žagar et al., 2006). Therefore, the difference between the ***A8*** and the ***DA*** is that there is almost no underestimation of the longer-than-diurnal motions, even in the coastal complex area (e.g. Figure 18b). The exception is the

Dubrovnik station (Figure 18d), which is very similar as it is for the ***A8*** model (Figure 17b). The energy simulated by the ***DA*** is higher at the mountain complex station (Figure 18f) than when simulated with the ***A8*** (Figure 17c), overestimating the longer-than-diurnal motions. Introducing the higher resolution orography into nearly flat continental topography results with very similar power spectral density curves for the ***DA***, as it is the case for the ***A8*** (e.g. Figure 18h, compared to Figure 17d). This is to be expected because the flatter the topography, the number of details added by increasing horizontal resolution is smaller. In the mountain complex topography (group II) results may be improved by using an even finer model resolution to represent local flows. However, the need for using 2- opposed to 8-km grid spacing for weak wind in the nearly flat continental topography (group III) may be less pronounced. Naturally, the post-processing methods are also exhibiting similar effect as it is the case of the ***A8*** model. Similarly, introducing a higher resolution field into the ***A2*** forecasts increases the power at all time scales. All the conclusions regarding power spectral density spectra that are valid for the ***DA*** longer-than-diurnal motions are valid for the ***A2*** model as well.

Additionally, due to the more complete package of physics parametrizations and non-hydrostatic effects, the ***A2*** model shorter-than-diurnal part of power spectral density spectra contains more energy than for the ***A8*** and the ***DA*** models, partially due to aliasing effect. Both the ***A8*** and the ***DA*** models severely underestimate the power at scales below diurnal, as reported by Žagar et al. [2006]. Unlike the ***A8*** and the ***DA*** models, the ***A2*** simulates well or even sometimes overestimates the shorter-than-diurnal motions. The exception is Dubrovnik station, where some underestimation of the shorter-than-diurnal motions can still be noticed.

Even when the model overestimates the shorter-than-diurnal motions, the analog-based predictions reduce the shorter-than-diurnal power, often leading to under-prediction of shorter-than-diurnal motions. When the shorter-than-diurnal motions are well simulated or underestimated by the model, the ***AN*** forecast often severely lacks power for these shorter-than-diurnal motions. The ***KFAS*** forecast, however, exhibits power spectral density values similar to the ***AN*** and observations for longer time scales, but it is similar to model values at shorter time scales. In other words, the ***KFAS*** forecast is less prone to underestimation of the shorter-than-diurnal motions than other analog-based predictions tested. This result is consistent regardless of the starting model.

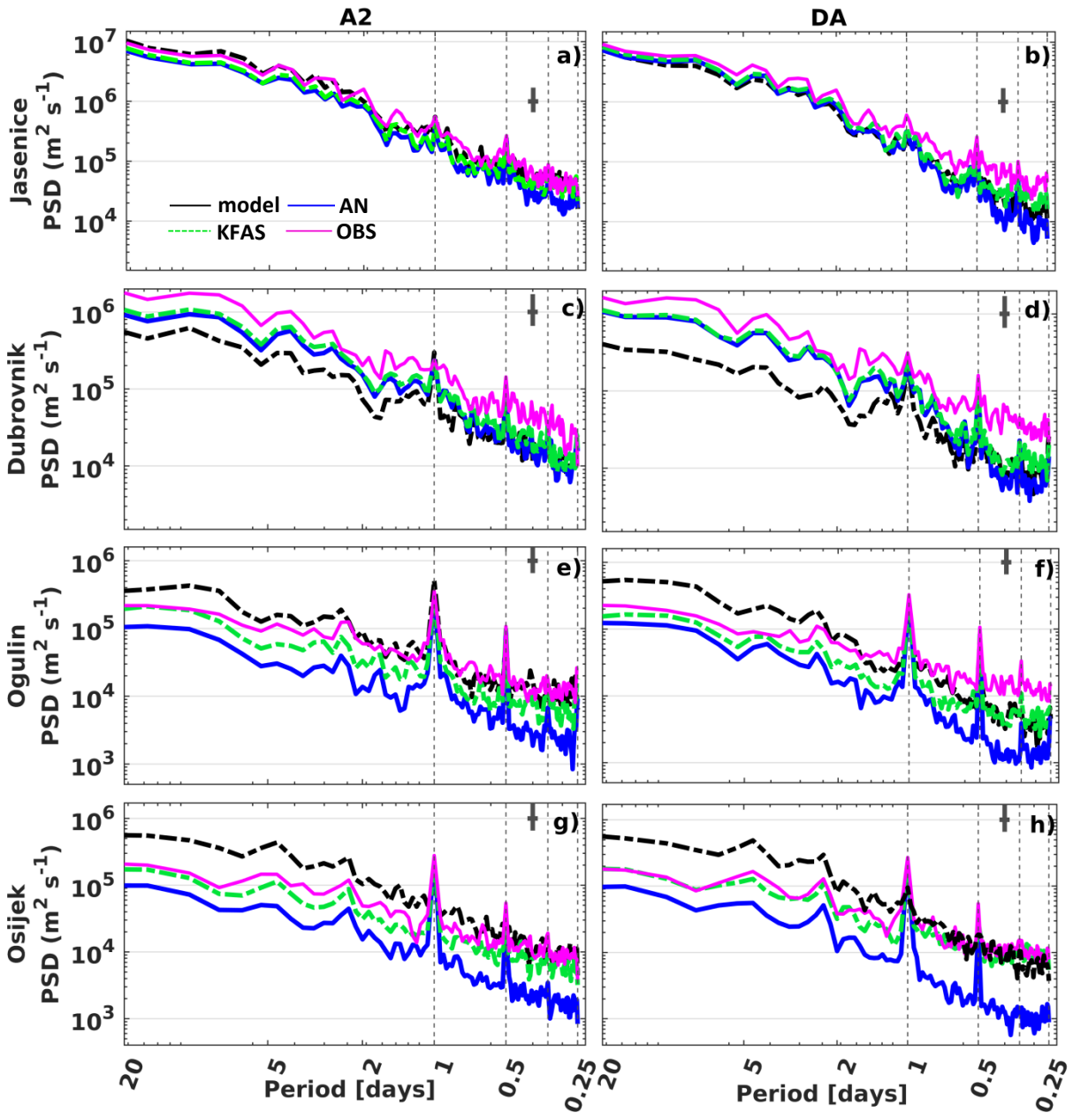

*Figure 18. The power spectral density of the observed 10-m wind speed, starting model forecasts (**A2** and **DA**) and the corresponding post-processing methods (**AN** and **KFAS**) for stations Jasenice, Dubrovnik, Ogulin and Osijek during year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

§ 4. POST-PROCESSING THE ENSEMBLE NWP (ENSEMBLE CALIBRATION)

4.1. Observations and climatology

The Austrian meteorological observation network, TAWES, consists of more than 300 sites across Austria. In this work, 29 TAWES sites are used representing the different Austrian climate zones, as listed in Table 4. The locations are selected based on the availability of wind speed measurements (10-minute average value) at 10 m above the ground in the selected time period. All sites monitor 2-m temperature, 10-m wind speed and direction, 2-m relative humidity, surface pressure, precipitation, and, depending on the site, different radiation measurements are carried out. Here, only 10-meter wind speed observations are used. The 2015 and 2016 wind speed observations are used for the analog-based method training period in this section. For the performance testing, two target months are chosen, January and July 2018. These months are selected to investigate the forecast performance during a winter and a summer period. The January and July 2017 wind speed observations are used for independent sensitivity testing (weight optimization), which is a procedure explained further below.

The observed average monthly wind speed is slightly higher in January (2.88 ms^{-1}) than in July (2.22 ms^{-1}), across all available stations and lead-times. Additionally, the standard deviation of the wind speed measurements is also higher on average in January (3.27 ms^{-1}) than in July (1.92 ms^{-1}).

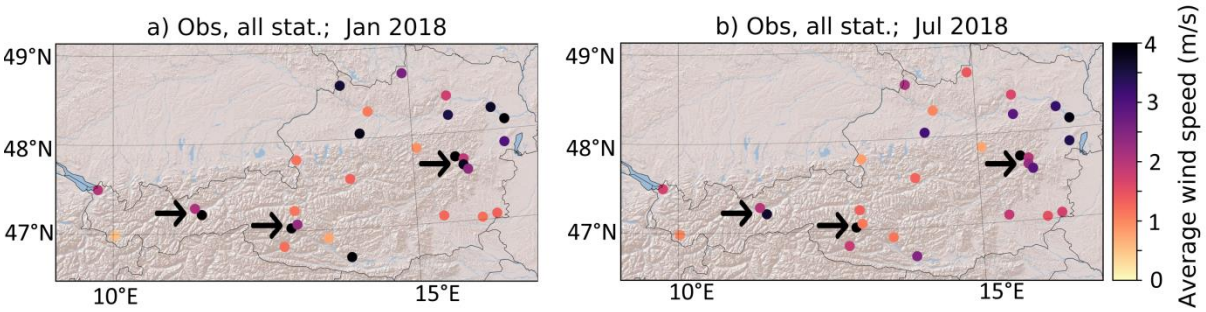

Figure 19. The spatial distribution of the observed monthly mean wind speed in the January (left) and July (right), 2018. The arrows mark mountain stations for later comparison.

Table 4. The list of the 29 stations providing the 10-m wind speed observations used in section 4 with main geographical features.

Station name	Latitude	Longitude	Altitude [m]
Weitra	14.9	48.7	572
Wien-Hohe Warte	16.4	48.3	198
Schwechat	16.6	48.1	183
Linz-Stadt	14.3	48.3	262
Krems	15.6	48.4	203
Bregenz	9.7	47.5	424
Gaschurn	10.0	47.0	976
Patscherkofel	11.5	47.2	2251
Lunz Am See	15.1	47.9	612
Rax/Seilbahnbergstation	15.8	47.7	1547
Eisenstadt-Nordost	16.5	47.9	184
Güssing	16.3	47.1	215
Lienz	12.8	46.8	661
Kanzelhöhe	13.9	46.7	1520
Fürstenfeld	16.1	47.0	271
Gmünd	13.5	46.9	738
Graz-Univ.	15.4	47.1	367
Innsbruck-Univ.	11.4	47.3	578
Sonnblick	13.0	47.0	3109
Kolm Saigurn	13.0	47.0	1626
Rauris	13.0	47.2	934
Salzburg/Freisaal	13.0	47.8	418
Bad Mitterndorf	13.9	47.6	814
Reichenau/Rax	15.8	47.7	488
Semmering	15.8	47.6	988
Hirschenkogel	15.8	47.6	1318
St. Pölten/Landhaus	15.6	48.2	274

The wind speed is weak and moderate (i.e. < 8.0 ms^{-1}) for both January (Figure 19a) and July (Figure 19b) at the majority of the stations. The average monthly wind speed increases towards the north-eastern part of Austria (Pannonian basin) for both January and July. Exceptions are the three mountain stations (arrows in Figure 19), where the average wind speeds are much higher if compared to the neighboring valley stations.

Most of the stations are located in or near the Alps, which significantly modulates the related local wind regimes. The complex topography of the Alpine area is characterized by a

variety of different wind processes such as foehn and downslope windstorms, gap winds, valley and slope winds, flow blocking and other. To investigate those phenomena, among other, Alpine region is also the target area to several major field experiments, such as ALPEX, MAP and TEAMx [Kuettner, 1986; Bougeault et al., 2001; Lehner and Rotach, 2018; Serafin et al, 2018, etc.]. Nevertheless, many challenges related to the NWP in complex topography still exist (e.g. Arnold et al. [2012]), including modeling wind climatology of the Alpine areas prone to such downslope windstorms (e.g. Horvath et al. [2011]) and objective foehn wind classification (e.g. Mayr et al. [2018]).

4.2. NWP model data

The numerical model used within section 4 is the ALADIN model configuration used in ALADIN-LAEF (Aire Limitée Adaptation dynamique Développement InterNational model – Limited-Area Ensemble Forecasting) [Wang et al., 2011, 2019] ensemble forecasting system. It is adjusted to fit the Austrian purposes and is running in operational mode since 2009. The NWP is initialized daily at 0000 and 1200 UTC with one hourly lead-time, up to 72 hours. Only the dataset corresponding to the model run initialized at 0000 UTC is used in this work.

The ALADIN-LAEF uses the underlying hydrostatic and spectral limited-area model (LAM) ALADIN-Austria [Wang et al., 2006]. It uses a two-time-level semi-Lagrangian advection scheme, semi-implicit time-stepping, fourth-order linear horizontal diffusion, Davies–Kalberg type relaxation and digital filter initialization, and set of parametetrizations of unresolved physics processes [Wang et al., 2006].

The ALADIN-LAEF integration domain covers the whole of Europe and a large part of the Atlantic, as shown in Figure 20. The resolution of 11 km on a Lambert conformal grid is used in the horizontal. In the vertical, 45 terrain-following pressure-based hybrid coordinate levels with on average nine levels within the lowest 1000 km above ground level are used.

For dealing with the initial uncertainties, a blending method is used [Wang et al., 2014], based on the idea of combining the large-scale perturbation from the ECMWF (European Centre for Medium-Range Weather Forecasts) singular vectors and the small-scale perturbations from the LAM native breeding vectors. The coupling with ECMWF-EPS

(Ensemble Prediction System) members are used for dealing with the lateral boundary condition uncertainties [Weidle et al., 2013]. A multi-physics is implemented to account for model uncertainties in the atmosphere. The perturbed initial land surface conditions, such as soil moisture and surface temperature, are obtained through an ensemble of land surface data assimilation [Belluš et al., 2016].

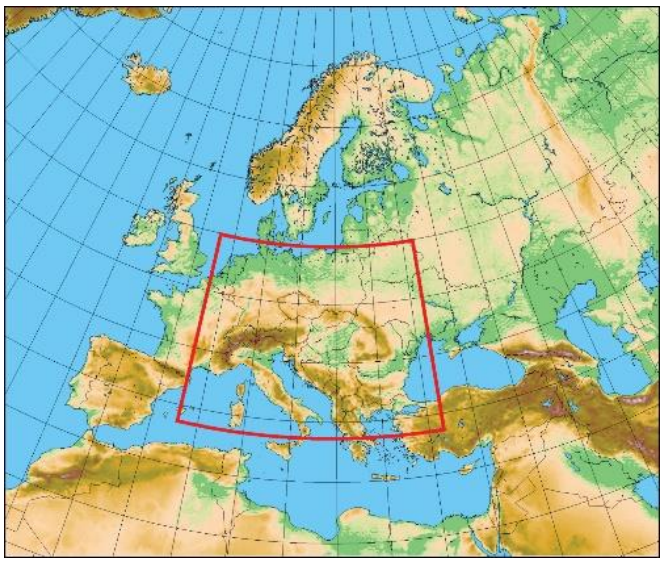

Figure 20. Domain and model topography of ALADIN-LAEF. (from Wang et al. [2019], page 3355. The inner limited-area domain in red represents the area authors used for verification of ensemble experiments).

The ALADIN-LAEF consists of 17 ensemble members: 16 perturbed members and one control run. The 16 perturbed members are driven by 16 ECMWF-EPS members. Given the structure and composition of the LAEF ensemble, it can be considered as a non-exchangeable ensemble. However, as shown by Baran and Lerch [2015] the differences between the treatment of a non-exchangeable ensemble as fully exchangeable did not worsen the results to statistically relevant size. Therefore, we decided to treat the ALADIN-LAEF ensemble as exchangeable.

A subset of six ALADIN-LAEF parameters to be used as an input to the analog-based method includes temperature (*t2m*), wind speed (*ws*) and direction (*dd*), relative humidity (*rH*), pressure (*p*) and precipitation (*prec*). The NWP datasets correspond to the observation datasets. From the four grid points surrounding the observation location, the closest model grid point is chosen. The 2-year long dataset (2015 – 2016) is used for training. January and

July 2017 are used for weight optimization. Finally, the results are given for the independent dataset consisting of January and July 2018.

4.3. Reference method: Ensemble model output statistics (EMOS)

The reference forecast for the analog-based method ensemble calibration is the ensemble model output statistics (EMOS). The EMOS is introduced by Gneiting et al. [2005] and adapted for wind by Messner et al. [2014]. Therefore, a non-homogeneous regression with a 30-day rolling training window is fitted on every lead-time and station. To capture the natural boundary of wind at 0 ms^{-1}, a left-censored logistic regression is used. In the EMOS the observed wind speed (y) is explained by a logistic distribution censored at zero ($\mathcal{L}_0$) with μ as a mean and σ as a spread. A logistic distribution has a similar bell shape as a Gaussian distribution but with slightly heavier tails. Additionally, censoring at zero states that no negative wind values can occur. Further details can be found in Messner et al. [2014]. Censoring and the linear regressions for μ and σ are defined as follows:

$$y \sim \mathcal{L}_0(\mu, \sigma), \tag{13}$$

$$\mu = \beta_0 + \beta_1 ws_\mu \tag{14}$$

$$\log(\sigma) = \gamma_0 + \gamma_1 \log(ws_\sigma), \tag{15}$$

with β_* and γ_* as the regression coefficients, ws_μ as an ensemble mean and ws_σ as an ensemble spread of the wind speed members. The logarithmic link function is used to ensure positive values. Further applications of the EMOS to wind speed can be found in Thorarinsdottir and Gneiting [2010], Baran and Lerch [2015] or Scheuerer and Möller [2015].

The 30 days rolling training window is often used for the ***EMOSws*** experiment, making it a good reference for the analog experiment that uses only the raw model wind speed data. However, since the other analog experiments use all available variables, a second reference is added. The second experiment (***EMOSstd***) uses all available variables. The boosting method of Messner et al. (2017), which is implemented in the R-package "*crch*", is applied to all variables and the whole dataset, instead of the rolling training window. Additionally, annual and biannual harmonic functions are added to capture a seasonal bias. A variable selection

method, such as boosting, is needed to prevent overfitting. The boosting is able to choose the most important variables and exclude the other variables using zero value. As a result, a single fit per station and lead time can be used to forecast both test months.

Concluding, whereas the ***EMOSws*** only uses the last 30 days as training and only the wind speed as an input, the ***EMOSstd*** uses all available training data and all variables including seasonal functions.

4.4. Sensitivity tests

In the previously described experiments (section 3) the predictors are chosen using the „trial-and-error“ approach, simply trying several combinations of available predictor variables and keeping the one that seems to be the most successful. Following the work of Delle Monache et al. [2013], all predictor weights are set to value 1.00. However, several authors in more recent work show that, instead of assigning the same importance to each predictor variable, the brute-force weight optimization can increase the AnEn performance. This is demonstrated in several applications, such as Junk et al. [2015] and Alessandrini et al. [2015a]. The weights' optimization is based on choosing the combination that minimizes the error (measured by the continuous rank probability score). For that reason, it is decided to include a predictor weighting strategy in the second part of this thesis.

Even though it is the best possible approach, due to the limited computational resources, not all the possible combinations are tested in this work. The forward selection algorithm is used instead, starting with weight value fixed at 1 for the wind speed parameter. Then, one by one (ensemble mean) predictor from a pre-selected subset of six ALADIN-LAEF parameters is added, optimizing the weights independently at each location by error minimization. The forward selection algorithm is computationally less demanding than testing all the possible combinations independently at each location. However, it needs to be noted that the algorithm makes a key assumption that is often not true - assuming that all predictors are independent of each other, which is generally not the case.

As already mentioned, six ALADIN-LAEF parameters are used as an input to the analog-based method: wind speed (*ws*) and direction (*dd*), temperature (*t2m*), relative humidity (*rH*), pressure (*p*) and precipitation (*prec*). They are tested using the forward selection algorithm

one after another, in the same order as listed. Five possible weight values (0.00, 0.25, 0.50, 0.75 and 1.00) are investigated for each predictor variable. The predictor weighting strategy is carried out for January and July 2017, using the 2015-2016 period for the training. Therefore, the optimization procedure uses a completely independent dataset from the period for which training, as well as for which forecasting is performed (January and July 2018). The independency of the datasets used is an important aspect that ensures the objective validity of the results.

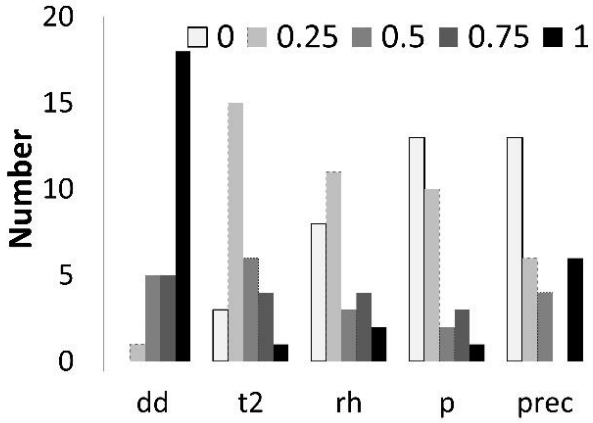

Figure 21. The histogram of the optimized weights for each predictor tested (using the AnEn mean values), at 29 stations in Austria in January and July 2017.

The results show that the wind direction is the most important predictor in addition to wind speed (Figure 21). Even though it seems the values are slightly higher in the complex topography, the values are quite high for all stations (Figure 22). The wind direction is followed by temperature and relative humidity parameters, especially in the more complex topography, such as the alpine area. The pressure and precipitation parameters are often optimized with the 0.00 weight, meaning that they are not carrying additional benefits at certain stations. But, that is not always the case. For instance, the pressure parameter is also optimized by taking higher values in the complex alpine area. For precipitation parameter, a similar behavior is found at the southern slopes of the Alps, a region prone to the convective precipitation. The increased importance of the precipitation predictor in this area might, for example, indicate the forecast improvement under foehn conditions, when foehn triggers the precipitation while approaching the southern Alps.

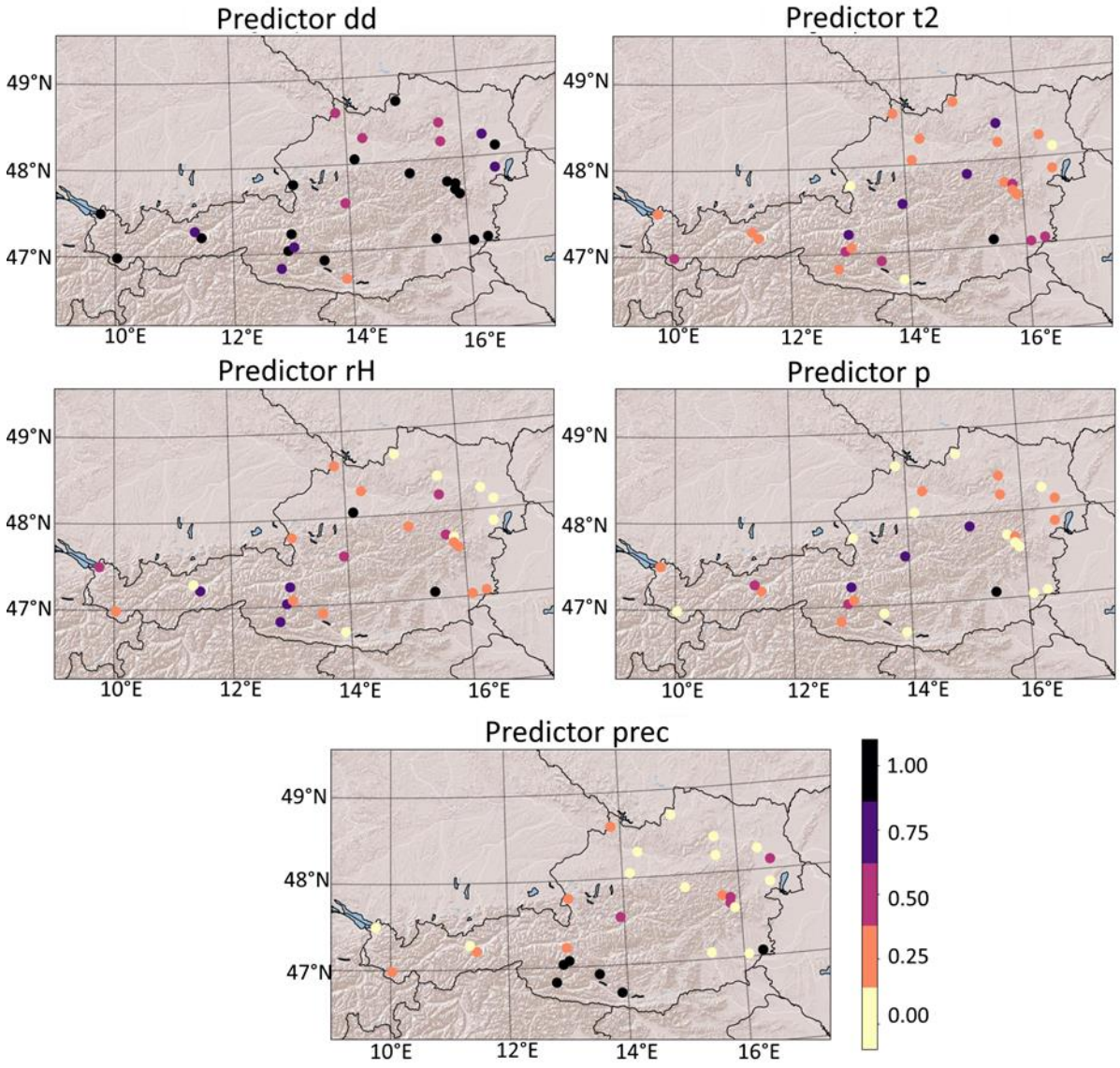

Figure 22. The spatial distribution of the optimized weights for each predictor tested, at 29 stations in Austria in January and July 2017.

Supplementary to using the mean value of 17 ALADIN-LAEF ensemble members for each meteorological parameter, the standard deviation of those 17 members can also be used as an additional predictor. Thus, the information on the starting model ensemble uncertainty is included in the analog search. The standard deviation predictors are optimized as one multiplying factor to the all pre-calculated weights for meteorological parameters, independently for each location. Five possible values of this multiplying factor are tested: 0.20, 0.40, 0.60, 0.80, and 1.00. If using neither of the values results in a forecast improvement, the value 0.00 is used as the best fit. In the following illustrative example, it is

assumed that the optimal weight for the ALADIN-LAEF temperature ensemble mean predictor is 0.75 at a particular location. Similarly, the weight for the relative humidity is optimized as 0.50, for precipitation as 0.00, etc. Then, the weight for the six ALADIN-LAEF ensemble standard deviation predictors is optimized as 0.20 value. The w_i in Eq.1. would be 0.20×0.75 for the temperature standard deviation predictor, 0.20×0.50 for the relative humidity standard deviation predictor, 0.20×0.00 for the precipitation, etc. The distribution for the optimized standard deviation multiplying factors is given at the (Figure 23). The result shows that the optimal contribution of the standard deviation predictors is about 40% of the ensemble mean predictors' contribution in the majority locations tested. However, no distinctive spatial distribution pattern regarding the optimal values is noticed.

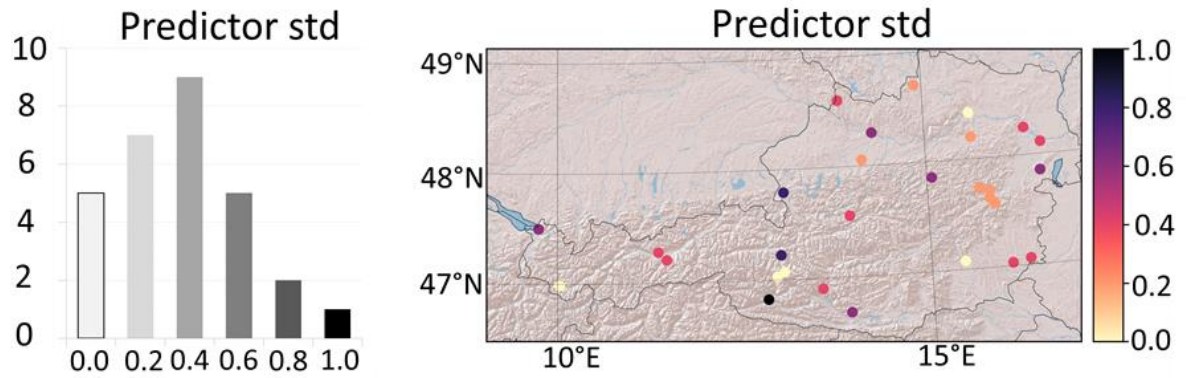

Figure 23. The histogram (left) and the spatial distribution (right) of the optimized weights for standard deviation predictor for different meteorological parameters tested at 29 stations in Austria in January and July 2018.

The AnEn can be affected by a conditional negative bias, especially when predicting events in the right tail of the forecast distribution. For that reason, the novel bias correction method is applied for these experiments, as proposed by Alessandrini et al. [2019]. The method is based on correction factor proportional to the linear regression coefficient between the wind speed observations and raw model forecast (i.e. ALADIN-LAEF wind speed ensemble mean) during training, as well as to the distance between the current raw model forecast and the average value of the previous raw model forecasts that correspond to the currently selected analogs in the AnEn. The lead-time-independent correction factor is added to all the members of the AnEn if the current raw model forecast is above a certain threshold value. If the threshold is set too low, the bias correction adjustment can become small and

noisy, leading to forecast performance degradation. After the simple minimizing the *RMSE*, the 95. percentile of the climatological raw model forecast distribution (during training period) is chosen as a threshold in this work.

4.5. Description of experiments

In total, six different input configurations using the observations and the ALADIN-LAEF ensemble data are investigated (see Table 5 for a summary). All six investigated configurations provide an AnEn forecast, consisting of the past observation corresponding to the 17 most similar past ALADIN-LAEF ensemble predictions. Thus, the new analog ensemble forecast provides the 17 ensemble members, equivalent to the original ALADIN-LAEF model. The chosen ensemble size does not only reflect the input NWP ensemble but is close to the optimal size of 15 members for the deterministic application of the analog ensemble found by Odak Plenković et al. [2018].

Table 5. The summary information for the experiments tested in section 4.

Name	Meteorological variables used	ALADIN-LAEF input (predictors)	Nb. of analog searches per lead-time
LAEFws	*ws*	X	X
EMOSws	*ws*	Ensemble μ and σ for one parameter, wind speed (2 predictors)	X
EMOSstd	*ws, dd, t2m, rH, p, prec*	Ensemble μ and σ for six parameters (12 predictors)	X
AnEnCtrl	*ws, dd, t2m, rH, p, prec*	Control ensemble member for six parameters (6 predictors)	1
AnEnWs	*ws*	17 ensemble wind speed members (17 predictors)	1
AnEnMu	*ws, dd, t2m, rH, p, prec*	Ensemble μ for six parameters (6 predictors)	1
AnEnStd	*ws, dd, t2m, rH, p, prec*	Ensemble μ and σ for 6 parameters (12 predictors)	1
AnEnAll	*ws, dd, t2m, rH, p, prec*	17 ensemble members for 6 parameters (6 × 17 predictors)	1
AnEnMem	*ws, dd, t2m, rH, p, prec*	1 ensemble member for every parameter (6 predictors)	17

Dabernig et al. [2015] show the value of an ensemble forecast compared to its deterministic control run. Therefore, the first experiment, the ***AnEnCtrl***, uses the ALADIN-LAEF control member for the six meteorological parameters available as six predictors. The ***AnEnWs***, uses all 17 ALADIN-LAEF ensemble member wind speed predictions (***LAEFws***) as 17 predictors. More meteorological variables are exploited in the ***AnEnMu*** experiment. In contrast to the ***AnEnWs***, only the ensemble mean μ for every parameter is used as a predictor in the ***AnEnMu*** experiment. For the ***AnEnStd*** ensemble forecasts, the ALADIN-LAEF ensemble uncertainty (σ) and the ensemble mean (μ) of the defined six meteorological parameters are used. The ***AnEnStd*** includes the aspects of error growth, represented dynamically by the used ensemble model, as explained in Eckel and Delle Monache [2016]. This adds additional information to the flow-dependent error growth already captured by the analog approach (e.g. in ***AnEnMu***).

In addition to the aforementioned experiments, two diverging ways of including all the ALADIN-LAEF information available are investigated. The first additional experiment, the ***AnEnAll***, uses every member of the ALADIN-LAEF ensemble for every defined meteorological predictor. Thus, in this experiment, 6 variables and 17 ensemble members are used, which equals 6 × 17 predictors. An important goal of this research is to evaluate if all probabilistic information is needed or summary measures, such as mean or spread, are already sufficient. The second additional experiment is the "member by member" approach ***AnEnMem***. Here, the analog search procedure is carried out for every ALADIN-LAEF member separately. Therefore, each raw model member is now distinguishable from the others. The analog-search procedure is independently done for each set of six pre-defined meteorological parameters, corresponding to the same raw model member. Thus, in ***AnEnMem*** the search procedure is performed 17 times in total. Only one analog is chosen in every analog search procedure per ensemble member, with verifying observation chosen as the member in the ***AnEnMem*** ensemble. This is the most demanding configuration presented in this research. An analog experiment similar to the ***AnEnMem*** experiment, but using more than one analog (e.g. 5 analogs) for each of the ALADIN-LAEF ensemble members, is also investigated. However, besides being even more computationally demanding, it did not provide any benefits justifying the additional computational costs. Therefore, these results are not discussed here.

All experiments use an analog search time window fixed at every lead-time individually, including one time step before/after to account for a trend.

To determine if the difference in scores between the experiments is statistically significant, the moving-block bootstrap technique, following the procedure of Wilks [1997] and using 1000 re-samples at a confidence level of 95%, is applied, except for correlation where pair bootstrap technique was used (as in Wilcox [2009]; see section 4.2).

4.6. Evaluation of the wind speed ensemble and probabilistic forecast

Even when evaluating the ensemble forecast, a useful starting point is to define a dominant source of error. The source can be specified when decomposing the *RMSE* to the bias of the ensemble mean, the bias of the standard deviation (σ bias) of the ensemble mean and the dispersion (phase) error of the ensemble mean, as previously explained. It needs to be noticed that the σ bias is defined as the bias of the standard deviation of the ensemble mean (regardless of the ensemble spread).

A particularly important aspect of ensemble forecasting is the information about the uncertainty in a forecast. The standard deviation of the ensemble members with respect to its mean is referred to as the spread of the ensemble. The spread describes the diversity of the ensemble forecast. In other words, the forecaster is confident that the ensemble mean is close to the eventual state of the atmosphere if the spread of the ensemble is small. On the contrary, if the ensemble members are all very different from each other, the future state of the atmosphere is more uncertain. To adequately represent the forecast uncertainty, the magnitude of ensemble spread should correspond to the magnitude of the error in the ensemble mean. A large difference between the ensemble spread and the *RMSE* of the ensemble mean is an indication of statistical inconsistency, while closeness is a measure of the statistical reliability [Buizza et al. 2005]. A good match between the ensemble spread and *RMSE* of the ensemble mean implies the greater predictability of ensemble mean skill, suggesting that the ensemble spread represents the ensemble uncertainty well.

The ensemble is consistent if the actual future atmospheric state behaves like a random draw from the same distribution that produced the ensemble [Anderson, 1997]. Then, the observation being predicted looks statistically like just another member of the forecast

ensemble. The probability forecasts derived from an ensemble is good (i.e., appropriately expresses the forecast uncertainty) to the extent that the consistency condition has been met. A necessary condition for the ensemble consistency is an appropriate degree of ensemble dispersion. The most common approach to evaluating whether a collection of ensemble forecasts satisfies the consistency condition is the construction of a verification rank histogram (or simply the rank histogram). The rank of the corresponding observation within the ensemble is tabulated, taking the value from 1 to $n+1$ for an n-member ensemble. Collectively, these verification ranks are plotted in the form of a histogram. If the consistency condition is met, the histogram of verification ranks is uniform, reflecting the equiprobability of the observations within their ensemble distributions [Wilks, 2011]. The exceptions are departures that are small enough to be attributable to sampling variations. Departures from the ideal of rank uniformity can be used to diagnose aggregate deficiencies of the ensembles [Hamill, 2001], as shown in Figure 24.

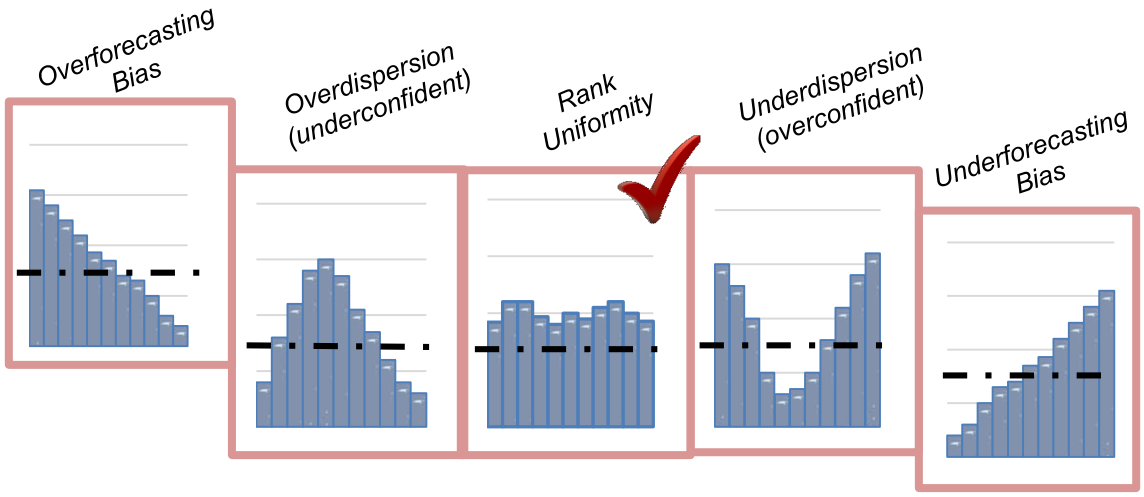

Figure 24. Example verification rank histograms for hypothetical ensembles of size 10, illustrating characteristic ensemble dispersion and bias errors. Perfect rank uniformity is indicated by the horizontal dashed lines, and the best match is noted with a "check" mark.

The Brier Skill Score (BSS) is a commonly used metric for the probabilistic forecast of binary event that uses climatology as a reference [Wilks, 2011; Jolliffe and Stephenson, 2011]. It is calculated using the following expression:

$$BSS = 1 - \frac{BS}{BS_{clim}}, \tag{16}$$

where the Brier score ($BS = \sum_i (f_i - o_i)^2 / n$) averages the squared differences between pairs of forecast probabilities f and the subsequent binary observations o over all n forecast – observation pairs. The Brier score is essentially the mean squared error of the probability forecasts, where the observation value is 1 (if the event occurs) or 0 (if the event does not occur). A binary event is defined using an exceedance threshold, i.e. of wind speed forecasted higher than 5 ms^{-1}. The closer the *BSS* is to the perfect number 1, the better the skill of the forecast is. Here, a threshold of 5 ms^{-1} is chosen for the *BSS* as it is reasonably high while, on the other hand, not being too rare. The Brier score is negatively oriented, with perfect forecasts having value 0. Since individual forecasts and observations are both bounded by 0 and 1, the score can take on values in the range between 0 and 1.

After some algebra, the Brier score can be expressed as the sum of the three terms: reliability (*REL*), resolution (*RES*), and uncertainty (*UNC*), as follows:

$$BS = RES - REL + UNC; \quad BSS = \frac{RES - REL}{UNC}, \tag{17}$$

The preferred outcome is as small as possible reliability term and as large as possible resolution term (in absolute value). The reliability term describes the calibration (or conditional bias) of the forecasts. The forecast probability in each subsample of the perfectly reliable forecast is exactly equal to the relative frequencies of the observed event in each subsample. The resolution term describes the ability of the forecasts to distinguish subsample forecast periods with relative frequencies of the event that are different from each other. In other words, the resolution term will be large if the forecasts sort the observations into subsamples having substantially different relative frequencies than the overall sample climatology (or vice versa). Since the uncertainty term depends only on the sample climatological relative frequency, it is unaffected by the forecasts. This term takes on value 0 when the climatological probability is either 0 or 1. Similarly, when the event being forecast almost never (or almost always) happens, the uncertainty in the forecasting situation is small. Then, forecasting the climatological probability gives good results. Contrary, the uncertainty maximum is achieved when the climatological probability is 0.5. In that case, there is substantially more uncertainty inherent in the forecasting situation.

The reliability diagram is a graphical device that shows the full joint distribution of forecasts and observations for probability forecasts of a binary event in terms of so-called

calibration-refinement factorization. The elements of the calibration-refinement factorization are the calibration (or conditional) distribution of the observation given each of the n allowable values of the forecast $p(o_1|f_i)$; and the refinement distribution $p(f_i)$ that describes the frequency of use of each of the possible forecasts. Here, the occurrence of a binary event is noted with index 1. The n calibration probabilities $p(o_1|f_i)$ define a calibration function that is usually the main aspect of the reliability diagram [Wilks, 2011]. The reliability diagram provides an insight into the unconditional and conditional biases, as shown in Figure 25.

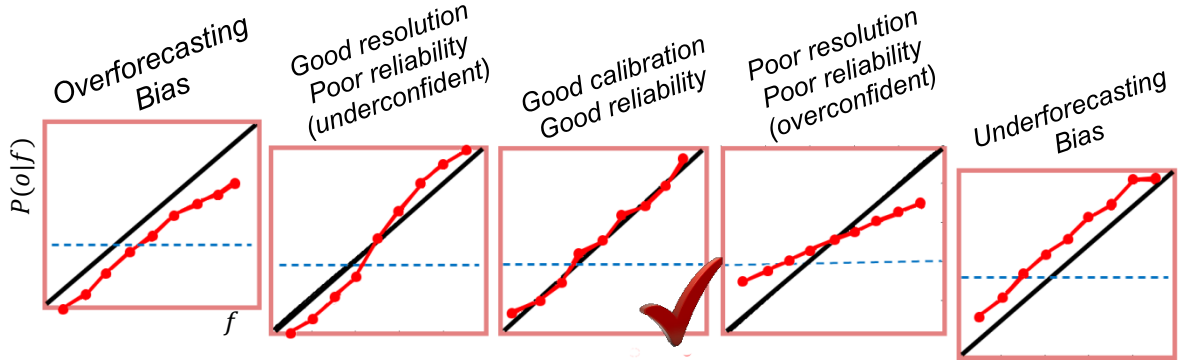

Figure 25. Example characteristic forms for the calibration-function $P(o|f)$ element of the reliability diagram. The black diagonal line represents the perfect reliability, while the blue dashed line represents the climatological frequency of the event. The most reliable forecast is indicated by a "check" mark. The arrangement of the panels corresponds to the calibration portions of the rank histogram in Figure 24.

The labels "underconfident" and "overconfident" are concerning the other elements of the reliability diagram: the refinement distribution $P(f_i)$ shown in the so-called sharpness diagram. The dispersion of the refinement distribution reflects the overall confidence of the forecaster. For example, forecasts that deviate rarely and quantitatively little from their average value exhibit little confidence. Forecasts that are frequently extreme (i.e. often 0% or 100% chance for the event occurrence) exhibit high confidence/sharpness [Wilks, 2011]. Characteristic forms are shown in Figure 26.

The continuous rank probability score *CRPS* is a summary metric that can be interpreted as the integral of the Brier score over all possible threshold values for the parameter under consideration:

$$CRPS = \int_{-\infty}^{\infty} [P_F(x) - P_o(x)]^2 \, dx, \tag{18}$$

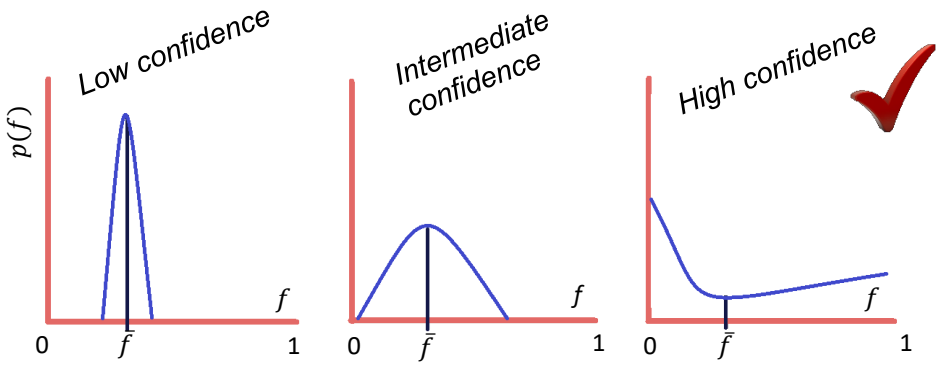

Figure 26. Example characteristic forms for the sharpness (refinement distribution) $p(f)$ element of the reliability diagram. Forecasts that are frequently extreme (i.e. often 0% or 100% chance for the event occurrence) exhibit high confidence/sharpness. On the other hand, narrow distribution close to the average forecast value exhibit low confidence.

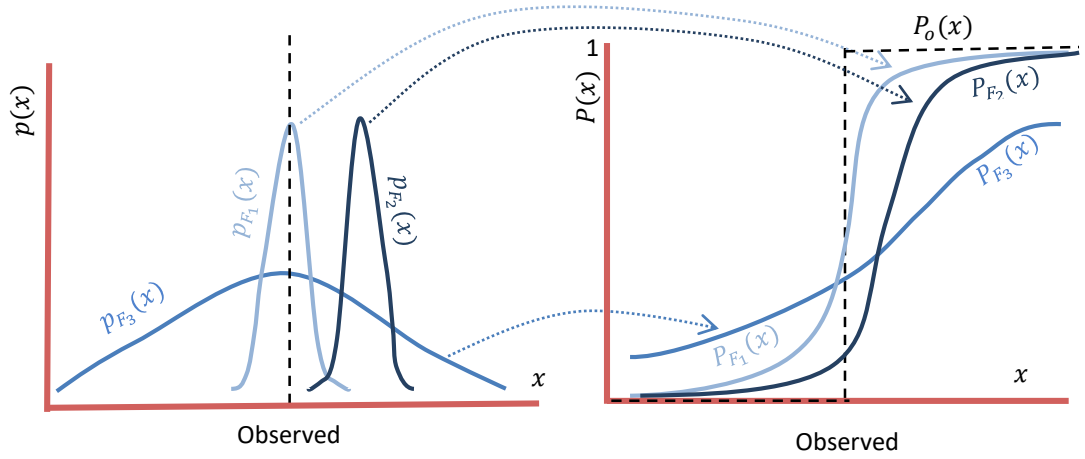

Figure 27. Schematic illustration of the continuous ranked probability score. Three forecast probability distribution functions $p_F(x)$ and corresponding cumulative distribution functions $P_F(x)$ are shown, together with the step-function cumulative distribution function for the observation $P_o(x)$ Distribution 1 would produce a small (good) score because $P_{F_1}(x)$ is the closest approximation to the step function (hence the smallest integrated squared difference). Distribution 2 concentrates probability away from the observation, and Distribution 3 is penalized for lack of sharpness even though it is centered on the observation.

where P_F stands for forecasted probability (cumulative distribution), while P_O is a cumulative-probability step function that jumps from 0 to 1 at the point where the forecast variable equals the observation. In other words, the *CRPS* can also be computed as the Brier score for binary events, integrated over all possible division points of the continuous variable y into the binary variable above and below the division point [Hersbach, 2000]. Additionally, for non-probabilistic forecasts, *CRPS* reduces to the (mean) absolute error.

The continuous rank probability score *CRPS* is a negatively oriented (the lower, the better) accuracy measure that is equivalent to the mean absolute error for deterministic forecast and also has a value of 0 for the perfect forecast. To better understand the meaning, an illustrative example of 3 forecast distributions is shown in Figure 27. Since the continuous rank probability score is the integrated squared difference between the cumulative distribution function and the step function representing the observation, cumulative distribution function that approximates the step function best (Distribution 1) produces relatively small integrated squared differences, and good scores. Distribution 2 is equally sharp but its displacement from the observation produces large discrepancies with the step function. This is especially the case for values of the predictand slightly larger than the observation, and hence very large integrated squared differences. Distribution 3 is centered on the observation, but much wider than the Distributions 1 and 2. Such a great width means that it is nevertheless a poor approximation to the step function and so also yields large integrated squared differences.

The *ROC* (relative operating characteristic, or receiver operating characteristic) diagram is another graphical forecast verification display. While the reliability diagram describes the calibration (distribution of observations conditioned on the forecast), the *ROC* diagram describes the likelihood (distribution of forecasts conditioned on the observation $p(f|o)$). Unlike the reliability diagram, it does not include the full information contained in the joint distribution of forecasts and observations. The base-rate (distribution of the observations $p(o)$) is not included and, hence, it is insensitive to conditional and unconditional biases (e.g., Jolliffe and Stephenson [2011]). To determine the *ROC* values, one contingency table is derived for several probabilistic thresholds (e.g. > 90%, >80%, >70%, etc). The probabilistic false alarm rate *F* and hit rate *H* (as defined in Eq.11) are calculated for a certain probability (e.g. 80%). Then, each *H* vs. F is plotted on the same graph to form the *ROC* curve. This curve must pass through points (0,0) and (1,1). It shows the ability of a set of probability forecasts to discriminate between the outcomes of a binary event (the event does or does not

occur). No-skill forecasts are indicated by a diagonal line (where *H=F*); the further the curve is towards the upper left-hand corner (where *H=1* and *F=0*) the better is the ability to discriminate the event. The example of two *ROC* curves on the *ROC* diagram is given in Figure 28.

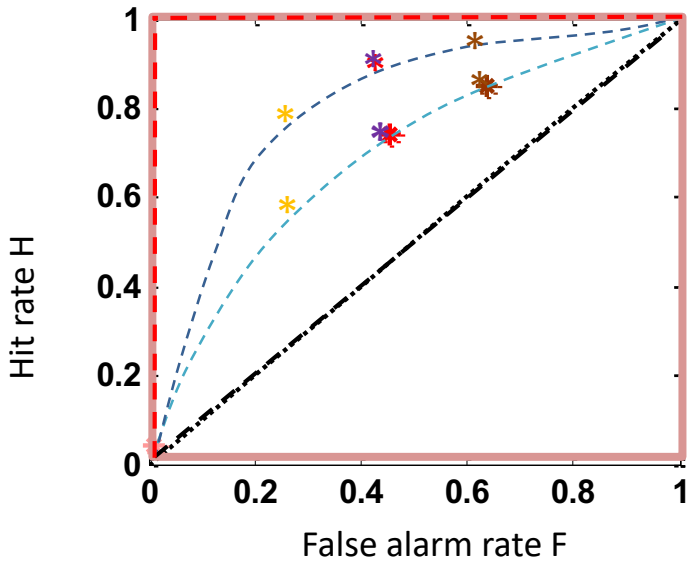

Figure 28. The illustrative example of the ROC diagram consisting of two ROC curves. Each curve is plotted through points derived from several (e.g. 3 marked with a different color for each forecast) contingency tables. Each contingency table uses a different probability threshold for probability forecasts of a pre-defined event. The black dashed line represents forecast with no skill, while the red dashed line represents the perfect forecast. Darker dashed blue line, by being closer to the top left corner, shows a better ability to discriminate different outcomes for a pre-defined event than the lighter dashed blue line.

4.6.1. Overall results

In total, six different analog-based ensemble experiments (see Table 5 for a summary) are carried out in this study. Results are evaluated against observations, the raw ensemble model, the ALADIN-LAEF (***LAEFws***) and the variations of the ***EMOS*** forecasts. The novelty of this approach is the usage of different types and setups of the probabilistic input model to give new insights into the analog-based methodology. Summarizing, all analog forecasts show an improvement compared to the raw forecasts during January (Table 6) and July (Table 7) 2018.

Moreover, most analog forecasts perform similar or even better than the ***EMOS*** methods. Furthermore, distinct differences between the analog configurations are found.

*Table 6. The average values and confidence interval (0.95 sig. level) of several verification measures for the different models at all available stations in Austria and all lead-times during January 2018. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value). The values significantly different from the **AnEnStd** forecast (0.05 sig. level) are marked with an asterisk sign.*

January	***LAEFws***	***EMOSws***	***EMOSstd***	***AnEnCtrl***	***AnEnWs***	***AnEnMu***	***AnEnStd***	***AnEnAll***	***AnEnMem***
Bias [ms^{-1}]	-0.210* *[-0.232, -0.185]*	-0.053* *[-0.069, -0.039]*	-0.160* *[-0.174, -0.146]*	-0.060* *[-0.072, -0.046]*	-0.036 *[-0.048, -0.022]*	-0.029 *[-0.042, -0.016]*	-0.023 *[-0.035, -0.011]*	-0.061* *[-0.075, -0.048]*	-0.048* *[-0.061, -0.034]*
CC	0.378* *[0.371, 0.385]*	0.831* *[0.826, 0.835]*	0.841* *[0.837, 0.845]*	0.841* *[0.837, 0.845]*	0.845* *[0.841, 0.849]*	0.861* *[0.858, 0.865]*	0.863 *[0.858, 0.865]*	0.863 *[0.860, 0.867]*	0.856* *[0.852, 0.860]*
Disp. Err [ms^{-1}]	2.670* *[2.645, 2.696]*	1.801* *[1.784, 1.826]*	1.705* *[1.681, 1.733]*	1.694* *[1.672, 1.715]*	1.705* *[1.682, 1.727]*	1.613 *[1.593, 1.633]*	1.608 *[1.589, 1.626]*	1.596* *[1.573, 1.618]*	1.634* *[1.612, 1.654]*
σ bias [ms^{-1}]	-1.501* *[-1.545, -1.458]*	-0.322* *[-0.378, -0.278]*	-0.454* *[-0.505, -0.404]*	-0.495* *[-0.546, -0.444]*	-0.391* *[-0.444, -0.340]*	-0.386 *[-0.438, -0.328]*	-0.372 *[-0.433, -0.314]*	-0.405* *[-0.455, -0.352]*	-0.420* *[-0.483, -0.367]*
RMSE [ms^{-1}]	3.070* *[3.029, 3.111]*	1.831* *[1.812, 1.851]*	1.772* *[1.748, 1.795]*	1.766* *[1.743, 1.792]*	1.749* *[1.729, 1.771]*	1.659 *[1.639, 1.677]*	1.650 *[1.632, 1.672]*	1.647 *[1.624, 1.667]*	1.688* *[1.670, 1.707]*
Spread [ms^{-1}]	0.850* *[0.846, 0.854]*	1.611* *[1.599, 1.622]*	1.605* *[1.592, 1.617]*	1.776* *[1.750, 1.779]*	1.663 *[1.650, 1.675]*	1.672 *[1.660, 1.686]*	1.667 *[1.655, 1.679]*	1.641* *[1.629, 1.654]*	1.728* *[1.714, 1.742]*
BSS (>5 ms^{-1})	-0.075* *[-0.093, -0.059]*	0.490* *[0.479, 0.500]*	0.515* *[0.505, 0.524]*	0.520* *[0.510, 0.529]*	0.513* *[0.504, 0.523]*	0.546 *[0.537, 0.555]*	0.549 *[0.541, 0.558]*	0.555 *[0.546, 0.563]*	0.526* *[0.517, 0.535]*
CRPS [ms^{-1}]	1.631* *[1.613, 1.648]*	0.883* *[0.875, 0.892]*	0.823* *[0.815, 0.831]*	0.814* *[0.806, 0.820]*	0.823* *[0.816, 0.831]*	0.777 *[0.770, 0.784]*	0.772 *[0.765, 0.779]*	0.769 *[0.762, 0.776]*	0.816* *[0.809, 0.823]*

Results show that the average bias of the ***LAEFws*** ensemble is small, underestimating the wind speed by 0.21 ms^{-1} in January and 0.23 ms^{-1} in July. The same results are found for the σ bias in July with 0.77 ms^{-1}, while it is a slightly more dominant source of error in January

with -1.50 ms^{-1}. The other evaluated scores such as the correlation coefficient (CC), which is on average higher in July than in January with 0.37, or the RMSE with 3.07 ms^{-1} in January and 1.79 ms^{-1} in July, indicate that the ***LAEFws***, in general, has realistic results, especially for the summer month. However, there are still some unresolved processes, as can be seen by the results of the dispersion error.

The main aim of any kind of the NWP model post-processing is to improve the results of the original model. This is the case here, too. The ***EMOS*** post-processing experiments are applied successfully, exhibiting the 0.46 maximum increase of the average correlation coefficient value. Moreover, the ***EMOS*** experiments are reducing all three error sources: the bias, the bias of the standard deviation (σ bias) and the dispersion error in comparison to ***LAEFws***. The ***LAEFws*** RMSE is, therefore, reduced by the ***EMOS*** experiments with the maximum 1.30 ms^{-1} difference among average values. The ***EMOSws*** is more successful in removing a systematic source of the error, while the ***EMOSstd*** is better in removing a dispersion error. All six analog-based experiments are able to outperform the ***LAEFws*** as well. Specifically, they can reduce all three error sources for the ensemble mean. Already the first and most "simple" experiments in terms of input data, the ***AnEnCtrl*** and the ***AnEnWs***, successfully remove the systematic errors in the bias and σ bias similar to the ***EMOS*** approach. Even more successful in removing the predominant dispersion source are the experiments with the additional predictors: ***AnEnMu***, ***AnEnStd***, and ***AnEnAll***.

In addition to improving the results for the ensemble mean, the average ensemble spread matches the average *RMSE* better after any post-processing. The ***AnEnStd*** exhibits the best spread among analog-based experiments in July, while ***AnEnAll*** shows better results in January. This might be related to the fact that wind speed shows greater variability (higher standard deviation of observations) and is probably harder to predict it correctly in January. For that reason, using more information from the raw model adds more variety to the ensemble members. This result also indicates that in the convective season most likely a horizontally and vertically higher resolved convection-permitting NWP model might add some additional information not present in the coarser ***LAEFws***.

In the selected two months, the observed frequency of the wind speed exceeding 5 ms^{-1} is higher for January with 18% cases than for July with 9%. Based on these observed numbers, the Brier skill score *BSS* value of the original ensemble (***LAEFws***) is -0.08 for January and 0.03 for July, indicating that the small differences are already present in the input data. It is

shown that the Brier skill score is improved by all post-processing experiments. This is especially the case in January, where the underlying climatology shows that the higher wind speed is more frequently observed than in July and the wind speed variance (higher standard deviation of observations) is higher. The ***AnEnMu, AnEnStd***, and ***AnEnAll*** experiments show a nearly similar improvement. The other post-processing approaches improve the Brier skill score *BSS* less.

*Table 7. The average values and confidence interval (0.95 sig. level) of several verification measures for the different models at all available stations in Austria and all lead-times during July 2018. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value). The values significantly different from the **AnEnStd** forecast (0.05 sig. level) are marked with an asterisk sign.*

July	***LAEFws***	***EMOSws***	***EMOSstd***	***AnEnCtrl***	***AnEnWs***	***AnEnMu***	***AnEnStd***	***AnEnAll***	***AnEnMem***
Bias [ms^{-1}]	-0.229* *[-0.242, -0.215]*	-0.001* *[-0.008, -0.010]*	-0.119* *[-0.129, -0.111]*	-0.012 *[-0.021, -0.001]*	-0.090* *[-0.099, -0.080]*	-0.055 *[-0.063, -0.046]*	-0.063 *[-0.072, -0.054]*	-0.088* *[-0098, -0.080]*	-0.043* *[-0.053, -0.033]*
CC	0.415* *[0.406, 0.422]*	0.750* *[0.745, 0.754]*	0.764* *[0.759, 0.768]*	0.752* *[0.748, 0.757]*	0.739* *[0.735, 0.744]*	0.770* *[0.766, 0.774]*	0.774 *[0.769, 0.778]*	0.774 *[0.770, 0.778]*	0.759* *[0.754, 0.763]*
Disp. Err [ms^{-1}]	1.602* *[1.589, 1.616]*	1.229* *[1.215, 1.240]*	1.144* *[1.132, 1.154]*	1.229* *[1.216, 1.241]*	1.262* *[1.250, 1.273]*	1.156* *[1.144, 1.167]*	1.145 *[1.136, 1.157]*	1.148* *[1.138, 1.159]*	1.183* *[1.172, 1.194]*
σ bias [ms^{-1}]	-0.773* *[-0.794, -0.754]*	-0.344* *[-0.368, -0.325]*	-0.474* *[-0.494, -0.452]*	-0.344* *[-0.364, -0.323]*	-0.331* *[-0.353, -0.308]*	-0.400* *[-0.418, -0.377]*	-0.409 *[-0.429, -0.387]*	-0.396* *[-0.416, -0.375]*	-0.403* *[-0.423, -0.383]*
RMSE [ms^{-1}]	1.794* *[1.775, 1.813]*	1.276* *[1.262, 1.288]*	1.244* *[1.234, 1.256]*	1.272* *[1.261, 1.284]*	1.307* *[1.294, 1.321]*	1.225 *[1.213, 1.237]*	1.219 *[1.208, 1.229]*	1.218 *[1.206, 1.228]*	1.251* *[1.238, 1.262]*
Spread [ms^{-1}]	0.651* *[0.648, 0.654]*	1.170* *[1.164, 1.176]*	1.138* *[1.133, 1.144]*	1.318* *[1.311, 1.326]*	1.256* *[1.248, 1.263]*	1.253 *[1.246, 1.261]*	1.244 *[1.236, 1.250]*	1.190* *[1.184, 1.197]*	1.301* *[1.294, 1.308]*
BSS (>5 ms^{-1})	0.032* *[0.009, 0.055]*	0.329* *[0.314, 0.345]*	0.337 *[0.322, 0.353]*	0.329* *[0.313, 0.344]*	0.319* *[0.303, 0.335]*	0.349 *[0.334, 0.365]*	0.355 *[0.341, 0.369]*	0.353 *[0.338, 0.369]*	0.325* *[0.310, 0.340]*
CRPS [ms^{-1}]	1.032* *[1.022, 1.042]*	0.648* *[0.643, 0.653]*	0.624* *[0.619, 0.629]*	0.636* *[0.631, 0.641*	0.650* *[0.645, 0.656]*	0.613 *[0.608, 0.618]*	0.610 *[0.605, 0.615]*	0.612 *[0.606, 0.617]*	0.635* *[0.630, 0.640]*

The ***LAEFws*** shows a higher continuous rank probability score *CRPS* (1.63 ms^{-1}) for January than for July (1.03 ms^{-1}). Again, the *CRPS* value is improved by all post-processing experiments, exhibiting better overall results for July than in January, when wind speed and its variance is higher on average. Similar to the Brier skill score *BSS*, the ***AnEnAll*** shows the highest skill during the winter month, while the ***AnEnStd*** is slightly better during the summer month. This indicates that not just that adding more input from the raw model increases the ensemble spread, but it also improves its accuracy. The ***AnEnMu*** follows both ***AnEnAll*** and ***AnEnStd*** results closely**.** The other post-processing experiments are not as successful, exhibiting significantly worse overall results for both months investigated.

4.6.2. Lead time performance

To investigate six analog-based ensemble experiments comparison further, a summary continuous rank probability score *CRPS* is considered for the individual lead-times (Figure 29). The result shows that there is no significant difference between the ***AnEnMu***, ***AnEnStd*** and ***AnEnAll*** performance during neither winter nor summer month. The ***AnEnCtrl, AnEnWs***, and ***AnEnMem*** are slightly outperformed by other analog-based experiments, especially for January. Even though the ***AnEnCtrl, AnEnWs***, and ***AnEnMem*** can improve the raw NWP forecasts, comparable to the ***EMOS*** approach, they are less promising than other analog-based experiments. The ***AnEnWs*** results show that it essential to use more than one meteorological variable as a predictor in the analog approach. This can be explained by the better ability of the analog method to distinguish different seasonal and synoptic situations. The analog-search pool in the ***AnEnMem*** experiment is smaller than in other analog experiments since the search is performed dependently for the same ensemble member. Possibly, that is why the ***AnEnMem*** would not increase the skill of the raw probabilistic input, as one would inherit undesirable properties of the input model, such as under-dispersion and lower resolution issues. Additionally, ***AnEnMem*** is the most computationally expensive setup. For these reasons, it is not shown or discussed further in the thesis (results can be found in the Appendix). Finally, even though the ***AnEnCtrl*** and the ***AnEnMu*** use the same number of the meteorological parameters as predictor variables, the

AnEnMu performs better for both months and at all lead times tested. Similar results are shown in Dabernig et al. [2015], where the ***EMOS*** results based on ensemble forecasts outperformed the forecasts using only the control run.

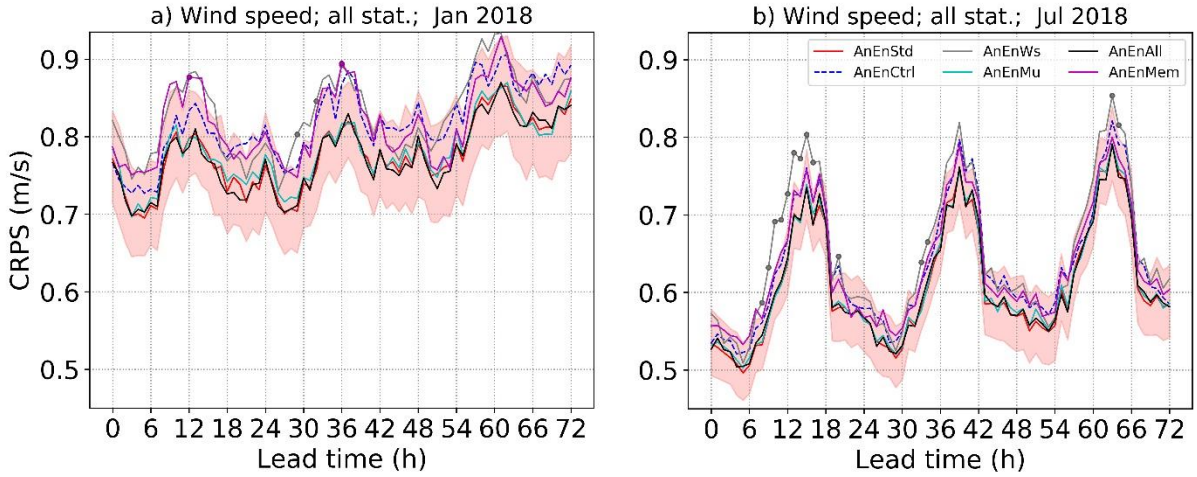

*Figure 29. Continuous rank probability score depending on lead-time for five different analog-based ensemble experiments during January (left) and July (right) 2018 at 29 stations in Austria. The markers are set for the results significantly different from the **AnEnStd** forecast (95% confidence level), while the red shaded area represents the AnEnStd 95% confidence interval calculated by the bootstrap percentile method [Jolliffe, 2007].*

Overall, the ***AnEnAll*** performs the best in post-processing for January whereas the ***AnEnStd*** setup performs the best for July. Among these experiments with a similar result, the ***AnEnStd*** is chosen as the best representative. The reason for this decision is that it is not computationally demanding as the ***AnEnAll***, while it includes the information about raw model spread (unlike the ***AnEnMu***). The information about the raw model error growth is considered as a very important aspect of the raw NWP ensemble forecast. Therefore, it is expected to be further developed in the near future, leading to greater differences between the ***AnEnMu*** and ***AnEnStd*** experiments. To determine if using summarized predictors, such in the ***AnEnStd*** experiment, leads to information loss and decreases the forecast quality, the results are compared to the ***AnEnAll*** experiment. In addition to overall comparison, the ***AnEnStd*** and ***AnEnAll*** experiments are also compared against the two different ***EMOS*** experiments and the ***LAEFws***, separated into lead-times using several verification metrics.

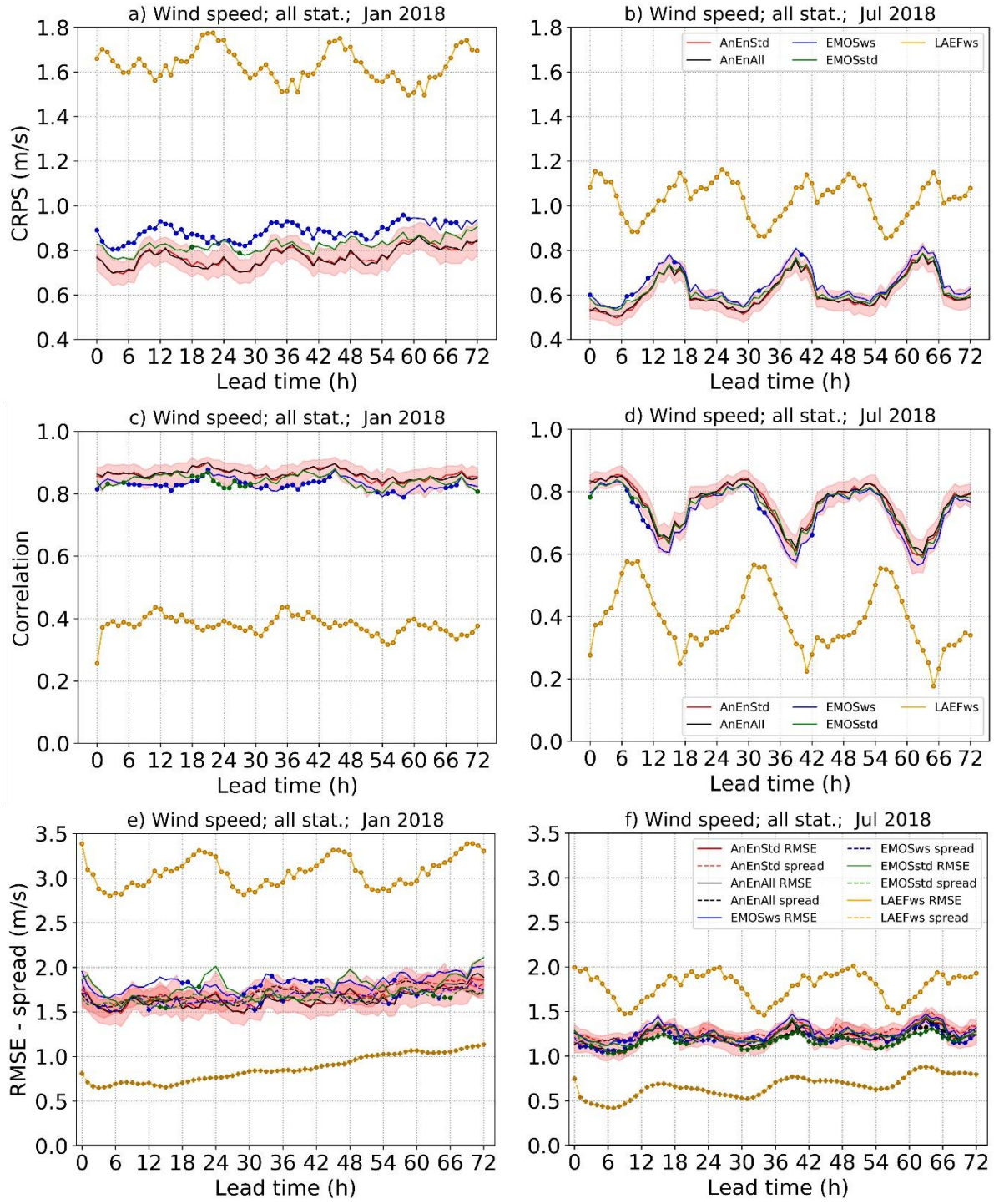

*Figure 30. Continuous rank probability score (top), the correlation coefficient for the ensemble mean (middle) and the spread-skill diagram (bottom) depending on lead-time for the raw **LAEFws** ensemble, the **EMOSws** and two different analog ensemble configurations at 29 stations in Austria tested for January (left) and July (right) 2018. The markers are set for the results significantly different from the AnEnStd forecast (95% confidence level), while the red shaded area represents the AnEnStd 95% confidence interval calculated by the bootstrap percentile method [Jolliffe, 2007].*

The continuous rank probability score *CRPS* shows that the ***LAEFws*** exhibits a higher skill during daytime (i.e. 0600 - 1800 UTC) than during nighttime, and higher during July (Figure 30b) than during January (Figure 30a). The ***EMOS*** and the analog-based experiments are more skillful during nighttime than during daytime. The improvement over the ***LAEFws*** after post-processing is greater in January for both the ***EMOS*** and the analog approach since the ***LAEFws*** is worse than in July.

However, the ***EMOS*** and the analog experiments are overall better in July, when the ***LAEFws***, which also served as input, is better. These results imply that the best result is achieved when the input model is also working better. The ***AnEnStd*** and ***AnEnAll*** show almost no difference. They are both more skillful than the two ***EMOS*** experiments. Even though the differences are often subtle, they are significant for the ***EMOSws*** at almost all lead-times during January and at several lead-times during July, especially within the first 24 hours.

Evaluating the dependency on the lead-time, the analog post-processing methods show considerable improvement over the ***LAEFws*** for both months tested with the correlation coefficient *CC* (Figure 30c-d). The analog approach outperforms the ***EMOS*** methods in terms of correlation, often significantly. This is especially the case for January when the correlation enlargement over ***EMOSws*** is significant for almost all lead-times and sometimes even over ***EMOSstd*** (i.e. during nighttime).

The analog-based forecasts exibit a major statistically significant reduction of the ***LAEFws*** *RMSE* at all lead-times (Figure 30e-f), similarly to the ***EMOS*** approach, with very few significant differences. The improvement is the most evident for the ***LAEFws*** *RMSE* maxima at 0000 UTC.

Similar results can be found in the spread-skill diagrams. These diagrams test if the average ensemble spread matches the average *RMSE*, representing the forecast uncertainty appropriately. The ***LAEFws*** experiment shows a strong underestimation of spread. All post-processing methods satisfactorily increase the spread. Here, both analog-based forecasts are showing a major imprevoment in spead-skill relationship with an almost perfect agreement between the *RMSE* and the spread, while the ***EMOS*** experiments are slightly under-dispersive, especially the ***EMOSws*** in January (Figure 30e). This can be related to the fact that it uses only the wind speed as a predictor and most likely, not enough dispersion

information is given. Additionally, the ***EMOSws*** only uses a 30-days training window, which also results in a small under-dispersion.

4.6.3. Spatial performance

The climatology in Figure 19 shows that the wind speed increases towards the northeastern part of Austria (Pannonian Plain) for both January and July, which also suggests a spatial pattern in forecast performance. Within this subsection, it is decided to show only results for January, since the previous results suggested the better distinction in the performance after post-processing. Even though not shown here, the spatial distribution of results for July is very similar to the ones for January. The results for both months are shown in Appendix B.

Additionally, due to very subtle and hardly notable differences among analog experiments, only the ***AnEnStd*** configuration is shown as a representative. The results for the ***AnEnMu*** and the ***AnEnAll*** experiments are almost indistinguishable from the ***AnEnStd***, while the ***AnEnCtrl***, ***AnEnWs***, and the ***AnEnMem*** are the same or slightly worse. Since the results for these experiments carry no new information within this subsection, they are not shown from this moment on (but can be found in Appendix B).

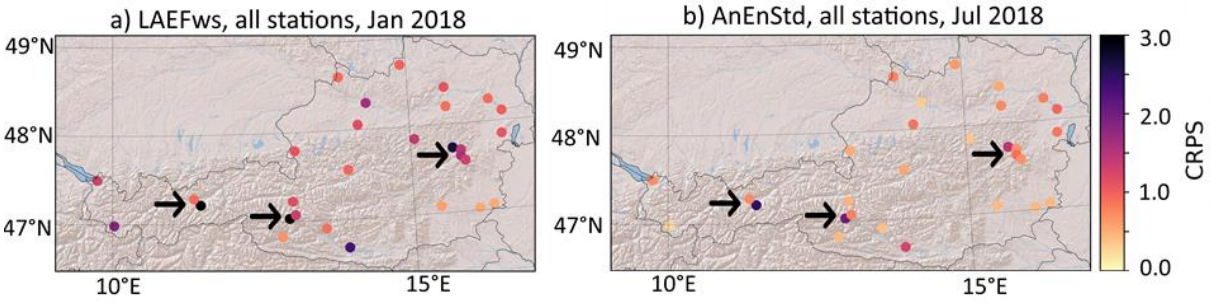

*Figure 31. The spatial distribution of the monthly mean continuous rank probability score for the raw **LAEFws** (left) and the **AnEnStd** (right) for January 2018. The arrows point to closely situated stations in the highly complex topography, where the valley stations exhibit much better results than the mountain stations.*

The value for the ***LAEFws*** monthly mean continuous rank probability score *CRPS* is following the climatological wind speed pattern, having higher values at the stations prone to higher winds. The plains are better represented by the ALADIN-LAEF topography and,

therefore, the performance of ALADIN-LAEF is, in general, better at lower altitudes and less complex topography. The error is reduced for the analog experiments (Figure 31b) compared to the ***LAEFws*** (Figure 31a), following a similar pattern. Additionally, there are large differences for the nearby stations situated in highly complex topography. The mapped *CRPS* values for any forecast tested show that the valley stations are better predicted than the mountain stations (arrows), especially for the ***LAEFws***. A close look at the two stations in Innsbruck (arrow in the west of Austria) shows, for example, that the ***AnEnStd*** *CRPS* at the valley station is improved by around 20% compared to the ***LAEFws***. As the ***LAEFws*** performance at mountain stations is not as efficient, this leaves room for improvement. Here, the *CRPS* can be improved by around 70% at e.g. Patscherkofel, the mountain station close to Innsbruck. A similar pattern is shown at the station Sonnblick (arrow in the middle) where the mountain station has much higher *CRPS* values (raw and post-processed) compared to the valley station. As an example, for the three sites located in the Semmering region (most eastern arrow), a mountain pass in the east of Austria, the different settings of the sites can be one of the factors. The site located at the pass is prone to the gap flows (e.g. Mayr et al. [2007]), whereas the site at the mountaintop is located within the skiing resort, somewhat shielded by the nearby hut and not represented in the model lower boundary conditions. The site located at the valley shows again the lower *CRPS* values. These differences in predictability are mainly related to the high wind speeds and the coarse resolution of the raw model. This suggests a large sensitivity of the models in the Alpine complex topography to the exact details of the mountain height and shape, as well as the incoming background layer, where subtle differences can result in a large range of responses in the downslope wind regime. In contrast, the stations in the north-east of Austria (around Vienna) are also climatologically prone to high wind speeds but show much better *CRPS* values.

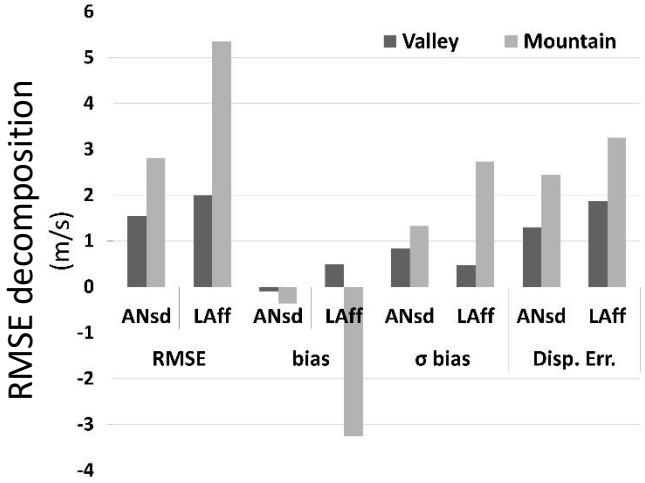

*Figure 32. The **AnEnStd** and the **LAEFws** performance comparison at mountain and valley stations by root-mean-square error decomposition into bias, σ bias, and dispersion error during January 2018.*

To evaluate the performance for valley and mountain stations, the stations marked with arrows (Figure 19 and Figure 31) are investigated separately. For the valley stations, the *RMSE* (1.50 ms^{-1}) shows that the ***LAEFws*** wind speed prediction performs adequately. However, for mountainous sites, the *RMSE* is 6.24 ms^{-1}, due to the aforementioned reasons. The RMSE is notably reduced by the analog approach, by 0.45 ms^{-1} at the valley and by 3.33 ms^{-1} at mountain stations. The *RMSE* decomposition (Figure 32) shows that the dispersion error is notably reduced by the analog approach, slightly more for the mountain than the valley sites. The ***LAEFws*** exhibits much larger systematic errors for the mountain than the valley stations. The ***LAEFws*** bias and the σ bias at the valley stations are very small, to begin with. The analog approach is therefore not able to make a large difference after post-processing. On the other hand, the ***LAEFws*** systematic sources of error at the mountain stations are much more pronounced than at the valley stations. These sources of error are yet again successfully removed by the analog approach. The *RMSE* reduction is therefore much more noticeable for the mountain stations than for the valley stations, due to the reduction of systematic sources of error, which are not as present in the raw model for the valley stations. However, the spatial distribution of forecast performance could be further investigated in future work.

4.6.4. Special diagrams: reliability, ROC and rank histograms

The reliability of a probabilistic forecast is the property of that forecast to predict probabilities that match the relative frequencies within the data. Here, it is evaluated for the probability of wind speed exceedance of > 5 ms^{-1}. Again, the ***LAEFws*** ensemble has lower reliability in January (Figure 33a) than in July (Figure 33b). Furthermore, it is below the no-skill line for the high probabilities in January. Both ***EMOS*** experiments improve ***LAEFws*** reliability***,*** ***EMOSstd*** improving it a bit more than ***EMOSws***. However, the analog experiments show an even higher resolution and reliability across all experiments, especially for the winter month. The differences can be noticed for the probabilities up to a 50% chance of wind speed to exceed 5 ms^{-1}, where the ***EMOSstd*** is slightly underconfident, or for the probabilities with a more than 40% chance, where the ***EMOSws*** is slightly overconfident. Between the analog experiments, only small and insignificant differences are found. Both analog-based experiments exhibit almost perfect reliability for winter month almost perfectly, while being slightly overconfident during summer.

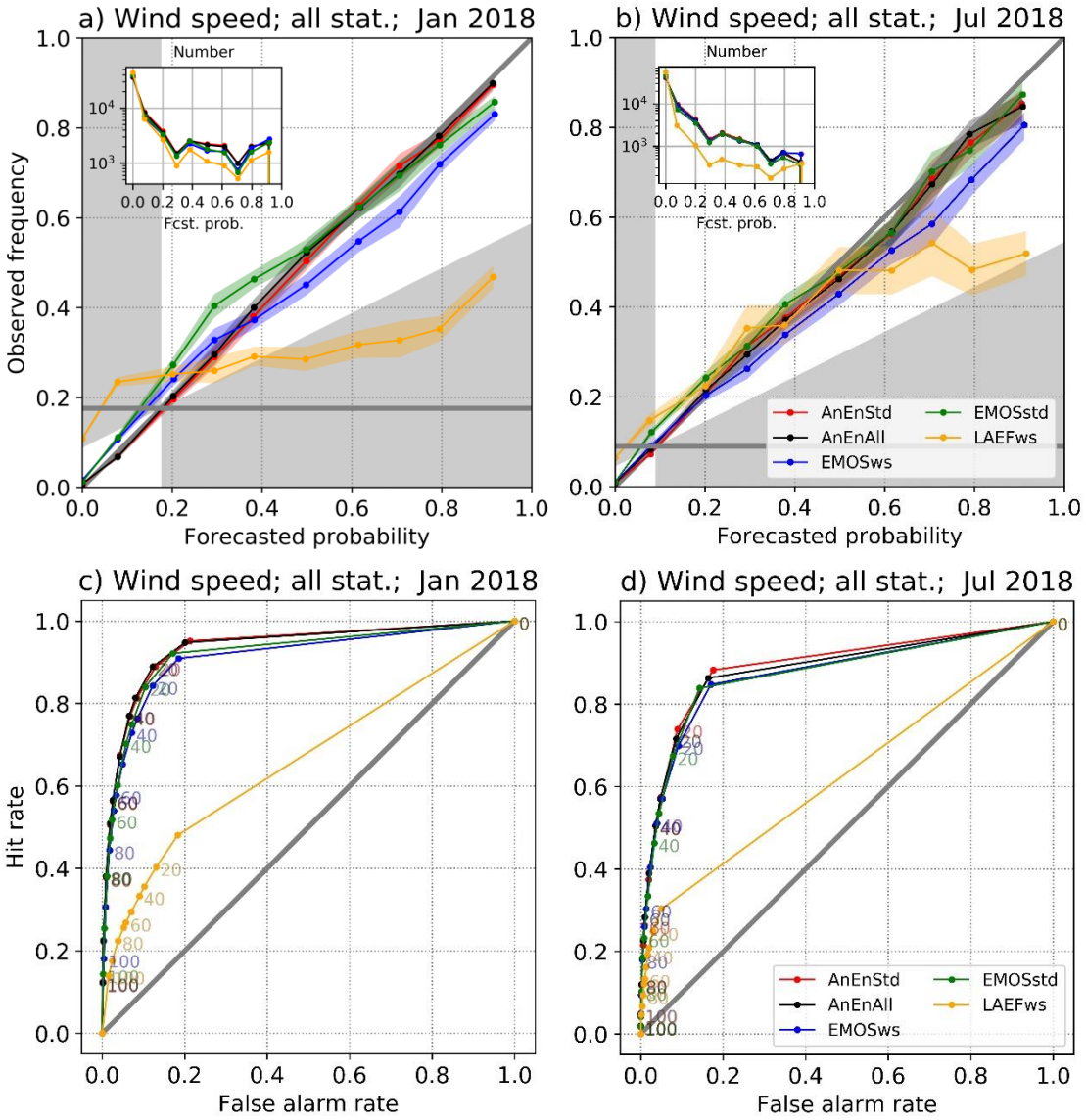

*Figure 33. Reliability diagrams (top) and relative operating characteristic (ROC) diagrams (bottom) for two different analog forecasts and a threshold of > 5 ms^{-1}, compared to the raw **LAEFws** and the **EMOSws** during January (left) and July (right) 2018 at 29 stations in Austria. The dashed lines in the reliability diagrams show a 95% confidence interval, while the sharpness diagrams are shown in the upper left corners.*

Besides a higher resolution of the analog experiments, one can notice that the sharpness property (the diagram in the upper left corner of the reliability diagram) is satisfactory for all approaches, exhibiting moderate to high forecast confidence. However, the ***LAEFws*** is a bit sharper than the post-processing experiments, indicating a higher tendency to forecast extreme probabilities. This is preferable because of the better forecast usability if the forecasts

are reliable. Still, the post-processing experiments are overall more accurate in terms of reliability.

The *ROC* curve shows a ratio of hit rate versus false alarm rate using a pre-defined threshold. Again, the threshold of 5 ms^{-1} is used. The *ROC* curve (Figure 33c-d) indicates that the analog methods, in general, improve the raw ***LAEFws*** forecasts comparable to or better than the ***EMOS.*** Unlike other measures, the reliability and discrimination property exhibit higher values for January than for July. However, this might be due to the higher climatological frequency of such wind speeds in January (18%) than in July (9%). For that reason, the differences among winter and summer month should not be investigated by using the fixed threshold. The results should be used for comparison among different experiments. The ***AnEnStd*** exhibits a slightly higher hit rate than the ***AnEnAll*** and ***EMOS*** experiments, especially for July.

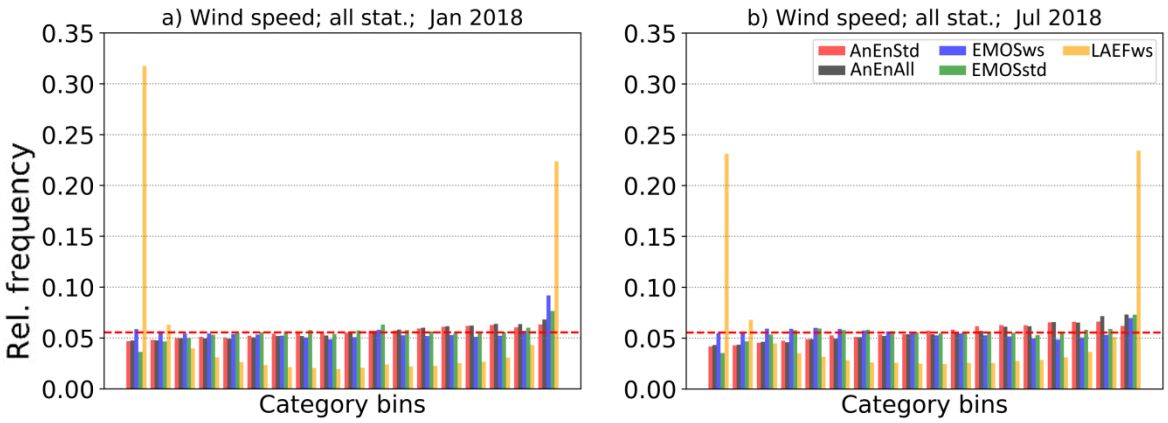

*Figure 34. Rank histograms for the **AnEnStd** and **AnEnAll** compared to the raw **LAEFws**, **EMOSws**, and **EMOSstd** forecasts during January (left) and July (right) 2018 at 29 stations in Austria.*

Evaluating the rank histogram (Figure 34), a clear under-dispersion of ***LAEFws*** is found, especially for January. This is not the case for the post-processed forecasts. It shows that the analog method is able to improve the dispersion of the original NWP ensemble.

Finally, it is shown that the analog approach outperforms the raw ***LAEFws*** model in terms of better accuracy, reliability, resolution, discrimination and spread for both winter and summer months. The results are very similar to or better than the ***EMOS*** experiments shown, with the larger differences during the winter month. The difference among analog experiments (***AnEnAll*** and ***AnEnStd***) is barely notable. Therefore, it is indicated that using

the summarized metrics of the raw model meteorological variable ensemble as a predictor in the analog approach barely sacrifices the forecast quality, while saving computational power. Special diagrams for the other analog-based experiments (***AnEnWs***, ***AnEnCtrl***, ***AnEnMu*** and ***AnEnMem***) can be found in Appendix C.

4.6.5. High wind speed predictions

The majority of measured wind speed values during the selected months are within 2-3 ms^{-1} range (30-40%), while the wind speeds higher than 10 ms^{-1} are rare (Figure 35c-d). However, it is not less important to properly forecast higher wind speeds as of their higher impact on people and damage on the property, road and air traffic disruptions, wind energy production, and others. For this reason, it is important that a probabilistic forecast is consistently good for several different thresholds. Besides the exceedance of 5 ms^{-1} the thresholds ranging from 0.5 ms^{-1} to 20 ms^{-1} are investigated (Figure 35a-b).

The Brier skill score *BSS* indicates that the ***LAEFws*** forecast is somewhat skillful in reproducing wind speeds of the order of 3 ms^{-1}, but shows much less skill, if any, for the higher and lower thresholds. The ***EMOS*** approach is more skillful than the ***LAEFws*** for any threshold value in January and up to 10 ms^{-1} (***EMOSws***) or even 15 ms^{-1} (***EMOSstd***) in July. The analog experiments are able to improve the forecast skills up to 10 ms^{-1} significantly better than the ***EMOS*** experiments. Approaches as in Baran and Lerch [2016] could be used to adjust EMOS to higher wind speeds but have not been tried. Furthermore, the ***AnEnStd*** and ***AnEnAll*** improve the ***LAEFws*** forecasts for all thresholds investigated for January. Again, the ***AnEnCtrl***, ***AnEnWs***, and ***AnEnMem*** do improve the ***LAEFws*** forecasts but are less skillful than the other analog experiments (shown in Appendix D). However, ***AnEnWs*** still provides a good result. It is, thus, recommended approach if only a reduced set of ensemble data is available or the computational resources are limited. These results reveal the potential for post-processing using the analog approach, even though one needs to be careful with the interpretation since the number of occurrences of high wind speed (i.e. around 20 ms^{-1}) is very small.

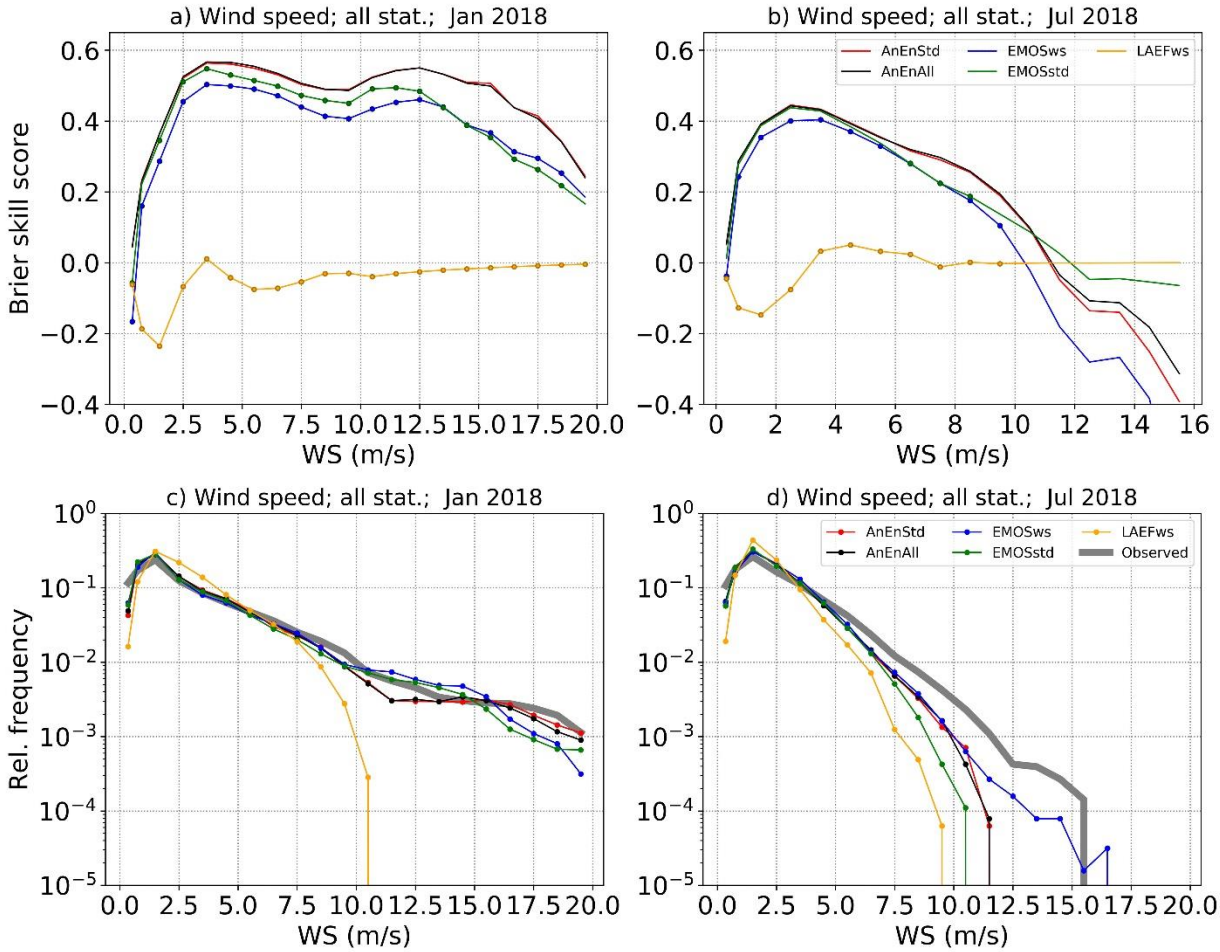

*Figure 35. Brier skill score (top) and relative frequency (bottom) depending on a wind speed threshold. The analog probabilistic forecasts are compared to the raw **LAEFws** and the **EMOS** forecasts during January (left) and July (right) 2018 at 29 stations in Austria. The markers are set for the BSS results significantly different from the **AnEnStd** forecast (95 % confidence level).*

§ 5. SUMMARY AND DISCUSSION

The development of suitable post-processing methods that reduce starting model errors at locations where measurements are available is needed since even state-of-the-art mesoscale models still exhibit considerable errors, especially in complex topography. The answer might lie in the several-decades-old idea to use analogies (i.e., similar past forecast, measurements or analysis) for forecasting future weather. The idea is based on an assumption that if two atmospheric states are initially very close, they will remain somewhat close for some time in the future. More recently developed formulation of an analog-based post-processing method is already proven to improve deterministic numerical weather prediction (NWP). The analog-based method uses a historical data set including NWP data and observations at a single site. The output of the analog-based method is an analog ensemble (AnEn), which can then be used to issue a deterministic forecast.

First, a deterministic NWP is tested as an input to the analog-based method – an operational limited-area mesoscale NWP model Aire Limitée Adaptation dynamique Développement InterNational (ALADIN) of the Croatian Meteorological and Hydrological Service. The deterministic output of analog-based experiments includes forecasting the mean (***AN***) and median (***ANM***) of the AnEn. Since the other experiments produce better results than the ***ANM***, and specific benefits are not achieved in tested cases presented in this work, results for the ***ANM*** are mostly discarded. The results for the ***AN*** are compared to a linear, adaptive and recursive Kalman filter (KF) post-processing approach. The KF algorithm is first applied using the same NWP and observations as in analog-based experiments, resulting in deterministic Kalman filter prediction (***KF***). Additionally, two experiments that combine analog and Kalman filter approaches are also performed. The first one is applying the KF to the time series of the ***AN*** forecasts, resulting in a new deterministic forecast called the ***KFAN***. Additionally, the KF is applied to the same historical set of NWPs and verifying observations but in the analog space, ordered from the worst to the best analog (Kalman Filter in Analog Space – ***KFAS.*** Therefore, the analog-based experiments include ***AN***, ***KFAN***, and ***KFAS***.

In this research, an in-depth analysis of the analog-based method over a complex topography is performed. The target area of this research is a coastal complex topography in Croatia, where the most significant portion of mesoscale energy is governed by strong

downslope windstorms as well as thermally induced land-sea circulations. Additionally, the research includes mountain complex topography and continental nearly flat topography.

By analyzing root-mean-square-error (*RMSE*), rank correlation coefficient (*RCC*) and bias of the mean it is shown that all tested post-processing methods improve results of the ***A8*** starting model. The best results are obtained for the analog-based method when using 15 analog ensemble members. The *RMSE* and bias growth is noticed for larger ensembles, probably due to climatological differences between training and verification period.

The ***KF*** and the ***KFAN*** are the most successful post-processing methods for bias reduction. That is expected result since the KF is constructed to remove the systematic error if it does not change rapidly (i.e. large hour-to-hour variations). However, the application of the KF can also lead to a decrease in the correlation coefficient (i.e. increase of the dispersion error). The dispersion error is noticeably reduced by the ***KF*** approach in the flat topography, where there are some indications of a systematic error influencing a large scale (i.e. period longer than 10 days) strong wind in the model. The ***KF*** is not as successful in reducing the dispersion error in the coastal complex area. The analog-based method reduces dispersion error (i.e., improve *RCC*) regardless of the topography complexity, showing greater adaptability than the ***KF*** forecast. The ***AN*** seems to be the most suitable post-processing method for *RCC* improvement. The standard deviation (σ) of the ***KF*** forecasts is closer to the observed standard deviation σ than for the raw model, especially in the complex topography. The analog-based method is also prone to the same underestimation but not as much as ***KF*** in the complex topography. The underestimation of the measured standard deviation σ for the analog-based method is partially explained by climatological differences between the training and testing period. The ***AN*** forecast is the most prone to systematic underestimation of the standard deviation σ among analog-based forecasts. This is due to additional averaging when forecasting the ensemble mean, thus naturally reducing the variability of the forecast. This systematic error is partially removed by the application of the KF in the ***KFAN*** forecast. The ***KFAS*** forecast exhibits the highest standard deviation σ among the analog-based experiments due to better adaptability. Finally, even though the analog-based method affects different aspects of the starting model, the *RMSE* reduction is very similar among them and superior to the ***KF*** approach. This is especially the case in the coastal complex topography.

The importance of the starting model resolution and formulation is investigated by using three configurations of ALADIN model run: with 8 km grid spacing (ALADIN; ***A8***) and 2 km

grid spacing (ALADIN; ***A2***, dynamical adaptation; ***DA***), as well as verifying observations of 10-m wind speed. Comparison of post-processing methods performance with starting models at 2 km grid spacing (***A2***, ***DA***) compared to the post-processing performance with the ***A8*** as a starting model shows that post-processing methods considerably improve numerical predictions for all tested model resolutions. We furthermore test the hypothesis that the greater the representation of physical processes directly simulated by the starting model, the better is the quality of the analogs. Even though the higher-resolution starting models yield better statistical results themselves in our target area (coastal complex topography), it is not necessarily the case for the analog-based forecasts generated by the higher resolution model. This may be due to the imperfections of the point-based verification metrics used that typically increase with a resolution at near-kilometer scale grid spacing of numerical models (i.e. high sensitivity to spatial and phase errors). Therefore, the categorical and spectral analyses are performed to explore the potentially undetected benefits of using a higher resolution model further.

To assess the performance of forecasts for different wind speeds, we performed a verification using three wind speed categories: weak, moderate and strong wind. The categories are divided by 50^{th} and 90^{th} percentile. The polychoric correlation coefficient for categorical forecasts leads to similar conclusions as to the rank correlation coefficient analysis. The ***DA*** and the ***A2*** exhibit higher association in the coastal complex but not in the other topography types. Association is significantly improved by almost all post-processing methods, except the ***KF*** forecast in the coastal complex topography. Averaged over all locations, the analog-based method achieves better both rank and polychoric correlation coefficient results than the ***KF*** in general, particularly the ***AN***.

Averaged over all locations, starting models forecast weak wind occurrence too rarely and moderate wind occurrence too often. For coastal complex topography, different starting models yield different frequency bias. Starting models ***A8*** and ***A2*** over-forecast the occurrence of moderate wind category while under-forecast the occurrence of the strong wind. The ***DA*** seems to be the least (frequency) biased model in the coastal complex topography. For other topography types (mountain complex and nearly flat continental) all starting models tested in this study under-forecast the frequency of weak wind and over-forecast the frequency of moderate and strong wind. All post-processing methods significantly reduce the frequency bias for climatologically common wind speed categories on average. While the

results for the ***KF*** are slightly less biased, the main challenge for the analog-based method seems to be the under-forecasting of strong wind occurrence. The ***KFAS*** seems to be the least biased analog-based experiment, showing the best result for strong wind while being as unbiased as the ***AN*** in the other two categories. It has to be noted that the results in the moderate and strong wind speed categories exhibit very large confidence intervals, providing only indications of the post-processing methods' ability to improve the starting model forecast.

The critical success index (*CSI*) is a measure of the relative accuracy of a categorical forecast. The ***KF*** has considerably higher relative accuracy than the starting models for weak wind category in the nearly flat continental and mountain complex topography, but not as much in the coastal complex topography. Results suggest that the relative accuracy result is improved for the moderate and strong winds as well. The analog-based experiments seem to outperform both starting models and corresponding ***KF*** forecasts for all the cases tested, except the prediction of the strong wind in the nearly flat continental topography. For the latter, the ***KF*** seems to be the best post-processing method once again suggesting consistent model error when predicting strong wind. The ***AN*** achieves the highest relative accuracy for weak wind, while the ***KFAN*** and the ***KFAS*** seem to be better in predicting the other categories.

Using a model at finer horizontal resolution leads to improvements in the relative accuracy for starting model predictions of the strong wind in the coastal complex topography. This confirms that finer resolution modeling in coastal complex topography leads to a better ability of the forecast in distinguishing low from moderate or unusually strong wind. This horizontal resolution increase yields mixed results for other categories and topography types, potentially due to the nature of time-space model errors and the related statistical imperfections of the metrics. This property is then inherited by all of the post-processing methods. However, the results corresponding to moderate and especially strong winds could be further reinforced using a larger sample size. However, enlarging the sample size is contradictory to the basic post-processing idea: it needs to be efficient but also quick and easy to implement. Every time there is an update in an NWP model, the method needs to be re-trained. That means that historical NWP forecasts need to be simulated, which is a computationally demanding procedure. Therefore, it is rarely done for periods longer than 1

or 2 years. That dataset is then used for training and new forecasts are issued. However, this is only up until the model is updated again (i.e. few years maximum).

The measure extremal dependence index (*EDI*) independent of the underlying climatology and, for that reason, also used to evaluate the forecast of rare events (i.e. strong wind). The results are generally consistent with the relative accuracy analysis (measured by *CSI*), with smaller confidence intervals. Overall, the analog-based method performs better than the ***KF***, especially the ***KFAN*** forecast. The analog-based method is more successful if it is started with ***A2*** than if it is started with ***A8*** or ***DA*** models, which is consistent with the previous results.

The spectral analysis suggested that the ***KF*** approach affects only (very) large scale motions (i.e. period longer 10 days) if the power spectral density function is biased. The ***KF*** thus enlarges the energy of the large-scale motions in the coastal area and reduces the energy of the large-scale motions at the nearly flat continental topography. However, the KF does not significantly influence the shorter time scales. The ***KF*** might be slightly adjusted by optimizing the parameters of the KF, affecting somewhat shorter scales (e.g. synoptic), However, the qualitative effect of affecting only large scale motions would presumably remain the same. In other words, the ***KF*** does not significantly influence the energy of the shorter time scale motions.

Unlike the ***KF*** approach, introducing past similar situations in the analog-based method leads to better forecasting processes on a longer-than-diurnal scale. The longer-than-diurnal scales are much more relevant than the larger scales (i.e. a period longer than 10 days) for forecasts up to 72 h ahead. The analog-based method improves model underestimation of the longer-than-diurnal motions in the coastal area and in the nearly flat topography when the model overestimates the longer-than-diurnal motions. The ***KFAS*** method is superior to the other post-processing methods because it maintains the modeled energy for shorter-than-diurnal part of the power spectra (unlike the ***AN***), while it improves both under- and overestimation of the longer-than-diurnal motion energy (just as good as or better than the ***AN***). Furthermore, higher-resolution models ***A2*** and ***DA*** generally contain more energy than ***A8***. Consequently, there are fewer situations with under-predicting large-scale motions. But when they do occur, the post-processing methods behave as presented for the lower resolution ***A8*** model. Even though the analog-based experiments often under-predict the shorter-than-diurnal motions, they simulate the correct amplitude of the diurnal cycle harmonics (24-h, 12-

h, and 8-h), similarly to model. Additionally, even if the model over-predicts the amplitudes of the diurnal cycle harmonics, the analog approach reduces them.

Table 8. The summarized results for the post-processing of the deterministic NWP regarding benefits and limitations the post-processing methods used in this thesis.

Forecast	**Benefits**	**Limitations**
KF	+ Bias reduction + The standard deviation unbiased in mountain and flat topography + Less prone to underestimate the occurrence of strong wind category + Best relative accuracy for the strong wind in the flat topography	– Possible correlation decrease (in complex topography) – Standard deviation underestimated in coastal complex topography – Affecting only the (very) large scale motions (i.e. period longer than 10 days)
AN	+ Best correlation increase + Best relative accuracy for the weak wind speed category + Better forecasting processes on a longer-than-diurnal scale	– Bias increase for large ensembles – Prone to underestimation of the variability (σ) in mountain and flat topography – Prone to underestimate the occurrence of strong wind category – Underestimates shorter-than-diurnal scale motions
KFAN	+ Correlation increase + Adequate relative accuracy for strong and moderate wind speed category + Better forecasting processes on a longer-than-diurnal scale	– Bias increase for large ensembles – Prone to somewhat underestimate the variability (σ) and the occurrence of strong wind category – Underestimates shorter-than-diurnal scale motions
KFAS	+ Bias reduction + Correlation increase + The least prone to underestimation of standard deviation overall + Less prone to underestimate the occurrence of strong wind category than other analog experiments + Adequate relative accuracy for strong and moderate wind speed category + Better forecasting processes on longer-than-diurnal scale + Better forecasting processes on a shorter-than-diurnal scale	– Prone to somewhat underestimate the variability (σ) and the occurrence of strong wind category

Finally, one can conclude that each post-processing method tested in this thesis successfully improves the deterministic NWP wind speed forecasts. However, each post-processing method also has its strengths and weaknesses, and the choice for operational use depends on the envisaged purpose. For that reason, the benefits but also the limitations of the post-processing method tested are listed in Table 8. Hence, one can decide on the most suitable approach according to the statistical properties of the starting model deterministic NWP and potential user-specific needs.

The availability of the quality data over mountain complex topography in Croatia is limited. Only three locations satisfy the necessary quality demands for the analog method testing and implementation in the first part of this research. Hence, the research is extended using 29 meteorological observation sites (TAWES) in Austria for winter (January) and summer (July) month of 2018. Additionally, after investigating wind speed as continuous and categorical predictand, the focus is shifted to the ensemble and probabilistic wind speed forecasting. In addition to using deterministic NWP input to analog-based method, the ability to calibrate the ensemble NWP is also investigated. Therefore, an in-depth analysis of the analog-based method applied to the Austrian ALADIN – LAEF (Aire Limitée Adaptation dynamique Développement InterNational model – Limited-Area Ensemble Forecasting) ensemble forecasts, is provided in the second part of this research.

The aim of this work is to test the potential improvement of the NWP ALADIN-LAEF ensemble forecasts for the 10-m wind speed (***LAEFws***) while maintaining low computational costs for the analog search. For that reason, several experiments using different forecast information of the Austrian ALADIN-LAEF ensemble as input to the analog method are thoroughly analyzed. First, the sensitivity tests are performed to determine the optimal influence a certain meteorological parameter used as a predictor should have in the analog search procedure. The results show that the wind direction is the most important predictor in addition to wind speed, followed by temperature and relative humidity parameters, especially in the more complex topography. Using an NWP ensemble enables the use of more meteorological variables (predictors) in more than one realization as input to the analog search. In addition, using summarized information such as the ensemble mean and/or the standard deviation of the ensemble or every single ensemble member can provide useful insights. If the standard deviation of the ensemble is used as a predictor, its optimal

contribution is about 40% of the ensemble mean predictors' contribution in the majority locations tested.

In total, six analog-based AnEn experiments are conducted using a different set of input information from the ALADIN-LAEF model as predictors to the analog-based method. The choice of predictors from raw NWP model includes:

- The ensemble control member of all available parameters (***AnEnCtrl***)
- All wind speed raw forecast ensemble members (***AnEnWs***)
- The ensemble mean of all available parameters (***AnEnMu***)
- The ensemble mean and spread of all parameters (***AnEnStd***)
- All ensemble members of all parameters (***AnEnAll***)
- All available parameters corresponding to only one (distinguishable) ensemble member (***AnEnMem***),

where the abbreviations for analog experiments are listed in the brackets.

All experiments provide the 17 members wind speed analog ensemble forecast. To better understand the impact on the raw forecasts, the two experiments using the ensemble model output statistic post-processing approach (***EMOS***) are used as a baseline. The ***EMOSws*** only uses the last 30 days as training and only the wind speed as an input, whereas the ***EMOSstd*** uses all available training data and all variables including seasonal functions. The ***EMOSws*** is slightly more successful in removing the systematic, while the ***EMOSstd*** the dispersion source of the error.

Results show that all AnEn experiments substantially improve the raw model forecast. However, the most computationally demanding "member by member" ***AnEnMem*** experiment proved to be the least successful. The undesirable properties of the raw model, such as under-dispersion and lower resolution, are inherited more easily for this than for the other analog experiments. That is probably due to the fact that the analog-search pool is smaller than when seeking among all members independently, as it is the case in the other analog experiments. Using only one predictor variable as input (the 17 members of ***LAEFws***) already improves the forecast skills and lowers the systematic error of the ensemble mean, better or comparable to the ***AnEnMem*** experiment. If the number of available parameters from the raw model is limited, the experiment using only wind speed ensemble members (***AnEnWs***) proved to be successful. Even better results are achieved when using more than one predictor variable. Therefore, similar or better results are achieved when using only the ensemble control

member as input (***AnEnCtrl***). In addition, using more than one ensemble member within the analog search procedure improves results even more. The results confirm the hypothesis that post-processing methods have a large potential to improve the raw ensemble forecast. Moreover, it is shown that often there is no need to use the full input spectrum of a raw probabilistic model, i.e. all ALADIN-LAEF members as predictors. Using basic information of an input ensemble, such as ensemble mean and ensemble standard deviation, improves the forecast skills almost as successful as using the full input spectrum of a raw probabilistic model as predictors, with very little significant differences, if any. Furthermore, it is computationally less demanding. This result confirms the additional hypothesis that the summary metric (e.g. mean and standard deviation) is the optimal way to add the aspects of error growth, that can be represented dynamically by the input ensemble model, to the flow-dependent error growth already captured by the analog approach. Therefore, it can be suggested as the most promising configuration among experiments tested in this work.

All post-processing experiments in this work provide better results than the raw input model, as expected, reducing the under-dispersion while increasing the reliability and discrimination. The best results for both the analog approach and the ***EMOS*** are achieved in July when the raw model performs better. The raw model under-spread is almost completely removed by all experiments. The ***EMOSws*** approach is slightly under-dispersive, especially in January, probably due to using only wind speed parameters and much shorter training than other post-processing experiments.

The accuracy of the ensemble forecast is measured by the *RMSE* for the ensemble mean and the continuous rank probability score (*CRPS*). The analog-based experiments outperform the raw ***LAEFws*** forecast in terms of significantly better accuracy for all forecast lead-times during both (winter and summer) months. They are more skillful during nighttime than during daytime. The analog-based method is comparable to or outperforms both ***EMOS*** experiments. The outperformance is noticed at short lead-times and during the winter month, especially in terms of correlation. The ***EMOSws*** is overconfident to a certain extent for the high probability forecasts, while ***EMOSstd*** is underconfident for low probability forecasts. The analog-based experiments are almost perfectly reliable. Additionally, discrimination is slightly better than the ***EMOS*** due to a higher hit rate. The difference among the analog experiments is less pronounced than when compared to the ***LAEFws*** and the ***EMOS*** experiments, confirming

that using basic information of an input ensemble, such as ensemble mean and standard deviation is often sufficient.

If considered spatially, the ***LAEFws*** error follows the climatological wind speed pattern, having higher values at the stations prone to higher winds. Also, the ***LAEFws*** error is more pronounced in the alpine complex topography than in the eastern plains. The accuracy of post-processing methods is improved when compared to the raw model forecast, following a similar pattern. Additionally, even though an improvement over the raw model forecast is evident, large differences among nearby stations are noticed in highly complex topography. The valley stations wind speed is better predicted by the raw model, and post-processing result is, therefore, overall better at the valley stations than at the mountain stations with the climatologically higher wind speeds. On the other hand, the relative improvement to the raw model is more pronounced at mountain stations due to the reduction of systematic sources of error by post-processing. These sources of error are less present in the raw model for the valley stations.

It is very important to assess the post-processing performance for high wind speed because of the impact on people and property. For that reason, several thresholds ranging from 0.5 ms^{-1} to 20 ms^{-1}, are used to test the skill of the post-processed forecasts. The result shows that the ***LAEFws*** forecast is skillful in reproducing wind speeds of the order of 3 ms^{-1} thresholds, but the same can not be concluded at higher or lower thresholds. The analog experiments are able to improve the raw forecast, exhibiting significantly higher skill than the ***EMOS***, up to 10 ms^{-1} wind speed threshold. Furthermore, the ***AnEnStd*** and the ***AnEnAll*** experiments significantly improve the raw model results for all thresholds tested in January. However, neither of the post-processing methods tested is an adequate tool to reproduce the wind speeds exceeding 15 ms^{-1}. For that purpose, further modifications of the proposed methods, their combination, or the usage of the additional calibration method, such as quantile regression forests, should be investigated.

To summarize, even the simple experiment ***AnEnWs***, which uses only one meteorological parameter (wind speed) as a predictor variable, significantly increases correlation with the measurements and decreases the error. Using more meteorological parameters as predictor variables improves the results even further, leading to substantial improvements in terms of correlation, reliability, spread-skill ratio and error reduction (measured by *RMSE* and *CRPS*). We confirmed our primary hypothesis that the analog

experiments can remarkably improve the raw ***LAEFws*** forecast. Overall, the experiments prove to be at least as successful as the EMOS post-processing approach, or even more.

It is shown that the optimal weight of the different predictor variables in the analog search procedure is location-dependent and every meteorological parameter tested is beneficial at least at certain areas, and, hence, should not be neglected. Furthermore, we demonstrate the importance of including the information on the raw ensemble uncertainty into analog search procedure in contrast to using only one raw ensemble member (***AnEnCtrl***) or the mean of the raw ensemble (***AnEnMu***). Since the two the most successful analog experiments, ***AnEnAll*** and ***AnEnStd,*** rarely differ significantly, we have proven the additional hypothesis that the summary metric is the optimal way to include the aspect of the error growth, that can be represented dynamically by the raw model.

The error reduction by the analog-based method is notable regardless of the topographic features due to (but not limited to) systematic error reduction. After demonstrating the applicability of the analog-based method in the coastal complex topography, the performance is hereby confirmed even for mountain complex topography tested for the alpine region. This makes the analog-based method a perfect candidate for the implementation in the Croatian Meteorological and Hydrological Service operational suite. Additionally, the importance of a predictor weighting strategy for a successful implementation is also highlighted. However, the post-processing methods tested in this thesis are not an adequate tool to reproduce the extremely high wind speeds (i.e. to issue warnings) in the proposed configuration. For that purpose, further modifications of the proposed methods or even the additional correction or calibration are advised.

§ 6. CONCLUSION

The performance of the analog-based post-processing method is tested in climatologically and topographically different regions in Croatia and Austria, for point-based wind speed predictions at 10 m above the ground. The target area is coastal-complex topography in Croatia. First, the analog-based method is applied to the deterministic numerical weather predictions (NWPs). The performed verification shows that all analyzed post-processing methods improve upon the starting model forecasts. The level of improvement depends on the type of topography, starting model and verification metric. Each tested post-processing method has its strengths and weaknesses and the choice for operational use of those methods depends on the envisaged purpose. The results are presented in such a manner that after a simple statistical analysis of the potential starting model, one can thus decide which post-processing method is the most applicable for a specific situation.

The forecasting using the mean of the analog ensemble exhibits the highest correlation with measurements. It is thus the most applicable if the model is unbiased, but there is a need to reduce the dispersion error. The applications of the Kalman filter directly on the NWP forecast (KF) or on the AN forecast (KFAN) are the most successful in removing bias, whereas the KFAN is better suited if the topography is more complex. The analog-based method exhibits better result than the Kalman filter approach in the complex topography in general, especially coastal area. If the focus is on the prediction of the weak wind, then the AN is the most suitable, whereas for somewhat higher wind speed the analog approach is better suited when combined with the KF (i.e. applying the Kalman filter to the sorted analogs - KFAS). The Kalman filter algorithm affects only the (very) large scale flows: enlarges the energy of these large-scale motions in the coastal area and reduces the energy at the nearly flat continental topography for the periods longer than 10 days. On the other hand, the analog-based method affects smaller scales. If the starting model power spectral density is biased, the KFAS method is superior to the other approaches. The superior adaptability of the KFAS results in better adaptability of the shorter than diurnal motions.

Additionally, results of the post-processing methods are further improved if higher-resolution (convection-permitting) starting model data are used in the coastal complex topography. Introducing the higher-resolution modeling in nearly flat continental topography results with very similar power spectral density curves. The experiments exhibit at least as

good results when using the coarser horizontal resolution, if not better. Therefore, the need for using 2-km as opposed to the 8-km grid spacing model may not be necessary. On the other hand, the higher-resolution modeling increased the energy available for all of the time scales in the mountain complex topography. The latter, however, yielded mixed results when using the other verification metrics for both the starting models and corresponding post-processing results. In this case, the results may be improved by using even finer model resolution than 2-km to represent local flows.

Due to the limited availability of measurements in the mountain complex topography, the second part of this research is performed using Austrian sites. After a thorough analysis of the analog-based method application to the deterministic NWP, the focus is now shifted to the application to the ensemble NWP. Naturally, the verification procedure in this part includes the scores suitable for the ensemble and probabilistic forecasting (i.e. Brier skill score, continuous rank probability score, spread-skill diagram, rank histograms) to analyze the most important aspects such as reliability, sharpness, discrimination, spread-skill ratio and statistical consistency.

Substantial improvements of raw model wind speed forecast are demonstrated in terms of correlation, reliability, spread-skill ratio and error reduction (measured by RMSE and CRPS). The benefits of using even the simple analog-based method implementation that uses only wind speed as a predictor variable are significant, and using more meteorological parameters as predictor variables further improves the results. Overall, the experiments are proved to be as successful as the ensemble model output statistic (EMOS) post-processing approach or better.

We demonstrate the importance of a predictor weighting strategy and also including the summarized information on the raw ensemble uncertainty into the analog search procedure in contrast to using only one raw ensemble member or the mean of the raw ensemble, but not necessarily the full input spectrum of a raw probabilistic model.

The error reduction is first demonstrated for coastal complex topography in Croatia and then confirmed even for mountain complex topography in the alpine region. encouraging the implementation of the analog-based method in the operational suite of the Croatian Meteorological and Hydrological Service. Finally, several possible future research avenues are proposed as a continuation of this research, such as investigating the implementation for forecast fields or extremely high wind speed forecasts.

§ 7. LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
A2	- operational limited-area mesoscale ALADIN model at 2-km horizontal grid spacing
A8	- operational limited-area mesoscale ALADIN model at 8-km horizontal grid spacing
ALADIN	- Aire Limitée Adaptation sdynamique Développement InterNational model
ALADIN-LAEF	- Aire Limitée Adaptation dynamique Développement InterNational model – Limited-Area Ensemble Forecasting
AN	- analog ensemble mean forecast
AnEn	- analog ensemble
AnEnAll	- analog-based experiment that uses all ensemble members of all meteorological parameters as an input
AnEnCtrl	- analog-based experiment that uses the ensemble control member of all available meteorological parameters from raw model as an input
AnEnMem	- analog-based experiment that uses all available meteorological parameters corresponding to only one (distinguishable) raw model ensemble member
AnEnMu	- analog-based experiment that uses the raw model ensemble mean of all available meteorological parameters as an input
AnEnStd	- analog-based experiment that uses the raw model ensemble mean and spread of all meteorological parameters as an input
AnEnWs	- analog-based experiment that uses all wind speed raw forecast ensemble members as an input
ANKF	- equivalent to KFAS forecast
ANM	- analog ensemble median forecast
ARPEGE	- Action de Recherche Petite Echelle Grande Echelle global model
BS	- Brier score

ABBREVIATION	DESCRIPTION
BSS	- Brier Skill Score
CRPS	- continuous rank probability score
CSI	- critical success index; equvalent to threat score
DA	- operational ALADIN high-resolution dynamical adaptation model
dd	- ALADIN-LAEF wind direction prediction
DIU	- diurnal (motions)
ECMWF	- European Centre for Medium-Range Weather Forecasts
EDI	- extremal dependence index
EMOS	- ensemble model output statistic
EMOSstd	- EMOS experiment that uses all available training data and all variables including seasonal functions
EMOSws	- EMOS experiment only using the last 30 days as training and only the wind speed as an input
EPS	- Ensemble Prediction System
F	- forecasts
F	- false alarm rate
FBias	- frequency bias
H	- hit rate
ISBA	- Interaction Soil Biosphere Atmosphere
KF	- Kalman filter; Kalman filter forecast (applied to starting model time series)
KFAN	- Kalman filter of the analog ensemble mean prediction
KFAS	- Kalman filter in analog space prediction
LAEFws	- ALADIN-LAEF ensemble wind speed predictions
LAM	- limited-area model
LTD	- longer than diurnal (motions)
N	- number of ensemble members
NOAA	- National Oceanic and Atmospheric Administration
NWP	- numerical weather prediction
O	- observations

ABBREVIATION	DESCRIPTION
p	- ALADIN-LAEF pressure prediction
PCC	- polychoric correlation coefficient
prec	- ALADIN-LAEF precipitation prediction
PSD	- power spectral density
RCC	- rank correlation coefficient
REL	- reliability term in the Brier score decomposition
RES	- resolution term in the Brier score decomposition
rH	- ALADIN-LAEF relative humidity prediction
RMSE	- root-mean-square error
ROC	- relative operating characteristic
SOI	- Southern Oscillation Index
STD	- shorter than diurnal (motions)
t2m	- ALADIN-LAEF temperature (2m) prediction
UNC	- uncertainty term in the Brier score decomposition
UTC	- coordinated universal time
WMO	- World Meteorological Organization
ws	- ALADIN-LAEF wind speed prediction
WSPD	- wind speed prediction
μ	- ensemble mean
σ	- standard deviation

.

§ 8. REFERENCES

- ALADIN International Team, 1997: The ALADIN project: Mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research. WMO Bulletin, 46, 317-324.

- Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015a: An analog ensemble for short-term probabilistic solar power forecast. Applied Energy, 157, 95-110.

- Alessandrini, S., Delle Monache, L., Sperati, S., Nissen, J, 2015b: ensemble novel application of an analog ensemble for short-term wind power forecasting. Renewable Energy, 76, 768-781.

- Alessandrini, S., Sperati, S., Delle Monache, L., 2019: Improving the Analog Ensemble Wind Speed Forecasts for Rare Events. Monthly Weather Review, 147, 2677–2692. doi:10.1175/MWR-D-19-0006.1

- Anderson, J.L, 1997: The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. Monthly Weather Review, 125, 2969-2983.

- Arnold, D., Morton, D., Schicker, I., Seibert, P., Rotach, M. W., Horvath, K., Dudhia, T., Satomura, T., Muller, M., Zangl, G., Takemi, T., Serafin, S., Schmidli, J., Schneider, S., 2012: Issues in high-resolution modeling in complex topography - The HiRCoT workshop, Croatian Meteorological Journal, 47, 1-12.

- Bajic, A., Ivatek-Sahdan, S., Horvath, K., 2007: Spatial distribution of wind speed in Croatia obtained using the ALADIN model. Croatian Meteorological Journal, 42, 67-77.

- Bajic, A., Ivatek-Sahdan, S., Zibrat, Z., 2008: ANEMO-ALARM iskustva operativne primjene prognoze smjera i brzine vjetra. GIU Hrvatski cestar. 109-114.

- Baran, S., Lerch, S., 2015: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind speed forecasting. Quarterly Journal of the Royal Meteorological Society, 141: 2289-2299. doi:10.1002/qj.2521.

- Belluš, M., Wang, Y., Meier, F., 2016: Perturbing surface initial conditions in a regional ensemble prediction system. Monthly Weather Review, 144, 3377–3390.

- Bougeault, P., P. Binder, Buzzi, A., Dirks, R., Houze, R., Kuettner, J., Smith, R. B., Steinacker, R., Volkert, H., 2001: The MAP Special Observing Period. Bulletin of the American Meteorological Society (BAMS), 82 (3), 433-462.

- Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M., Zhu, Y., 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction system. Weather and Forecasting, 14, 168-189.

- Burgers, G., van Leeuwen, P. J., Evensen, G., 1998: Analysis scheme in the ensemble Kalman filter. Monthly Weather Review, 126, 1719 – 1724.

- Catry, B., Geleyn, J.-F., Tudor, M., Benard, P., Trojakova, A., 2007: Flux conservative thermodynamic equations in a mass-weighted framework. Tellus 59A, 71-79.

- Dabernig, M., Mayr, G.J., Messner, J.W., 2015: Predicting Wind Power with Reforecasts. Weather Forecasting, 30, 1655–1662, doi:10.1175/WAF-D-15-0095.1.

- Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., Stull, R. B., 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias-correction. Journal of Geophysical Research 111, D05308.

- Delle Monache, L., Wilczak, J., McKeen, S., Grell, G., Pagowski, M., Peckham, S., Stull, R., McHenry, J., McQueen, J., 2008: A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. Tellus B, 60, 238-249.

- Delle Monache, L., Nipen, T., Liu, Y., Roux, G., Stull, R., 2011: Kalman filter and analog schemes to post-process numerical weather predictions. Monthly Weather Review, 139, 3554-3570.

- Delle Monache, L., Eckel, T., Rife, D., Nagarajan, B., 2013: Probabilistic weather prediction with an analog ensemble. Monthly Weather Review 141, 3498-3516.

- Djalalova, I., Delle Monache, L., Wilczak, J., 2015: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. Atmospheric Environment 108, 76-87.

- Drosdowsky, W., 1994: Analog (nonlinear) forecasts of the Southern Oscillation index time series. Weather Forecasting, 9, 78–84.

- Eckel, F.A., Delle Monache, L., 2016: A Hybrid NWP–Analog Ensemble. Monthly Weather Review, 144, 897–911.

- Ekström, Joakim, 2011: On the Relation Between the Polychoric Correlation Coefficient and Spearman's Rank Correlation Coefficient. Department of Statistics Papers, UCLA Department of Statistics, 15 pp.

- Esterle, G. R., 1992: Adaptive, self-learning statistical interpretation system for the central Asian region. Annales geophysicae, 10, 924–929.

- Ferro, C.A.T., Stephenson, D.B, 2011: Extremal Dependence Indices: improved verifiation measures for deterministic forecasts of rare binary events. Weather Forecasting, 26, 699-713.

- Gao, L., Ren, H., Li, J., Chou, J., 2006: Analogue correction method of errors and its application to numerical weather prediction. Chinese Physics, 15, 882–889.

- Geleyn, J.-F., Hollingsworth, A., 1979: An economical analytical method for the computation of the interaction between scattering and line absorption of radiation. Beitraege zur Physik der Atmosphaere, 52, 1-16.

- Geleyn, J.-F., 1987: Use of a modified Richardson number for parameterizing the effect of shallow convection. Matsuno Z. (ed)., Short and medium range weather prediction, Special volume of Journal of the Meteorological Society of Japan, 141-149.

- Geleyn J.-F., 1988: Interpolation of wind, temperature and humidity values from model levels to the height of measurement. Tellus, 40A, 347-351.

- Geleyn, J.-F., Bazile, E., Bougeault, P., Deque, M., Ivanovici, V., Joly, A., Labbe, L., Piedelievre, J.-P., Pirou, J.-M., Royer, J.-F., 1994: Atmospheric parametrizations schemes in Meteo-France's ARPEGE NWP model. ECMWF seminar proceedings on Parametrization of sub-grid scale physical processes, pp. 385-402.

- Geleyn, J.-F., Vána, F., Cedilnik, J., Tudor, M., Catry, B., 2006: An intermediate solution between diagnostic exchange coefficients and prognostic TKE methods for vertical turbulent transport, Research Activities in Atmospheric and Oceanic Modelling. J. Côté, Ed., World Meteorological Organization, 4.11–4.12.

- Geleyn, J.-F., Catry, B., Bouteloup, Y., Brožkova, R., 2008: A statistical approach for sedimentation inside a microphysical precipitation scheme. Tellus 60A, 649-662.

- Gerard, L., 2007: An integrated package for subgrid convection, clouds and precipitation compatible with the meso-gamma scales. Quarterly Journal of the Royal Meteorological Society, 133, 711-730.

- Gerard, L., J.-F. Geleyn, 2005: Evolution of a subgrid deep convection parametrization in a limited area model with increasing resolution. Quarterly Journal of the Royal Meteorological Society, 131, 2293-2312.

- Gerard, L., Piriou, J.-M., Brozkova, R., Geleyn, J.-F., Banciu, D., 2009: Cloud and Precipitation Parameterization in a Meso-Gamma-Scale Operational Weather Prediction Model. Monthly Weather Review, 137, 3960-3977.

- Giard, D., E. Bazile, 2000: Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. Monthly Weather Review, 128, 997-1015.

- Gneiting, T., Raftery, A.E., Westveld, A.H. Goldman, T., 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly Weather Review, 133, 1098–1118. doi:10.1175/MWR2904.1

- Grisogono, B., Belusic, D., 2009: A review of recent advances in understanding the meso- and micro-scale properties of the severe Bora wind. Tellus A, 61, 1, 1-16.

- Hall, P., Horowitz, J.L., Jing, B.-Y., 1995: On blocking rules for the bootstrap with dependent data. Biometrika, 82, 561–574.

- Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550-560.

- Hamill, T. M., Whitaker, J. S., 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Monthly Weather Review, 134, 3209–3229.

- Hamill, T. M., Whitaker, J. S., Mullen, S. L., 2006: Reforecasts: An important dataset for improving weather predictions. Bulletin of the American Meteorological Society (BAMS), 87, 33–46.

- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecasting, 15, 559-570.

- Homleid, M., 1995: Diurnal corrections of short-term surfacetemperature forecasts using Kalman filter. Weather Forecasting, 10, 689–707.

- Hopson, T. M., 2005: Operational flood-forecasting for Bangladesh. Ph.D. thesis, University of Colorado, 225 pp.

- Hopson, T. M., Webster, P. J., 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. Journal of Hydrometeorology, 11, 618–641.

- Horvath, K., Bajić, A., Ivatek-Šahdan, S., 2011: Dynamical downscaling of wind speed in complex terrain prone to bora-type flows. Journal of Applied Meteorology and Climatology, 50, 1676-1691.

- Horvath, K., Ivatek-Sahdan, S., Ivancan-Picek, B., Grubisic, V., 2009: Evolution and structure of two severe cyclonic Bora events: contrast between the northern and southern Adriatic. Weather Forecasting, 24, 946-964.

- Horvath, K., Koračin, D., Vellore, R., Jiang, J., Belu, R., 2012: Sub-kilometer dynamical downscaling of near-surface winds in complex terrain using WRF and MM5 mesoscale models. Journal of Geophysical Research, 117, D11111, 19 pp.

- Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., Hansen, B., 2005: Atmospheric data assimilation withan Ensemble Kalman Filter: Results with real observations. Monthly Weather Review, 133, 604 – 620.

- Hrastinski, M., Horvath, K., Odak Plenkovic, I., Ivatek-Sahdan, S., Bajic, A., 2015: Verification of the operational 10 m wind forecast obtained with the ALADIN mesoscale numerical weather prediction model. Hrvatski meteoroloski casopis, 50/50, 105-120.

- Ivatek-Sahdan, S. Tudor M. 2004: Use of high-resolution dynamical adaptation in operational suite and research impact studies. Meteorologische Zeitschrift, 13 (2), 99-108.

- Ivatek-Sahdan, S., Ivancan-Picek, B., 2006: Effects of different initial and boundary conditions in ALADIN/HR simulations during MAP IOPs. Meteorologische Zeitschrift, 15, 187-197.

- Ivatek-Šahdan, S., Stanešić, A., Tudor, M., Odak Plenković, I., Janeković, I., 2018: Impact of SST on heavy rainfall events on eastern Adriatic during SOP1 of HyMeX. Atmospheric Research, Volume 200, 36-59 (https://doi.org/10.1016/j.atmosres.2017.09.019.)

- Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. Weather Forecasting, 22, 637–650.

- Jolliffe I. T., Stephenson, D.B., 2011: Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, 274 pp.

- Junk, C., Delle Monache, L., Alessandrini, S., von Bremen, L., Cervone, G., 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. Meteorologische Zeitschrift, 24, 361–379.

- Juras, J., Pasaric, Z., 2006: Application of tetrachoric and polychoric correlation coefficients to forecast verification. Geofizika, 23, 59–81.

- Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82, 35-45.

- Keller, D. E., Fischer, A. M., Liniger, M. A., Appenzeller, C., Knutti, R., 2017: Testing a weather generator for downscaling climate change projections over Switzerland. International Journal of Climatology, 37, 928–942.

- Klausner, Z., Kaplan, H., Fattal, E., 2009: The similar days method for predicting near surface wind vectors. Meteorological Applications, 16, 569–579.

- Koopmans, L. H., 1974: The Spectral Analysis o f Time Series. Academic Press, New York, 382 pp.

- Kuettner, J. P., 1986: The aim and conduct of ALPEX (Alpine Experiment (ALPEX)). WMO Proceedings of the Conference on the Scientific Results of the Alpine Experiment (ALPEX), ICSU-WMO, GARP Publication Series, 27, 3-13.

- Lehner, M., Rotach, M.W., 2018: Current Challenges in Understanding and Predicting Transport and Exchange in the Atmosphere over Mountainous Terrain. Atmosphere 2018, 9, 276.

- Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences, 26, 636-646.

- Louis, J.-F., Tiedke, M., Geleyn, J.-F., 1982: A short history of PBL parameterization at ECMWF. Proceedings from ECMWF workshop on planetary boundary layer parameterization. 25-27 November 1981, pp. 59-79.

- Mass, C. F., Ovens, D., Westrick, K., Colle, B. A., 2002: Does increasing horizontal resolution produce more skillful forecast? Bulletin of the American Meteorological Society, 83, 407-430.

- Mayr, G. J., Armi, L. , Gohm, A. , Zängl, G. , Durran, D. R., Flamant, C. , Gaberšek, S. , Mobbs, S., Ross, A., Weissmann, M., 2007: Gap flows: Results from the Mesoscale Alpine Programme. Quarterly Journal of the Royal Meteorological Society, 133: 881-896. doi:10.1002/qj.66

- Mayr, G. J., Plavcan, D., Armi, L., Elvidge, A., Grisogono, B., Horvath, K., Jackson, P., Neururer A., Seibert P., Steenburgh, J. W., 2018: The Community Foehn Classification Experiment. Bulletin of the American Meteorological Society, 99, 11, 2229-2235. doi:10.1175/BAMS-D-17-0200.1.

- Messner, J.W., Mayr, G.J., Wilks, D.S., Zeileis, A., 2014: Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored. Monthly Weather Review, 142, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1

- Messner, J.W., Mayr, G.J., Zeileis, A., 2017: Nonhomogeneous Boosting for Predictor Selection in Ensemble Postprocessing. Monthly Weather Review, 145, 137–147, doi: 10.1175/MWR-D-16-0088.1.

- Mugume, I., Mesquita, M.D.S., Bamutaze, Y., Ntwali, D., Basalirwa, C., Waiswa, D., Reuder, J., Twinomuhangi, R., Tumwine, F., Jakob Ngailo, T., Ogwang, B.A., 2017: Improving Quantitative Rainfall Prediction Using Ensemble Analogues in the Tropics: Case study of Uganda. Preprints 2017. doi: 10.20944/preprints201710.0199.v1.

- Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient, Monthly Weather Review, 116, 2417–2424.

- Nagarajan, B., Delle Monache, L., Hacker, J., Rife, D., Searight, K., Knievel, J., Nipen, T., 2015: An evaluation of analog-based post-processing methods across several variables and forecast models. Weather and Forecasting, 30, 1623-1643.

- Noilhan, J., Planton, S., 1989: A simple parametrization of land surface processes for meteorological models. Monthly Weather Review, 117, 536-549.

- Odak Plenković, I., Delle Monache, L., Horvath, K., Hrastinski, M., 2018: Deterministic Wind Speed Predictions with Analog-Based Methods over Complex Topography. Journal of Applied Meteorology and Climatology, 57, 2047–2070. doi:10.1175/JAMC-D-17-0151.1.

- Odak Plenković, I, Schicker, I, Dabernig, M, Horvath, K, Keresturi, E., 2020: Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain. Q J R Meteorol Soc. 2020; 1-19 (https://doi.org/10.1002/qj.3769).

- Panziera, L., Germann, U., Gabella, M., Mandapaka, P. V., 2011: NORA—Nowcasting of orographic rainfall by means of analogues. Quarterly Journal of the Royal Meteorological Society, 137, 2106–2123.

- Papoulis, A., 1984: Probability, Random Variables, and Stochastic Processes. 4th ed. (2002), McGraw-Hill, New York, 852 pp.

- Poje, D., 1995: Bura (bora) and burin in Split. Croatian Meteorological Journal, 30(30), 1-19.

- Redelsperger, J. L., F. Mahé, P. Carlotti, 2001: A simple and general subgrid model suitable both for surface layer and free-stream turbulence, Boundary-Layer Meteorology, 101, 375–408, doi:10.1023/A:1019206001292.

- Ren, H., Chou, J., 2006: Analogue correction method of errors by combining statistical and dynamical methods. Acta Meteorologica Sinica, 20, 367–373.

- Ren, H., Chou, J., 2007: Strategy and methodology of dynamical analogue prediction. Science in China Series D: Earth Sciences, 50, 1589–1599.

- Rife, D. L., Davis, C. A., Liu, Y., 2004: Predictability of low-level winds by mesoscale meteorological models. Monthly Weather Review, 132, 2553–2569.

- Ritter B., Geleyn, J.-F., 1992: A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. Monthly Weather Review, 120, 303-325.

- Roeger, C., Stull, R. B., McClung, D., Hacker, J., Deng, X., Modzelewski, H., 2003: Verification of mesoscale numericalweather forecast in mountainous terrain for application toavalanche prediction. Weather Forecasting, 18, 1140–1160.

- Rossa, A., Nurmi, P., Ebert, E., 2008: Overview of methods for the verification of quantative precipitation forecasts. Precipitation: Advances in Measurement, Estimation and Prediction. Springer, 419-452.

- Rousteenoja, K., 1988: Factors affecting the occurrence and lifetime of 500 mb height analogues: A study based on a large amount of data. Monthly Weather Review, 116, 368-376.

- Scheuerer, M., Möller, D., 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. Annals of Applied Statistics, 9, no. 3, 1328--1349. doi:10.1214/15-AOAS843.

- Serafin, S., Adler, B., Cuxart, J., De Wekker, S.F.J., Gohm, A., Grisogono, B., Kalthoff, N., Kirshbaum, D.J., Rotach, M.W., Schmidli, J., Stiperski, I., Večenaj, Ž., Zardi, D., 2018: Exchange Processes in the Atmospheric Boundary Layer Over Mountainous Terrain. Atmosphere 2018, 9, 102.

- Sperati, S., Alessandrini, S. Delle Monache, L. 2017: Gridded probabilistic weather forecasts with an analog ensemble Quarterly Journal of the Royal Meteorological Society, 143: 2874-2885. doi:10.1002/qj.3137

- Stanesic, A., 2011: Assimilation system at DHMZ: development and first verification results. Croatian Meteorological Journal, 44/45, 3-17.

- Telišman Prtenjak, M., Grisogono, B., 2007: Sea/land breeze climatological characteristics along the northern Croatian Adriatic Coast. Theoretical and Applied Climatology, 90, 201–215.

- Teremonia, P., 2008: Scale-selective digital filter initialization. Monthly Weather Review, 136, 5246-5255.

- Thorarinsdottir, T. L., Gneiting, T., 2010: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173, 371-388, doi:10.1111/j.1467-985X.2009.00616.x.

- Tudor, M., Ivatek-Sahdan, S., 2002: The MAP-IOP 15 case study. Croatian Meteorological Journal 37, 1-14.

- Tudor, M., Ivatek-Sahdan, S., Stanesic, A., Horvath, K., Bajic, A., 2013: Forecasting weather in Croatia using ALADIN numerical weather prediction model. Climate Change and Regional/Local Responses, InTech, 247 pp., 59-88.

- Van den Dool, H. M., 1989: Anew look at weather forecast through analogs. Monthly Weather Review, 117, 2230–2247.

- Vanvyve, E., Delle Monache, L., Rife, D., Monaghan, A., Pinto, J., 2015: Wind resource estimates with an analog ensemble approach. Renewable Energy, 74, 761-773.

- Wang, Y., Haiden, T., Kann, A., 2006: The operational limited area modelling system at ZAMG: ALADIN-AUSTRIA. Österreichische Beiträge zu Meteorologie und Geophysik, 37, 1–33.

- Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F., Kann, A., Ivatek-Sahdan, S., Tian, W., Ma, X., Tascu, S., Bazile, E., 2011: The Central European limited-area ensemble forecasting system: ALADIN-LAEF. Quarterly Journal of the Royal Meteorological Society, 137(655): 483-502.

- Wang, Y., Belluš, M., Geleyn, J.F., Tian, W., Ma, X., Weidle, F., 2014: A new method for generating initial perturbations in regional ensemble prediction system: blending. Monthly Weather Review, 142, 2043–2059.

- Wang, Y, Belluš, M, Weidle, F. et al., 2019: Impact of land surface stochastic physics in ALADIN-LAEF. Quaterly Journal of Royal Meteorological Society, 145, 3333–3350.https://doi.org/10.1002/qj.3623

- Weidle, F., Wang, Y., Tian, W., Wang, T., 2013: Validation of strategies using clustering analysis of ECMWF EPS for initial perturbations in a limited area model ensemble prediction system. Atmosphere–Ocean, 51, 284–295.

- Welch, P., 1967: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics, 15/2, 70 - 73.

- Wilcox, R. R., 2009: Comparing Pearson Correlations: Dealing with Heteroscedasticity and Nonnormality. Com. in S-S and Comp., 38, 2220-2234.

- Wilks, D.S., 1997: Resampling hypothesis tests for autocorrelated fields. Journal of Climate, 10, 65–82.

- Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. 3rd ed., Academic Press, 676 pp.

- WMO, 2008: WMO Guide to Meteorological Instruments and Methods of Observation (Updated in 2010). WMO-No. 8, 716 pp.

- Wu, W. and Coauthors, 2012: Statistical downscaling of climate forecast system seasonal predictions for the southeastern Mediterranean. Atmospheric Research, 118, 346–356.

- Xavier, P. K., Goswami, B. N., 2007: An analog method for real-time forecasting of summer monsoon subseasonal variability. Monthly Weather Review, 135, 4149–4160.

- Zagar, M., Rakovec, J. 1999: Small-scale surface wind prediciton using dynamical adaptation. Tellus, 51A, 489-504.

- Zagar, N., Zagar, M., Cedilnik, J., Gregoric, G., Rakovec, J., 2006: Validation of mesoscale low-level winds obtained by dynamical downscaling of ERA40 over complex terrain. Tellus, 58A, 445-455.

- Zaninovic, K. and Coauthors, 2008: Klimatski Atlas Hrvatske 1961-1990 : 1971-2000 (Climate Atlas of Croatia 1961-1990 : 1971-2000). Meteorological and Hydrological Service, 200 pp.

- Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Hodge, B.-M., 2015: Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. Applied Energy, 156, 528-541.

.

§ 9. APPENDIX

9.1. Appendix A – spectral analysis

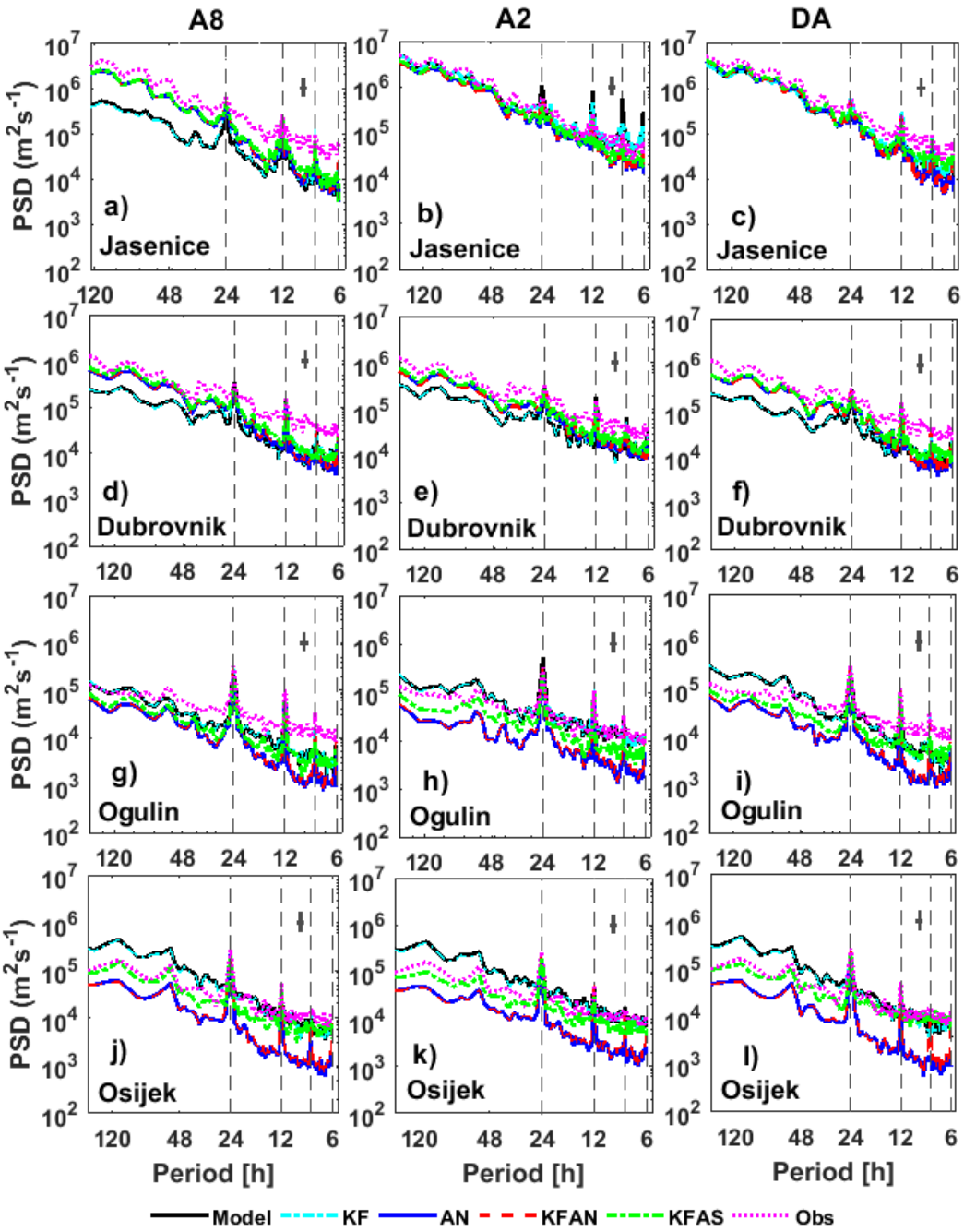

*Fig. I. The power spectral density of the observed 10-m wind speed, starting model forecasts (**A8**, **A2** and **DA**) and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**) for stations Jasenice, Dubrovnik, Ogulin and Osijek during year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

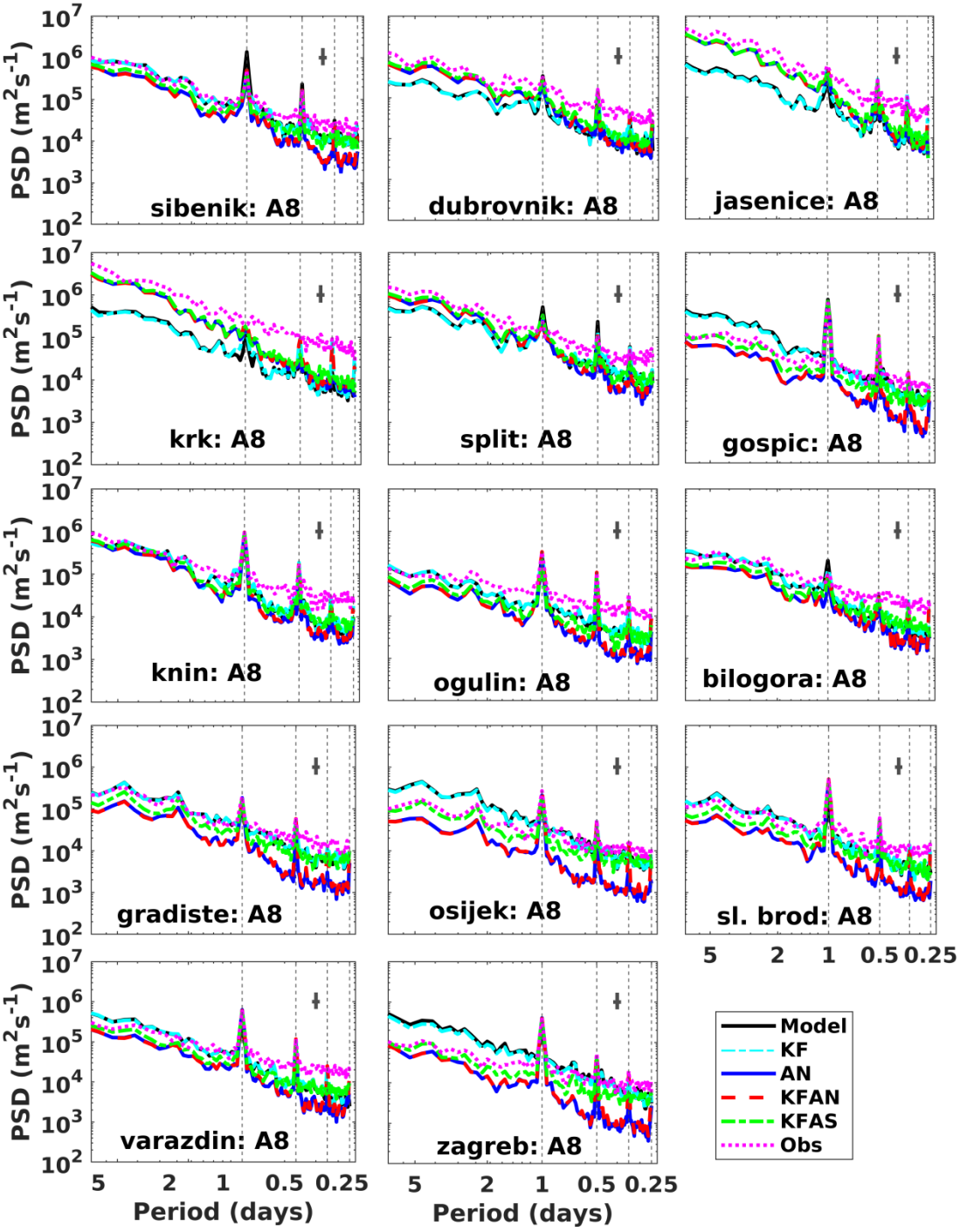

*Fig. II. The power spectral density of the observed 10-m wind speed, starting model forecasts **A8** and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**) for 14 stations in Croatia during year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

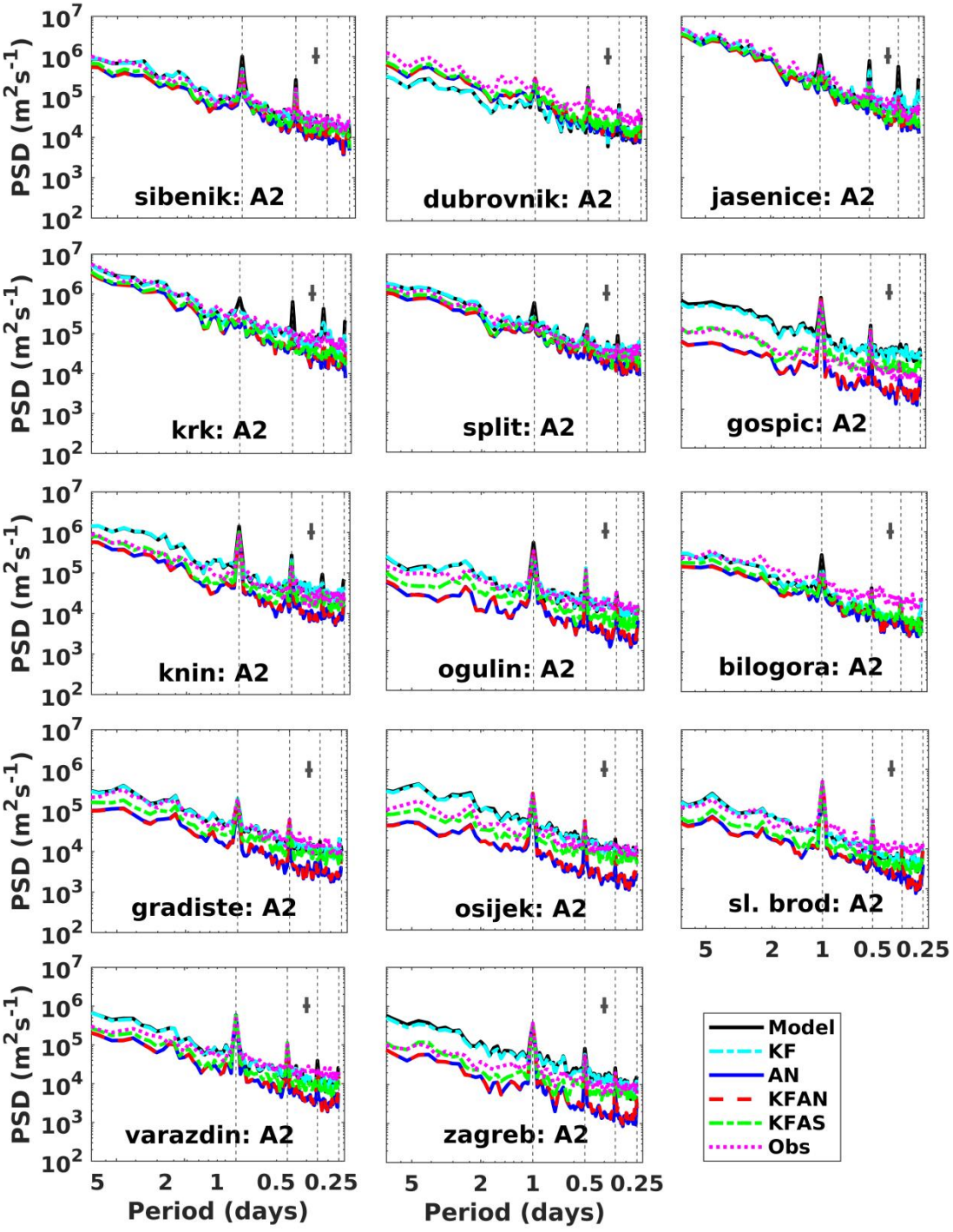

*Fig. III. The power spectral density of the observed 10-m wind speed, starting model forecasts **A2** and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**) for 14 stations in Croatia during year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

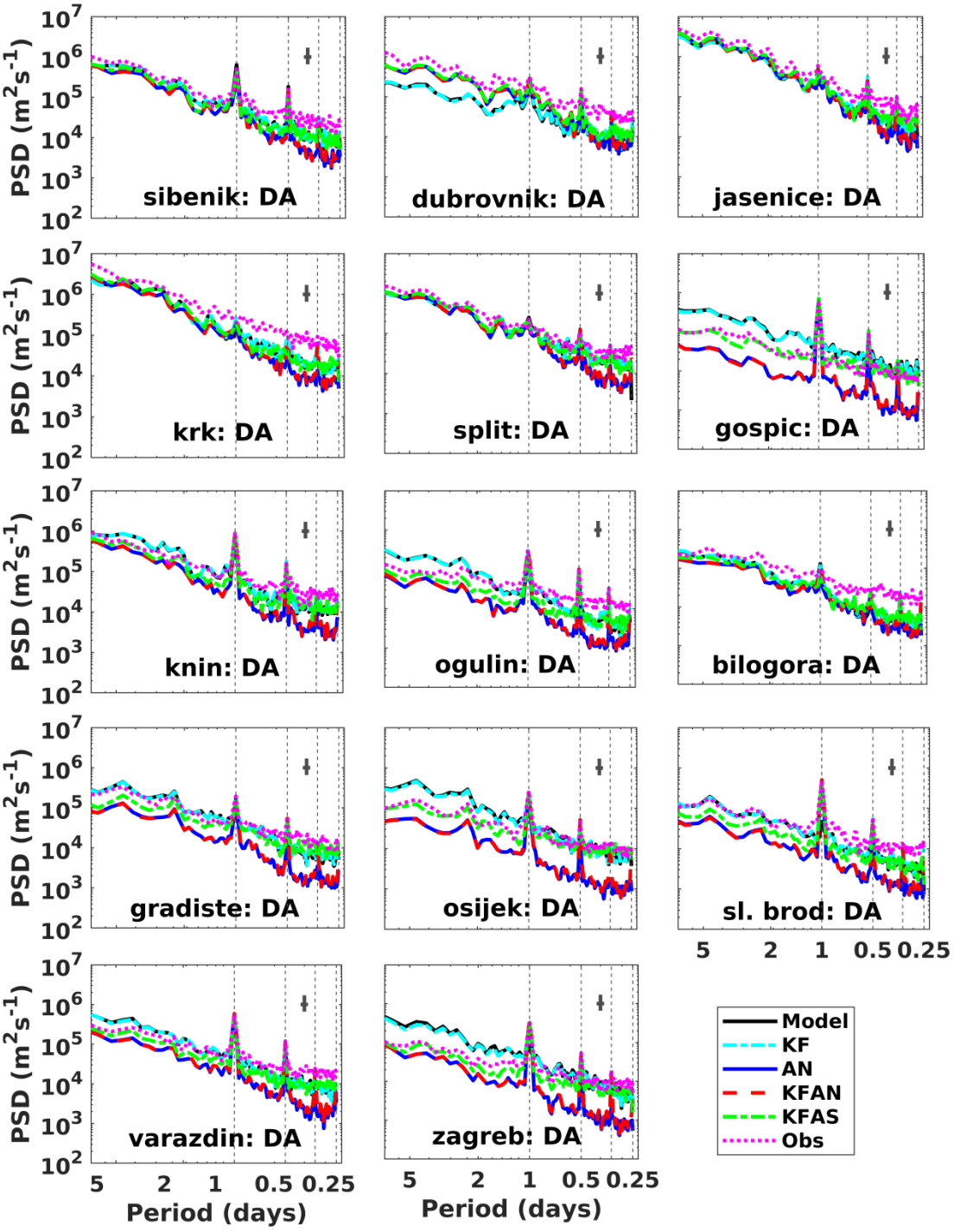

*Fig. IV. The power spectral density of the observed 10-m wind speed, starting model forecasts **DA** and the corresponding post-processing methods (**KF**, **AN**, **KFAN** and **KFAS**) for 14 stations in Croatia during year 2012. The confidence intervals (in the logarithmic scale) are noted by the cross-like symbol in the upper right corner.*

9.2. Appendix B – spatial performance

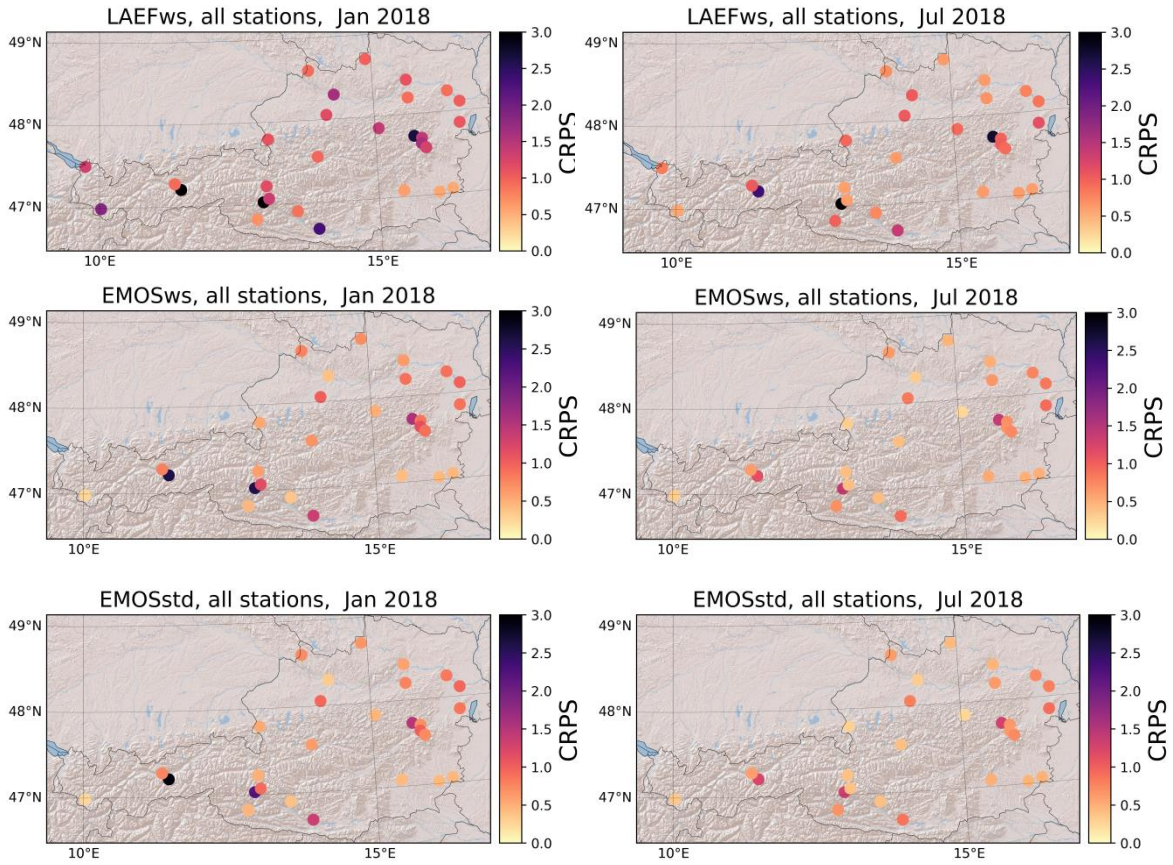

*Fig. V. The spatial distribution of the monthly mean continuous rank probability score for the raw **LAEFws, EMOSws** and **EMOSstd** forecasts for January (left) and July (right) 2018.*

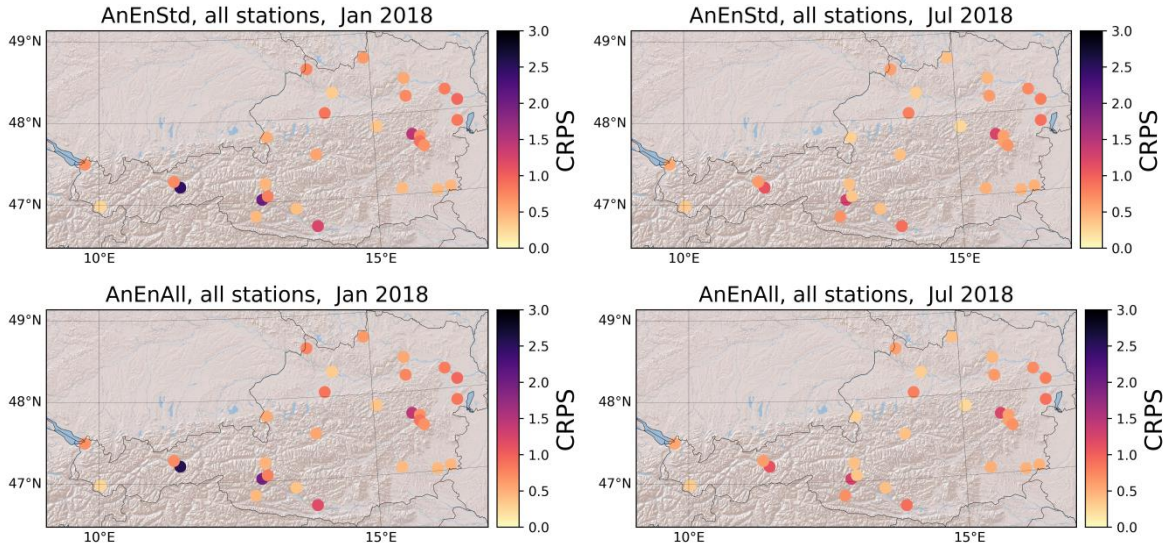

*Fig. VI. The spatial distribution of the monthly mean continuous rank probability score for the **AnEnStd** and **AnEnAll** forecasts for January (left) and July (right) 2018.*

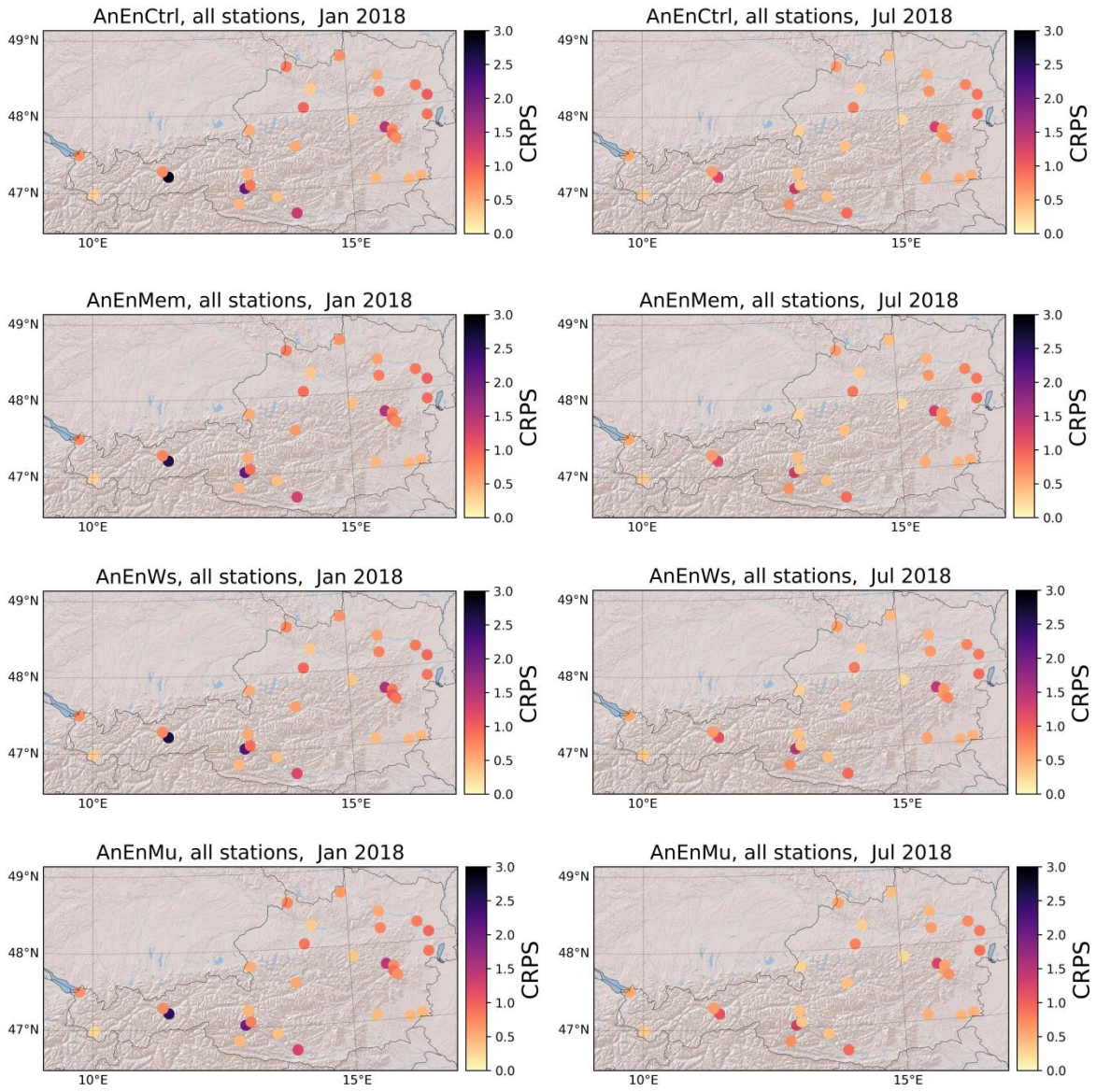

*Fig. VII. The spatial distribution of the monthly mean continuous rank probability score for the **AnEnCtrl, AnEnMem, AnEnWs** and **AnEnMu** forecasts for January (left) and July (right) 2018.*

9.3. Appendix C – special diagrams

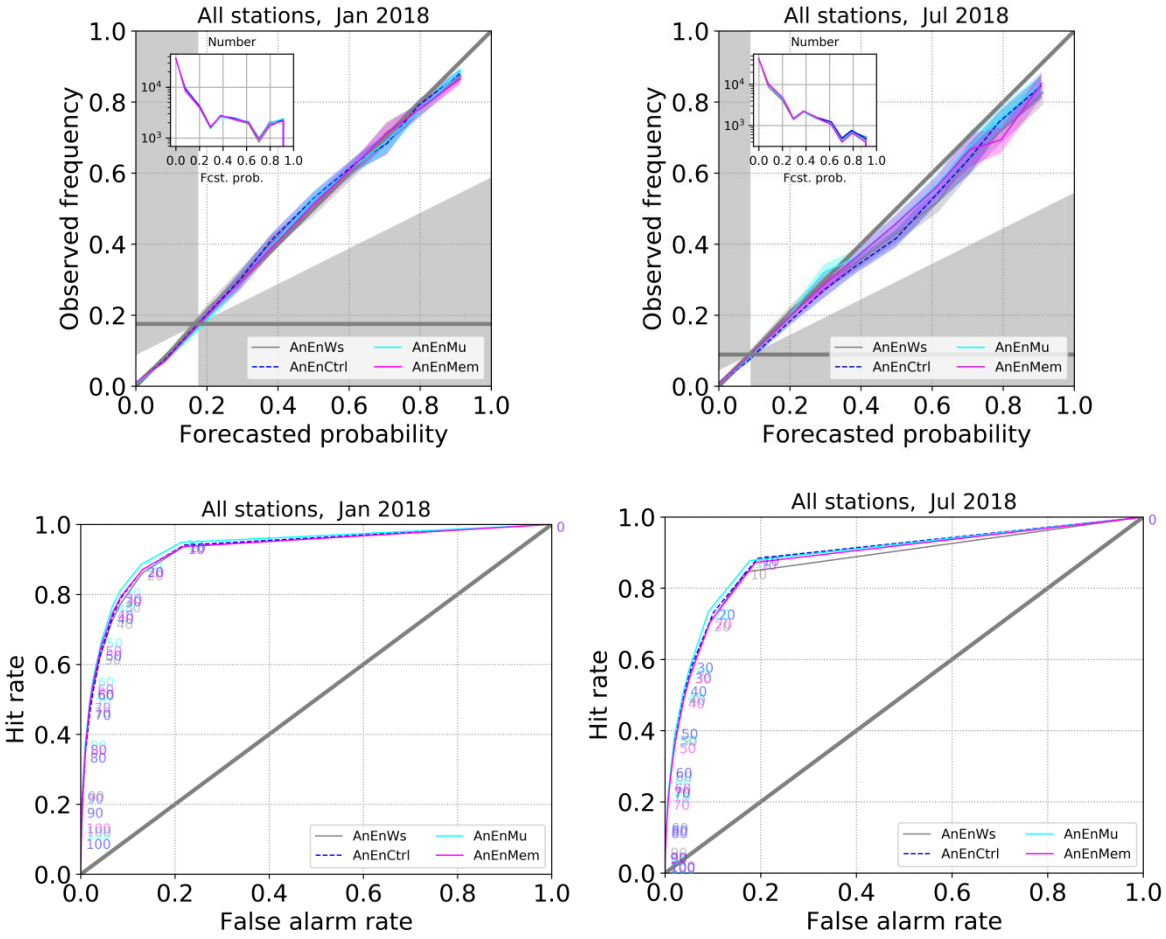

Fig. VIII. Reliability diagrams (top) and relative operating characteristic (ROC) diagrams (bottom) for four different analog forecasts and a threshold of > 5 ms^{-1} during January (left) and July (right) 2018 at 29 stations in Austria. The dashed lines in the reliability diagrams show a 95% confidence interval, while the sharpness diagrams are shown in the upper left corners

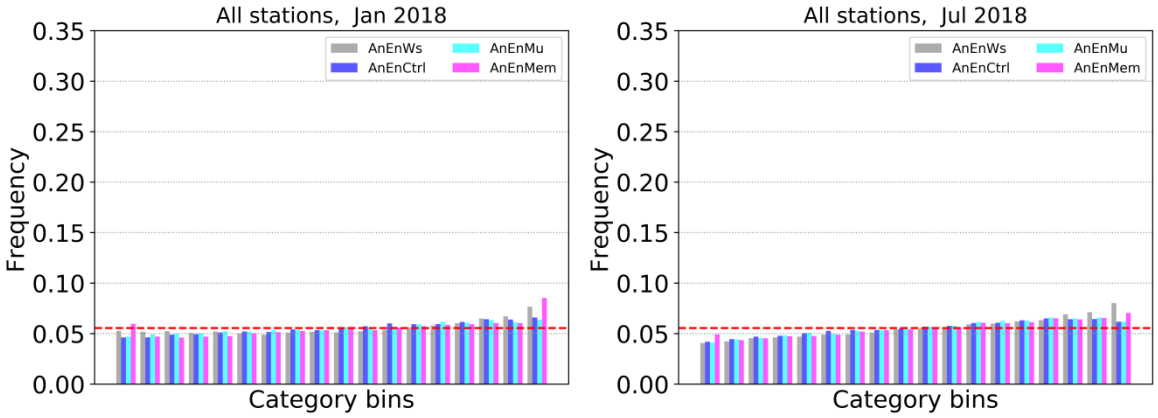

*Fig. IX. Rank histograms for the **AnEnWs, AnEnCtrl, AnEnMu** and **AnEnMem** forecasts during January (left) and July (right) 2018 at 29 stations in Austria.*

9.4. Appendix D – high wind speed predictions

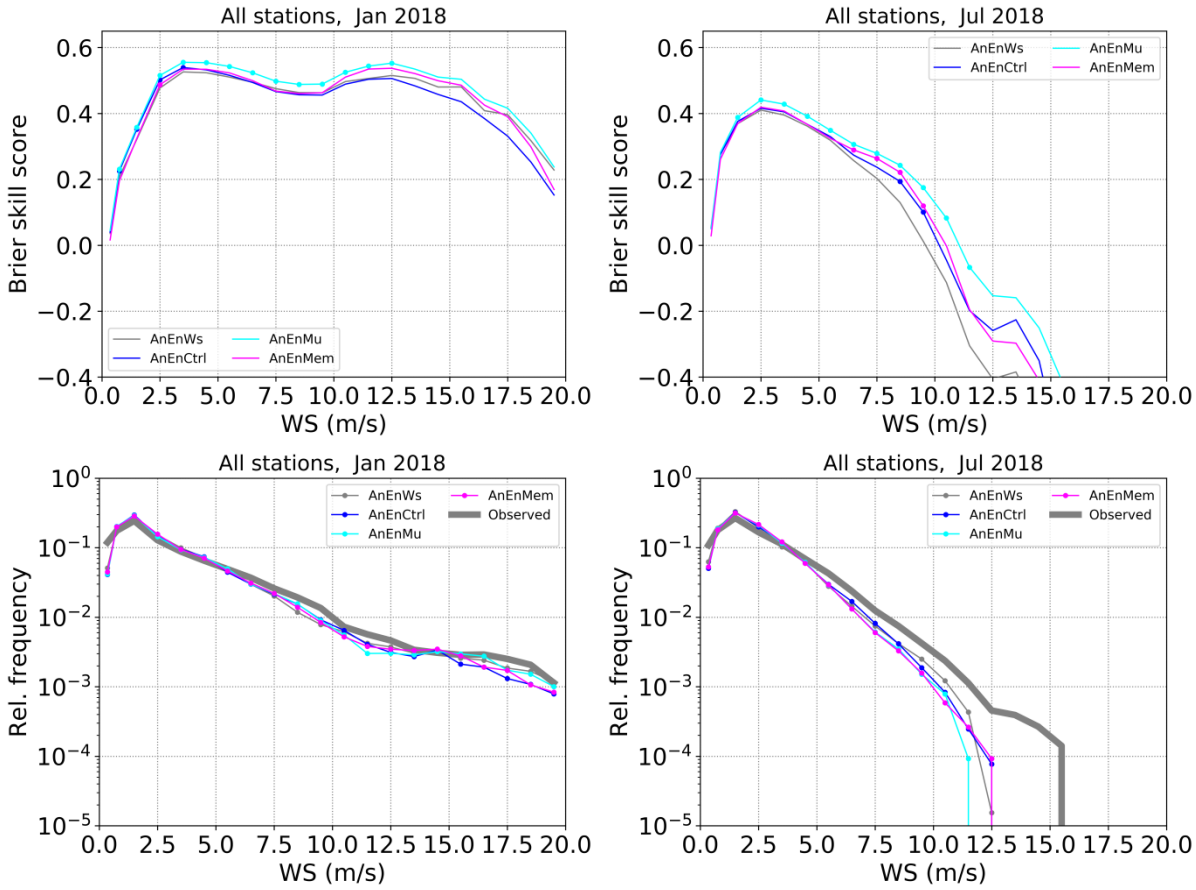

*Fig. X. Brier skill score (top) and relative frequency (bottom) depending on a wind speed threshold. The analog probabilistic forecasts shown for January (left) and July (right) 2018 at 29 stations in Austria. The markers are set for the BSS results significantly different from the **AnEnWs** forecast (95 % confidence level).*