Emulacija variacijske asimilacije meritev pri napovedovanju vremena z variacijskim avtokodirnikom

Boštjan Melinc*, Žiga Zaplotnik**

Povzetek

Variacijska asimilacija je pogosto uporabljena metoda za določitev začetnih pogojev za numerično napoved vremena. Metoda določa cenilko, ki upošteva prejšnjo vremensko napoved in nedavne meritve ozračja, upoštevajoč njihove napake, z minimizacijo te cenilke pa dobimo optimalni začetni pogoj, analizo. V članku predstavimo predelavo tradicionalne tridimenzionalne variacijske asimilacije (3D-Var), tako da se minimizacija njene cenilke izvede v reduciranem latentnem prostoru variacijskega avtokodirnika (VAE) namesto v polnem fizičnem prostoru modelskih točk. Za primer preproste persistenčne napovedi v latentnem prostoru demonstriramo, da je njej pripadajoča kovariančna matrika napake ozadja kvazidiagonalna. Z eksperimenti asimilacije posameznih psevdomeritev pokažemo, da ista kovariančna matrika napake ozadja za naš asimilacijski postopek ustrezno opiše tako ravnovesja v tropih, kot v zmernih geografskih širinah, poleg tega tudi ustrezno prenese vpliv merjene fizikalne količine na njej sklopljeno fizikalno količino. Ocenjena napaka ozadja v fizičnem prostoru je odvisna predvsem od letnega časa, nekoliko pa tudi od trenutnega stanja atmosfere. Z boljšim prognostičnim modelom bi lahko svojo metodo razširili na 4D-Var.

Ključne besede: asimilacija meritev, nevronske mreže, variacijski avtokodirnik, kovariančna matrika napake ozadja, 3D-Var, inkrement analize

Keywords: data assimilation, neural networks, variational autoencoder, background-error covariance matrix, 3D-Var, analysis increment

Uvod

Od prvega zapisa diferencialnih enačb za dinamiko atmosfere (Bjerknes, 1904) in uspešne vremenske napovedi z njihovo integracijo (Charney in sod., 1950) temelji operativno napovedovanje vremena na numeričnem reševanju fizikalno pogojenih parcialnih diferencialnih enačb. Dueben in Bauer (2018) sta pred vsega petimi leti kot prva omenila možnost napovedovanja vremena z nevronskimi mrežami (NN), ki nimajo nobene eksplicitno podane informacije o fiziki atmosfere, temveč njihova napoved temelji zgolj na učenju preslikav med polji atmosferskih spremenljivk, ki so nekaj ur (npr. 3, 6, 24 ur) narazen. Vremensko napoved torej dobimo avtoregresivno s serijo takšnih preslikav. Zaradi dostopnosti kvalitetnih podatkov (npr. reanaliz, tj. rekonstroukcij vremena za nazaj s sodobnimi numeričnimi modeli) za učenje tovrstnih modelov je sledil ekstremno hiter razvoj tovrstnih modelov, ki so v vsega štirih letih dosegli enak napredek v kvaliteti svoje napovedi kot klasični numerični modeli v več desetletjih, s koncem leta 2022 pa je bil objavljen prvi model (Bi in sod., 2023), ki je po nekaterih objektivnih metrikah (koren povprečne napake in anomalni korelacijski koeficient za številne fizikalne količine) dosegel najboljše sodobne deterministične numerične modele.

*Univerza v Ljubljani, Fakulteta za matematiko in fiziko, Jadranska ul. 12, Ljubljana
**Evropski center za srednjeročne vremenske napovedi (ECMWF)

Največja težava trenutnih napovednih modelov z NN je njihovo glajenje izhodnih polj, s čimer se izgubi njihova uporabnost za operativno napovedovanje vremena. Modeli nevronskih mrež tako na primer dobro napovedo pot tropskega ciklona, vendar močno podcenjujejo njegovo globino in posledično vetrove v njegovem središču (Bouallegue in sod., 2023). Prav tako podcenjujejo divergentno komponento horizontalnega vetra, ki v ozračju spodbuja vertikalna gibanja in padavine. S časom so tako modelske padavine vedno šibkejše (Bonavita, 2023). Kljub vsemu klasični dinamični eksperimenti prilagajanja ozračja na umetno vnesene motnje, torej eksperimenti, ki se v učni množici ne nahajajo, jasno pokažejo, da se modeli nevronskih mrež naučijo realistične fizike (Hakim in Masanam, 2023).

Druga težava trenutnih napovednih modelov z NN je v samostojni pripravi začetnih pogojev za izračun napovedi - te namreč še vedno jemljejo iz operativnih vremenskih centrov. Slednji izračunajo najboljši približek trenutnega stanja atmosfere, ki služi kot začetni pogoj za naslednjo napoved, z *asimilacijo meritev*, v kateri statistično objektivno združijo predhodno kratkoročno napoved stanja atmosfere z novimi meritvami (Kalnay, 2003; Lahoz in Schneider, 2014). Kljub temu da so meritve običajno precej natančne, same po sebi ne zadoščajo za pripravo začetnega pogoja, saj jih je premalo, poleg tega so v prostoru neenakomerno razporejene in nekaterih atmosferskih spremenljivk sploh ne moremo pomeriti neposredno.

Variacijska asimilacija je ena izmed najpogostejših metod za asimilacijo meritev v operativnih centrih (uporabljajo jo npr. na Evropskem centru za srednjeročne vremenske napovedi (ECMWF, 2023)). Njena tridimenzionalna verzija (3D-Var) temelji na izračunu najverjetnejšega stanja atmosfere preko metode največjega verjetja (angl. *maximum likelihood approach*) ob predpostavki, da so predhodna kratkoročna napoved (zanjo bomo uporabili izraz *ozadje*) in nove (sočasne) meritve med seboj neodvisne ter da za oboje veljajo gaussovske lastnosti (Kalnay, 2003). To metodo se lahko razširi na 4D-Var, ki upošteva meritve, ki so veljavne ob različnih časih, vanjo pa mora biti vključen tudi model za razvoj stanja atmosfere v času.

V tem delu se bomo osredotočili na izvedbo variacijske asimilacije v latentnem prostoru variacijskega avtokodirnika (VAE). Zaradi izjemne računske potratnosti (več v poglavju *3D-Var v latentnem prostoru*) se stanje atmosfere pri klasični asimilaciji običajno transformira v manjši prostor z določenimi analitičnimi predpostavkami (Bannister, 2008b). Slednje pogosto temeljijo na poenostavitvah atmosferske dinamike, ki imajo omejeno aplikativnost. Zato je ideja naše metode zamenjava analitičnih transformacij z nevronsko mrežo, ki bi se sama naučila optimalne transformacije iz stanja atmosfere v *fizičnem prostoru* v manjši *latentni prostor*, ter v nasprotni smeri.

Doslej je bilo v literaturi predstavljenih zelo malo primerov o asimilaciji meritev v latentnem prostoru, dobljenem z NN. Mack in sod. (2020) so predstavili teoretični algoritem za 3D-Var v latentnem prostoru standardnega avtokodirnika (AE), ki pa ima nekaj omejitev. Število meritev mora biti enako dimenziji latentnega prostora ali pa je treba meritve interpolirati v latentni prostor, kar storijo z dodatno NN za interpolacijo opazovanj. To je smiselna rešitev za primer, ko je za asimilacijo vedno na voljo enako število opazovanj na istih lokacijah, nikakor pa ne za numerično napovedovanje vremena, kjer število opazovanj in njihova lokacija vseskozi varira, kar bi pomenilo vnovični trening NN za vsako novo postavitev. Tej težavi so se izognili Peyron in sod. (2021), ki so implementirali Kalmanov filter v latentnem prostoru AE za Lorenzov '96 model. Implementacijo Kalmanovega filtra v latentnem prostoru AE za bolj praktičen primer so izvedli Amendola in sod. (2021), in sicer za asimilacijo meritev v modelu onesnaženosti zraka v zaprtem prostoru, kjer so znova lahko predpostavili konstantno število meritev (ves čas so imeli na voljo meritve iz sedmih senzorjev). Prvi primer asimilacije z nevronskimi mrežami za napovedovanje vremena sta

predstavila Melinc in Zaplotnik (2024) za asimilacijo meritev temperature na tlačni ploskvi 850 hPa. Njun pristop se lahko uporabi za poljubno število meritev na poljubnih lokacijah brez vnovičnega treniranja nevronske mreže, v tem članku pa ga bomo razširili na asimilacijo meritev v multivariatnem primeru z asimilacijo poljubne količine izmed trojice zonalni veter, meridionalni veter ter geopotencial na tlačni ploskvi 200 hPa.

Podatki

Ker smo asimilacijske eksperimente izvajali na dveh množicah fizikalnih količin, smo natrenirali dva variacijska avtokodirnika – enega za avtokodiranje temperature na nivoju 850 hPa (T850; v nadaljevanju *univariatni primer*) in enega za avtokodiranje trojice zonalni veter, meridionalni veter ter geopotencial, vse na višini 200 hPa (u200, v200, Z200; v nadaljevanju *multivariatni primer*). Vse podatke za treniranje smo dobili iz reanalize ERA5 (Hersbach in sod., 2020). Za vsako količino smo izračunali njeno dnevno povprečje na njihovi izvirni regularni mreži z resolucijo 0.25°. Slednja ima težave s singularnostmi pri polih, ki predstavljajo težavo za trening nevronskih mrež s konvolucijskimi plastmi (Perkan, 2023), zato smo podatke interpolirali na regularno mrežo z enako resolucijo, vendar brez polov.

Vhodi v nevronsko mrežo so bili standardizirani, tako da smo jim za vsak posamezen dan v letu odšteli lastno klimatološko povprečje za obdobje 1981-2010 ter jih delili s klimatološkim standardnim odklonom za isto obdobje. Klimatološko povprečje in standardni odklon sta bila računana za vsako točko na mreži in vsak dan v letu posebej.

Podatke smo razdelili na sledeči način: obdobje 1979-2014 je bilo uporabljeno kot množica za trening nevronske mreže, obdobje 2015-2018 kot validacijska množica ter obdobje 2019-2022 kot testna množica.

Variacijski avtokodirnik

Za razumevanje koncepta variacijskega avtokodirnika (angl. *variational autoencoder*, VAE) najprej razložimo delovanje standardnega avtokodirnika (angl. *autoencoder*, AE). AE je nevronska mreža, katere cilj je na izhodu poustvariti vhod vanjo. Sestavljena je iz dveh delov: kodirnik E pretvori vhod $\mathbf{x}^{\mathrm{in}}$ v latentni vektor $\mathbf{z}$ v prostoru zmanjšane dimenzije (t. i. *latentni prostor*), torej $\mathbf{z} = E(\mathbf{x}^{\mathrm{in}})$, dekodirnik D pa pretvori latentni vektor ponovno v izvirni prostor vhoda (*fizični prostor*), torej $\mathbf{x} = D(\mathbf{z}) = D(E(\mathbf{x}^{\mathrm{in}}))$, kjer je $\mathbf{x}$ izhod iz AE. Za izhod iz popolnega AE bi veljalo $\mathbf{x} = \mathbf{x}^{\mathrm{in}}$.

Tako kot AE je tudi VAE sestavljen iz kodirnika in dekodirnika, vendar tokrat kodirnik ne vrne točne vrednosti elementov latentnega vektorja temveč parametre normalnih porazdelitev, iz katerih se te žreba kot (Kingma in Welling, 2022)

$$z_i = \mu_i + \hat{z}_i \sigma_i, \quad \hat{z}_i \sim \mathcal{N}(0,1), \tag{1}$$

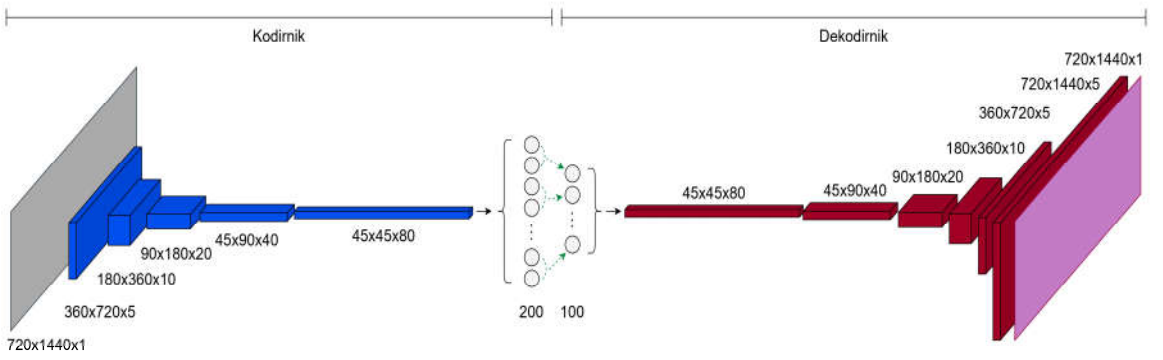
kjer je i indeks elementa latentnega vektorja, μ_i in σ_i sta povprečje in standardni odklon iz kodirnika, $\hat{z}_i$ pa je izžreban po standardni normalni porazdelitvi. Izžrebani latentni vektor $\mathbf{z}$ nato vstopi v dekodirnik, izhod iz dekodirnika pa je ponovno v fizičnem prostoru. Zaradi žreba je VAE stohastičen – za dva enaka vhoda bo vrnil dva različna izhoda, ki sta perturbirani verziji vhoda. Če je VAE reprezentativen, je vsak njegov izhod fizikalno smiseln, množica njegovih izhodov pa ima smiselne statistične lastnosti (v našem primeru to pomeni, da mora množica izhodov slediti klimatološkim značilnostim za izbrani datum).

Zaradi stohastičnih lastnosti VAE ima funkcija izgube $\mathcal{L}$, ki jo minimiziramo med treniranjem nevronske mreže, dva člena (Kingma in Welling, 2019): *rekonstrukcijski člen* $\mathcal{L}^{\mathrm{rec}}$ in *regularizacijski člen* $\mathcal{L}^{\mathrm{reg}}$, skupaj torej

$$\mathcal{L} = \mathcal{L}^{\mathrm{rec}} + \mathcal{L}^{\mathrm{reg}}. \tag{2}$$

$\mathcal{L}^{\mathrm{rec}}$ je enak edinemu členu v funkciji izgube za AE: njegov cilj je, da je izhod iz nevronske mreže čim bolj podoben vhodu. $\mathcal{L}^{\mathrm{reg}}$ pa spodbuja stohastične lastnosti VAE (tj. veča σ_i) in vsiljuje predpisano porazdelitev latentnega vektorja kot celote. Običajno to pomeni, da je porazdelitev vrednosti vseh elementov $\mathbf{z}$ čim bolj podobna standardni normalni porazdelitvi (Goodfelow in sod., 2016; Kingma in Welling, 2019). $\mathcal{L}^{\mathrm{reg}}$ posredno zagotovi tudi gladkost latentnega prostora. To pomeni, da z dekodiranjem dveh latentnih vektorjev, ki sta si relativno blizu v latentnem prostoru, dobimo dve polji v fizičnem prostoru, ki sta si prav tako relativno blizu. To nam omogoča (1) generiranje novih fizikalno smiselnih polj v fizičnem prostoru zgolj z žrebanjem celotnega $\mathbf{z}$ po standardni normalni porazdelitvi; (2) generiranje celotnega ansambla napovedi na podlagi enega samega člana, ki ga pripeljemo na vhod VAE (demonstrirano v Grooms (2021)); (3) asimilacijo meritev s cenilko definirano v latentnem prostoru, kar bomo pokazali v tem članku. Gladkost latentnega prostora je pri asimilaciji meritev glavna prednost VAE pred standardnimi AE. Brez te lastnosti lahko že majhna sprememba latentnega vektorja vodi v velike spremembe polja v fizičnem prostoru (Grooms, 2021), pri asimilaciji meritev pa povzročamo ravno popravke latentnega vektorja.

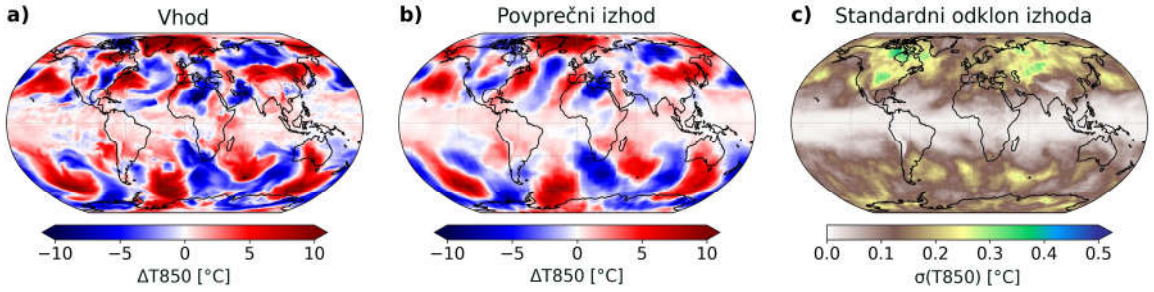
Arhitektura naše nevronske mreže je prikazana na sliki 1 in temelji na arhitekturi iz Brohan (2022) z nekaj manjšimi prilagoditvami. Pri VAE za T850 je imel latentni vektor $N = 100$ elementov, za multivariatni primer (u200, v200, Z200) pa smo uporabili $N = 200$. Kodirnik je bil večinoma sestavljen iz 2D konvolucijskih, dekodirnik pa 2D transponiranih konvolucijskih plasti, v obeh primerih pa so te plasti vsebovale jedra velikosti 3×3, večanje in manjšanje vmesnih polj pa je bilo doseženo z ustreznim korakanjem teh jeder. V 2D konvolucijskih plasteh je bilo poskrbljeno za oblaganje vmesnih plasti v skladu z robnimi pogoji ob datumski meji in na polih. Na izhodnih plasteh iz kodirnika in dekodirnika je bila uporabljena linearna aktivacijska funkcija, v gosti plasti dekodirnika je bila uporabljena aktivacijska funkcija ReLU, povsod drugod pa ELU. Tehnični pojmi iz tega odstavka so pojasnjeni v Perkan (2023).



Slika 1: Arhitektura variacijskega avtokodirnika za vhodno polje T850. Vhodno (sivo) polje s 720×1440 točkami v zonalni in meridionalni smeri vstopi v kodirnik. Ta z 2D konvolucijskimi plastmi postopno zmanjšuje velikost polja in mu dodaja kanale. Vmesna polja so obarvana modro, številke pod njimi predstavljajo velikost teh polj in število kanalov. Končno polje vstopi v gosto povezano plast (črna puščica) z 200 nevroni, ki predstavljajo po 100 povprečij (μ) in logaritmov varianc ($\log \sigma^2$), tj. po en par (μ_i, $\log \sigma_i^2$)

za vsak element latentnega vektorja. Ti elementi so nato žrebani po normalni porazdelitvi s pripadajočimi parametri (μ_i, log σ_i^2) (zelene črtkane puščice). Vhod v dekodirnik je z gosto povezano plastjo transformiran v polje dimenzije 45×45×80, ki je nato z 2D transponiranimi konvolucijskimi plastmi postopno pretvorjeno (vmesna polja so obarvana rdeče) v izhod (vijolično), ki je enake oblike kot vhod. V multivariatnem primeru (vhodna polja (u200, v200, Z200)) je bilo število kanalov v vmesnih plasteh mreže podvojeno, prav tako je bilo zaradi dvakrat večjega latentnega prostora podvojeno število nevronov na izhodu iz kodirnika in vhodu v dekodirnik. (Prirejeno po: Melinc in Zaplotnik, 2024)

Na sliki 2 je prikazano značilno obnašanje izhoda iz našega VAE. Zaradi boljše preglednosti je namesto T850 risano polje njenih anomalij (tj. odstopanj od klimatološkega povprečja) ΔT850. Podobno kot pri številnih poskusih rekonstrukcij vremenskih polj z nevronskimi mrežami (npr. Weyn in sod., 2020) je tudi naše povprečno izhodno polje zglajena verzija vhodnega. VAE dobro opiše prostorske variacije temperature na veliki, sinoptični prostorski skali, variacij na majhnih skalah pa v rekonstruiranem izhodnem polju večinoma ni. Ponekod so opazne pravokotne strukture, ki so posledica konvolucijskih plasti v kodirniku in dekodirniku in bi se jim lahko izognili z uporabo drugačnih plasti. Značilne vrednosti standardnega odklona izhodnega polja so red velikosti manjše od značilnih vrednosti povprečja izhodnega polja in s tem praktično zanemarljive.



Slika 2: (a) Vhodno polje v VAE (predpostavljena resnica za ΔT850 iz ERA5 reanalize za 15. 4. 2019). (b) Povprečno izhodno polje iz VAE. (c) Standardni odklon izhodnih polj iz VAE. Povprečje in standardni odklon sta računana na podlagi 150 izhodov iz VAE. (Prirejeno po: Melinc in Zaplotnik, 2024)

Asimilacija psevdomeritev v latentnem prostoru variacijskega avtokodirnika

3D-Var v latentnem prostoru

Cilj 3D-variacijske asimilacije (3D-Var) je poiskati čim bolj točen približek dejanskemu stanju ozračja (predstavimo ga z vektorjem $\mathbf{x}$), pri čemer statistično objektivno združimo naše poznavanje atmosfere iz predhodne napovedi (zanj uporabimo izraz *ozadje* in ga označimo z $\mathbf{x}_b$) z opazovanji ($\mathbf{y}$), kar doseže z minimizacijo cenilke (Lorenc, 1986; Kalnay, 2003)

$$\mathcal{J}(\mathbf{x}) = \mathcal{J}_b + \mathcal{J}_o = \\ = (\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \{\mathbf{y}-H(\mathbf{x})\}^{\mathrm{T}}\mathbf{R}^{-1}\{\mathbf{y}-H(\mathbf{x})\}\,, \qquad (3)$$

kjer člen ozadja $\mathcal{J}_b$ meri razdaljo med stanjem atmosfere in ozadjem, člen opazovanj $\mathcal{J}_o$ pa meri razdaljo med opazovanji in iskanim stanjem atmosfere, ki ga z operatorjem

opazovanj H propagiramo v t. i. prostor opazovanj, $\mathbf{B}$ je matrika kovarianc napak ozadja ter $\mathbf{R}$ matrika kovarianc napake meritev. Stanju, ki da minimum cenilke (3), pravimo *analiza* ($\mathbf{x}_a = \arg\min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$) in služi kot začetni pogoj za novo napoved vremena, razliki med analizo in ozadjem pa *inkrement analize* ($\delta\mathbf{x} = \mathbf{x}_a - \mathbf{x}_b$).

V operativnem napovedovanju vremena predstavlja glavno težavo pri iskanju minimuma cenilke (3) izračun inverza matrike $\mathbf{B}$, saj ima $\mathbf{x} \sim 10^9$ elementov ter zato $\mathbf{B}$ dimenzijo $\sim 10^9 \times 10^9$ in s tem $\sim 10^{18}$ elementov, kar bi bilo za numerični izračun $\mathbf{B}^{-1}$ pogubno. Zato se v operativnih centrih zatekajo k transformaciji stanja $\mathbf{x}$ v reducirano stanje $\boldsymbol{\xi}$ (transformaciji pravimo *transformacija kontrolnih spremenljivk*, reduciranemu prostoru pa *kontrolni prostor*) z bistveno manjšim številom elementov (Bannister, 2008b, 2021), kar dosežejo z analitičnimi poenostavitvami atmosferskih tokov, npr. z nelinearno ravnotežno enačbo, kvazi-geostrofsko omega enačbo in termodinamskimi ravnovesji (ECMWF, 2023).

Transformacijo stanja $\mathbf{x}$ v nižje-dimenzionalni prostor lahko dosežemo tudi z nevronsko mrežo. Za izpeljavo cenilke za 3D-Var potrebujemo predpostavki, da so ozadje in opazovanja med seboj neodvisni ter da je napaka obojih gaussovsko porazdeljena. Nobena izmed teh dveh predpostavk ni kršena, če je ozadje definirano v latentnem prostoru VAE. Zato definiramo 3D-Var cenilko v latentnem prostoru VAE, kjer člen ozadja $\mathcal{J}_{bz}$ tokrat meri razdaljo med vektorjem stanja atmosfere v latentnem prostoru $\mathbf{z}$ in ozadjem v latentnem prostoru $\mathbf{z}_b$, člen opazovanj $\mathcal{J}_{oz}$ pa razdaljo med opazovanji $\mathbf{y}$ in stanjem $\mathbf{z}$, ki ga propagiramo nazaj v fizični prostor:

$$\begin{aligned}\mathcal{J}_z(\mathbf{z}) &= \mathcal{J}_{bz} + \mathcal{J}_{oz} = \\ &= (\mathbf{z}-\mathbf{z}_b)^{\mathrm{T}}\mathbf{B}_z^{-1}(\mathbf{z}-\mathbf{z}_b) + [\mathbf{y} - H\{D(\mathbf{z})\}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y} - H\{D(\mathbf{z})\}]\,, \end{aligned} \quad (4)$$

kjer je $\mathbf{B}_z$ matrika kovarianc napak ozadja v latentnem prostoru, D je dekodirnik, H pa v našem primeru predstavlja bilinearno interpolacijo dekodiranega polja na lokacijo opazovanja. Kot v primeru klasičnega 3D-Var dobimo analizo z minimizacijo cenilke (4), $\mathbf{z}_a = \arg\min_{\mathbf{z}} \mathcal{J}_z(\mathbf{z})$.

Podobno kot v primeru analitične transformacije v kontrolni prostor smo s premikom minimizacijskega problema v latentni prostor močno olajšali njegovo numerično zahtevnost. Vektor $\mathbf{x}$ bi že v univariatnem primeru imel $720 \times 1440 = 1\,036\,800$ elementov, pripadajoča matrika $\mathbf{B}$ posledično več kot 10^{12} elementov, latentni vektor $\mathbf{z}$ pa ima 100 elementov in njemu pripadajoča polna matrika $\mathbf{B}_z$ le še 10^4 elementov.

Postavitev eksperimenta

Shema naših eksperimentov za asimilacijo psevdomeritev v latentnem prostoru VAE je predstavljena na sliki 3. Tako ozadje kot opazovanja smo generirali iz ERA5 reanalize, ki smo jo predpostavili kot absolutno resnico, analizo pa smo dobili z minimizacijo cenilke (4).

Opazovanja za dan d na izbrani lokaciji smo dobili tako, da smo resnico za ta dan $\mathbf{x}_t^d$ interpolirali s H na to lokacijo, nato pa meritvi dodali naključni šum ε_o, izžreban iz normalne porazdelitve s povprečjem 0 in standardnim odklonom σ_o ($\varepsilon_o \sim \mathcal{N}(0, \sigma_o)$). Vektor meritev je torej bil

$$\mathbf{y} = H\left(\mathbf{x}_t^d\right) + \boldsymbol{\varepsilon}_o\,. \quad (5)$$

V tem članku so sicer prikazani samo rezultati eksperimentov za študijo vpliva posameznih opazovanj, kjer smo ta raje generirali tako, da smo interpolirani vrednosti iz ozadja za dan d prišteli konstantni *inkrement opazovanja* $\delta\mathbf{y}^o$, torej

$$\mathbf{y} = H\left(\mathbf{x}_b^d\right) + \delta\mathbf{y}^o + \boldsymbol{\varepsilon}_o\,. \quad (6)$$

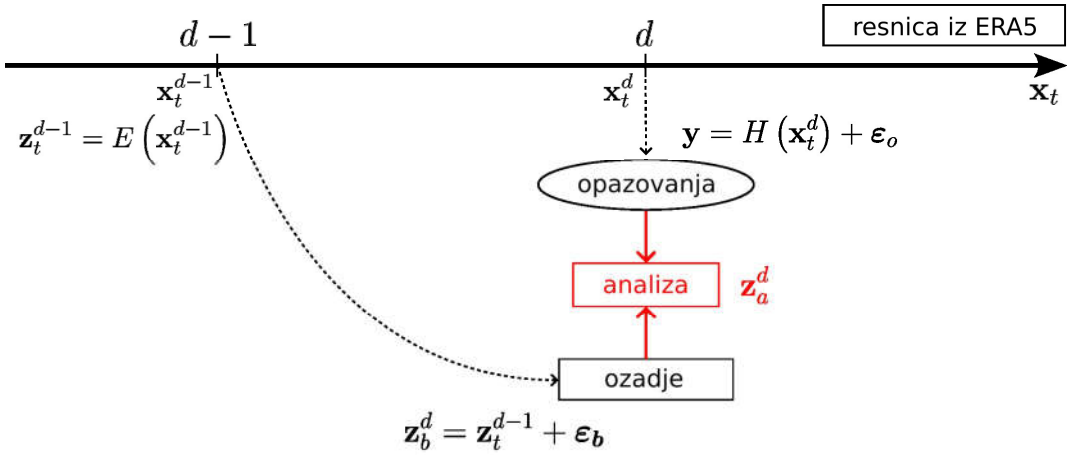
Zaradi preprostosti smo za ozadje v latentnem prostoru za dan d predpostavili, da je to enako resnici v tem prostoru za predhodnji dan $d-1$, torej

$$\mathbf{z}_b^d = \mathbf{z}_t^{d-1} = E\left(\mathbf{x}_t^{d-1}\right) \,. \tag{7}$$

Običajen izraz za predpostavko o konstantnosti stanja atmosfere je *persistenčni model*. Tudi latentnemu vektorju za ozadje smo dodali naključni gaussovski šum, kjer smo σ_{bi} dobili kot koren i-tega elementa diagonale matrike $\mathbf{B}_z$, povprečje pa iz izhoda iz kodirnika:

$$z_{bi} = \mu_i + \hat{z}_i\, \sigma_{bi} \,. \tag{8}$$

Končni rezultat smo dobili iz ansambla asimilacij (EDA, Bonavita in sod., 2012), kjer smo za 150 parov zašumljenih vektorjev opazovanj in ozadja izračunali minimum cenilke (4).



Slika 3: Postavitev simuliranega eksperimenta. (Prirejeno po: Melinc in Zaplotnik, 2024)

Modeliranje kovariančne matrike napake ozadja v latentnem prostoru

Kovariančno matriko napake ozadja po definiciji dobimo s povprečenjem vnanjega produkta napake ozadja same s seboj preko številnih parov,

$$\mathbf{B} = \left\langle (x_t - x_b)\ (x_t - x_b)^{\mathrm{T}} \right\rangle , \tag{9}$$

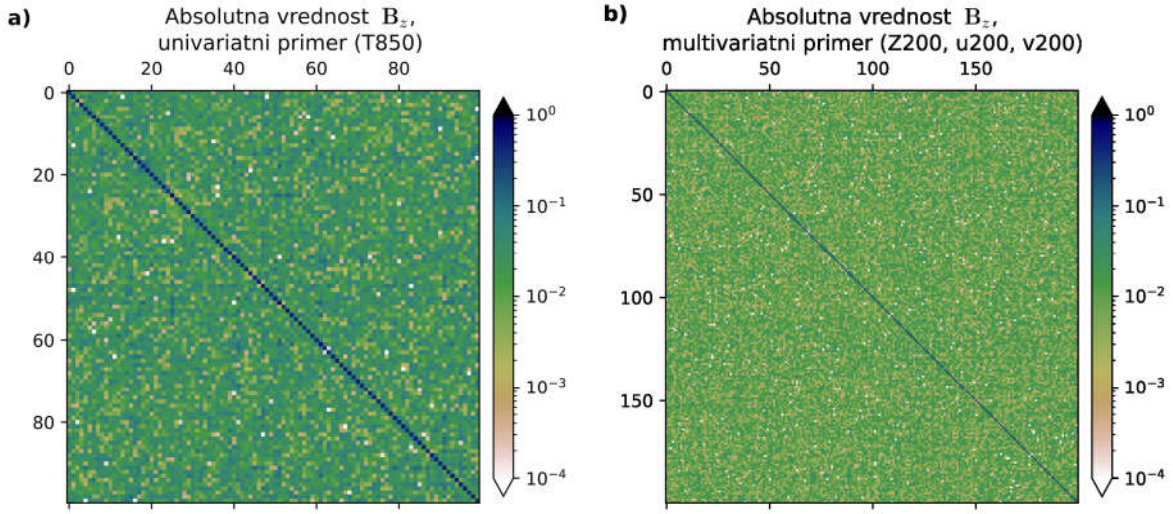
kjer z $\mathbf{x}_t$ označujemo resnično stanje atmosfere. Enačbo (9) lahko za primer, ko je ozadje definirano v latentnem prostoru, prepišemo kot

$$\begin{aligned} \mathbf{B}_z &= \left\langle (z_t - z_b)\ (z_t - z_b)^{\mathrm{T}} \right\rangle \\ &= \left\langle \left(z_t^d - z_t^{d-1}\right)\ \left(z_t^d - z_t^{d-1}\right)^T \right\rangle , \end{aligned} \tag{10}$$

kjer smo v drugi vrstici upoštevali naš način modeliranja ozadja (persistenčni model).

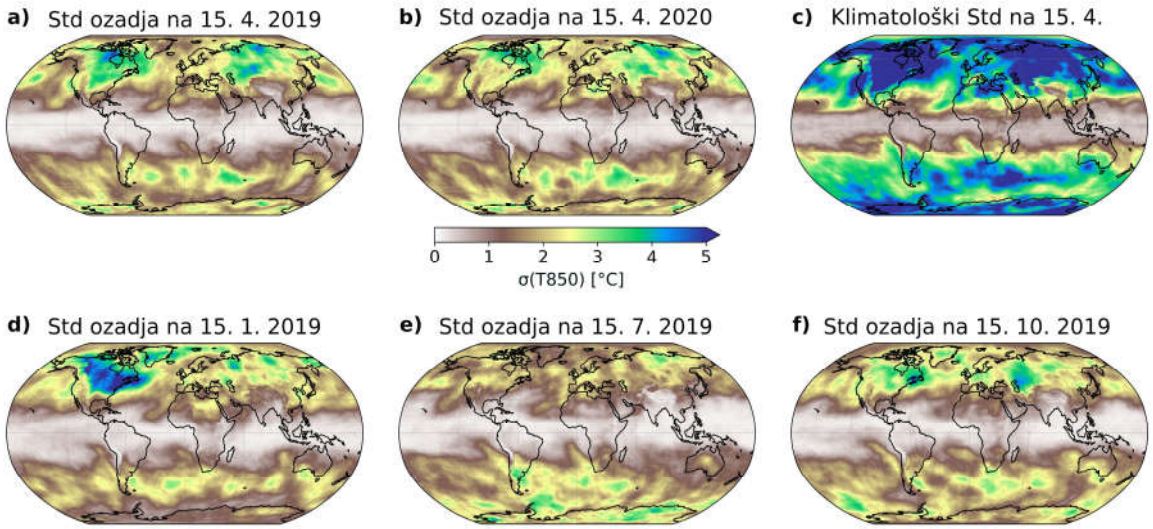
Na sliki 4 sta prikazani matriki $\mathbf{B}_z$ za univariatni in multivariatni primer, ki smo ju dobili s povprečenjem po celotni validacijski množici in ju nato uporabili pri asimilacijskih eksperimentih. Izračun inverza matrike $\mathbf{B}_z$, ki ga potrebujemo pri minimizaciji cenilke (4) bi bil poceni že za polno matriko dimenzije 100×100 (oz. 200×200), vendar se zavedamo, da bi bila v primeru operativnega napovedovanja vremena velikost latentnega prostora zagotovo precej večja. S slike 4 je očitno, da so diagonalni elementi matrike $\mathbf{B}_z$ precej večji

od izvendiagonalnih, zato smo si že v naših eksperimentih privoščili predpostavko o diagonalni strukturi $\mathbf{B}_z$, saj to bistveno olajša izračun njenega inverza. V nekaj korakih smo tako računsko potraten problem obračanja matrike dimenzije $10^6 \times 10^6$ (oz. $3 \cdot 10^6 \times 3 \cdot 10^6$) poenostavili na izračun obratnih vrednosti vsega stotih (oz. dvestotih) skalarjev.



Slika 4: Matrika , dobljena z enačbo (10) s povprečenjem po vseh datumih v množici za validacijo (1. 1. 2015-31. 10. 2018) za (a) univariatni primer in (b) multivariatni primer. (Panel (a) prirejen po: Melinc in Zaplotnik, 2024)

Kljub temu da pri zašumljenju ozadja v latentnem prostoru (enačba (8)) vedno uporabimo isto $\mathbf{B}_z$ in je zato standardni odklon ozadja v latentnem prostoru ves čas enak, to ne velja za standardni odklon ozadja v fizičnem prostoru σ_b. Na sliki 5 vidimo, da je slednji odvisen predvsem od dneva v letu oz. klimatološkega standardnega odklona za ta dan. To je posledica standardizacije vhoda v VAE, ki je odvisna od dneva v letu. Vendar pa se σ_b zaradi nelinearnosti dekodirnika nekoliko razlikuje tudi na isti zaporedni dan v letu v različnih letih, kar se na panelih (a) in (b) vidi predvsem nad Severno Ameriko in južnim Atlantskim oceanom. Zaključimo lahko, da statična $\mathbf{B}_z$-matrika v latentnem prostoru zaradi nelinearnosti dekodirnika opiše v fizičnem prostoru napake ozadja, ki so odvisne od trenutnega stanja ozračja, opisanega z latentnim vektorjem $\mathbf{z}$.

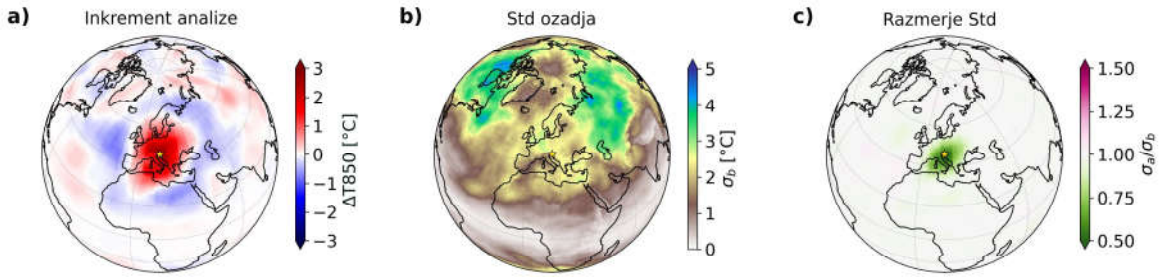


Slika 5: Standardni odklon ozadja na različne datume in klimatološki standardni odklon na 15. april. Barvna lestvica na sredini slike velja za vse panele. Standardni odkloni ozadja so bili izračunani na podlagi 150 članov ansambla. (Prirejeno po: Melinc in Zaplotnik, 2024)

Rezultati

Asimilacija posamezne meritve v sistemu (T850)

Atmosferska dinamika v tropih se precej razlikuje od dinamike zmernih širin, kar je predvsem posledica Coriolisove sile, ki je odvisna od geografske širine. V zmernih širinah zato na sinoptičnih skalah prevladuje ravnovesje termalnega vetra, ki povezuje vertikalno spreminjanje geostrofskega vetra s horizontalnim spreminjanjem temperature, značilne vzorce na velikih razdaljah pa predstavljajo planetarni Rossbyjevi valovi. V tropih pa prevladuje ravnovesje med vertikalnim vetrom in diabatnim gretjem s kondenzacijo, zato tam povezave na velikih skalah določajo ekvatorialni valovi (Matsuno, 1966). Zato se analitične funkcije, s katerimi bi opisali ravnovesja v $\mathbf{B}$-matriki, močno razlikujejo in se v praksi v operativnem napovedovanju vremena upoštevajo samo tista za zmerne širine (Bannister 2008b, 2021; EMCWF 2023). Z naslednjimi primeri bomo pokazali, da $\mathbf{B}_z$-matrika smiselno upošteva tako ravnovesja v zmernih širinah kot ravnovesja v tropih.



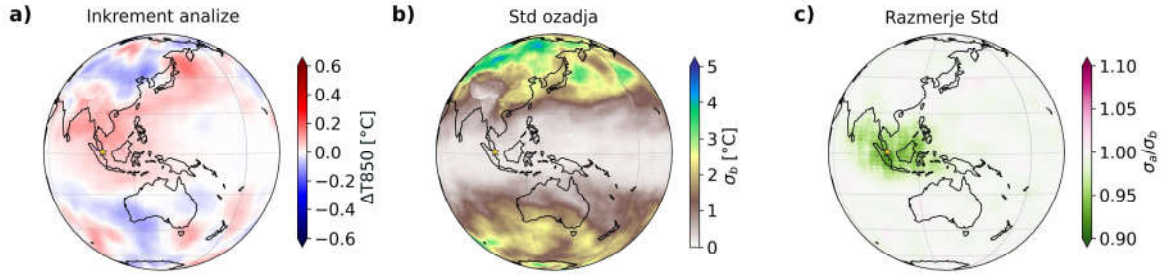
Slika 6: Asimilacija meritve T850 nad Ljubljano z $\delta\mathbf{y}^o$=3 K in σ_o=1 K z ozadjem za 15. 4. 2019. (a) Povprečni inkrement analize (tj. povprečna analiza v fizičnem prostoru, ki ji odštejemo popvprečno ozadje v fizičnem prostoru). (b) Standardni odklon ozadja v fizičnem prostoru (σ_b). (c) Razmerje med standardnim odklonom analize in standardnim odklonom ozadja (σ_a / σ_b). Lokacija opazovanja je označena z zlato zvezdo. Asimilacija je bila izvedena za 150 članov ansambla. (Prirejeno po: Melinc in Zaplotnik, 2024)

Opazovanja v univariatnem sistemu (T850) smo generirali po enačbi (6), kjer smo nastavili $\delta\mathbf{y}^o = 3\,\mathrm{K}$ in $\sigma_o = 1\,\mathrm{K}$. Na sliki 6 je primer asimilacije meritve v zmernih širinah, in sicer nad Ljubljano. Panel (a) prikazuje inkrement analize. Ta je smiseln zaradi sledečih lastnosti:

1. Ima največjo vrednost v točki opazovanja in relativno veliko vrednost nad celino, kjer je tudi standardni odklon ozadja σ_b nekoliko večji kot nad Sredozemljem (panel (b), večji σ_b teoretično pomeni manjši vpliv ozadja na analizo, zaradi česar lahko meritev bolj popravi ozadje). Zmanjšanje σ_b nad morjem je značilno za klimatološke $\mathbf{B}$ matrike zaradi sorazmerno počasnega spreminjanja temperature površine morja, ki vpliva na temperaturo v spodnji troposferi.

2. Območje s pozitivnim inkrementom analize je raztegnjeno v smeri jugozahod-severovzhod, kar pojasnimo z značilnim jugozahodnim vetrom na tem območju in advekcijo z njim.
3. Območje s pozitivnim inkrementom analize obkroža plitkejše območje negativnega inkremeta. To ponovno pojasnimo s klimatološkim izvorom $\mathbf{B}_z$ matrike, ki pozitivni inkrement opazovanja povezuje s prostorskim premikom planetarnega Rossbyjevega vala (Fisher, 2003).

Amplituda inkrementa analize naprej od območja negativnega inkrementa je zanemarljivo majhna v primerjavi s standardnim odklonom analize. Iz panela (c) lahko sklepamo, da je opazovanje resno vplivalo na analizo zgolj na območju pozitivnega inkrementa, saj samo tam beležimo resen upad razmerja med standardnim odklonom analize σ_a in ozadja.

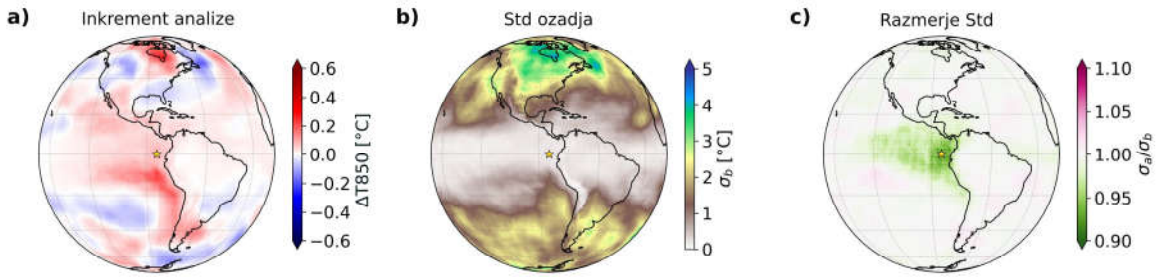


Slika 7: Kot slika 6, le da je meritev nad Singapurjem.
(Prirejeno po: Melinc in Zaplotnik, 2024)

Na sliki 7 je primer asimilacije v tropih, in sicer za meritev nad Singapurjem. Za klimatološke $\mathbf{B}$ matrike je značilno, da je njihov σ_b podcenjen v tropih in precenjen v zmernih širinah (Bannister, 2008a), kar je demonstrirano v panelu (b), posledice tega pa se odražajo v majhni magnitudi inkrementa analize na lokaciji meritve. Tam je namreč $\sigma_b \approx 0.19$ in s tem $\sigma_o \gg \sigma_b$. V zmernih širinah pa je magnituda inkrementa analize kljub oddaljenosti primerljiva z magnitudo na lokaciji meritve. Tam je namreč $\sigma_o \ll \sigma_b$, ozadje je torej manj natančno določeno in meritev ga lahko popravi toliko bolj.

V zmernih širinah pa ne pride do njenega upada, kljub temu da je meritev od njih precej oddaljena. Zanimiva je asimetrična oblika inkrementa analize (predvsem raztegnjenost pozitivnega inkrementa proti Japonski in Severni Ameriki), ki spominja na odziv na diabatno segrevanje nad Malajskim otočjem v obliki Rossbyjevega valovnega vlaka (Matsuno, 1966; Gill, 1980; Kosovelj in sod., 2019). Poleg tega je upad σ_a v primerjavi s σ_b rahlo raztegnjen proti vzhodu, kar povezujemo s tipičnimi vzhodnimi vetrovi nad Singapurjem in vzhodno od njega.

Še en primer asimilacije v tropih je predstavljen na sliki 8, kjer je opazovanje generirano vzhodno od obale Ekvadorja. Inkrement analize nad Tihim oceanom se sklada z značilno obliko El-Nino južne oscilacije (ENSO), upad σ_a v primerjavi s σ_b pa je močno razpotegnjen proti zahodu, tj. po celotnem območju spodnje veje vzhodnega krila Pacifiške Walkerjeve cirkulacije.



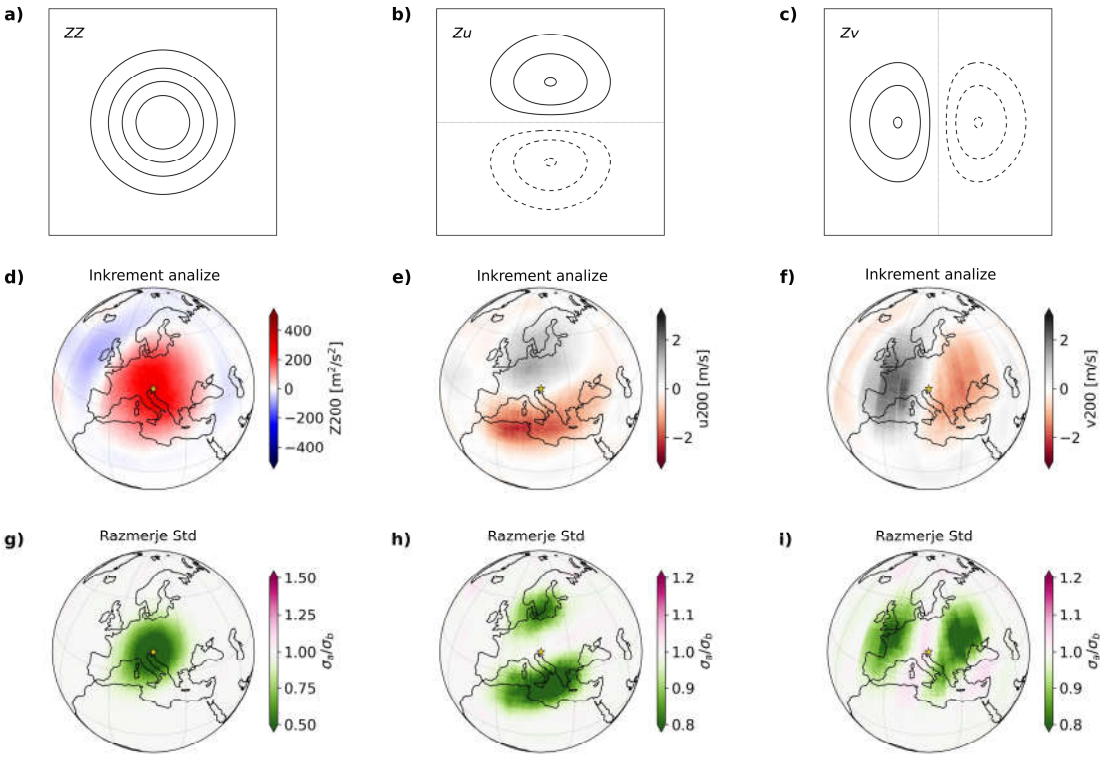
Slika 8: Kot slika 6, le da je meritev vzhodno od obale Ekvadorja.
(Prirejeno po: Melinc in Zaplotnik, 2024)

Slike 6, 7 in 8 dokazujejo, da z matriko $\mathbf{B}_z$ dobimo smiselne inkremente analize tako v zmernih širinah kot v tropih. Z drugimi besedami, $\mathbf{B}_z$ matrika uspešno upošteva tako ravnovesja v zmernih širinah kot ravnovesja v tropih.

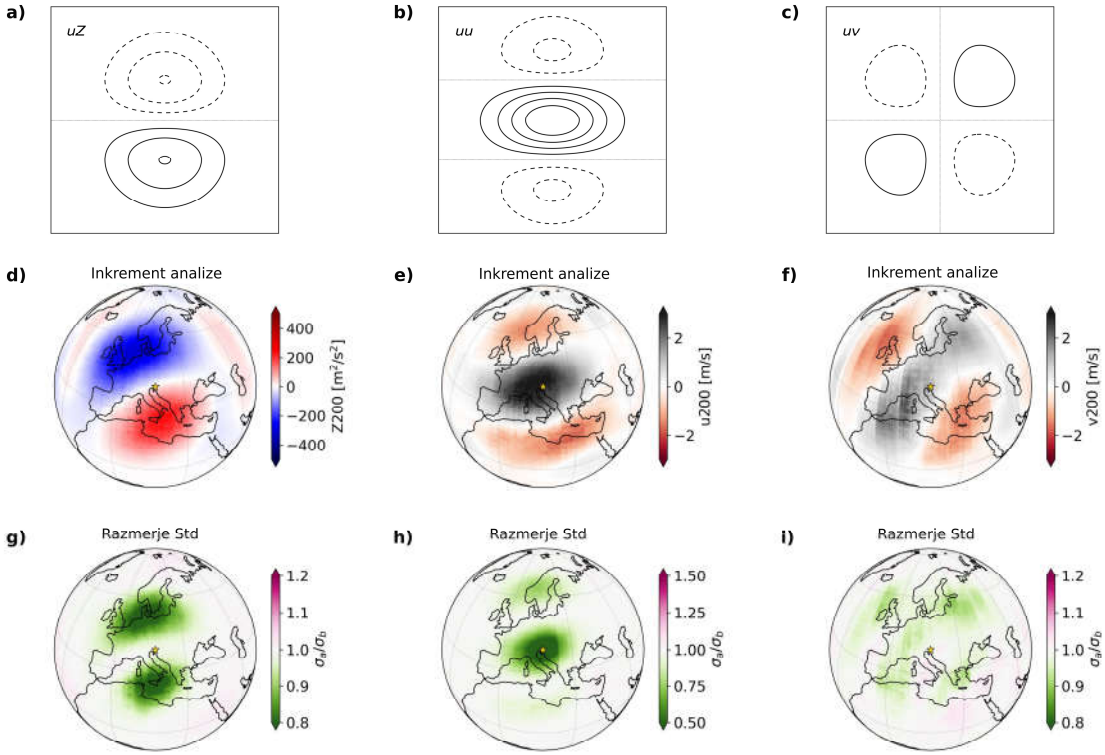
Asimilacija posamezne meritve v sistemu (u200, v200, Z200)

Spremenljivke v atmosferi so med seboj povezane. Na 200 hPa tlačni ploskvi je v zmernih širinah npr. polje horizontalnega vetra (u200, v200) dobro sklopljeno s poljem geopotenciala (Z200) prek geostrofskega ravnovesja. Podobno kot spremenljivke so med seboj povezane tudi napake teh spremenljivk. Kovariance napak opišemo v matriki B. Posledično meritev neke količine v ozračju ne vpliva le na lastno analizo, temveč tudi na analizo ostalih, sklopljenih količin. Pri eksperimentih, opisanih v tem razdelku, smo z VAE predstavili globalna polja (u200, v200, Z200) z enim samim latentnim vektorjem z 200 elementi. V nadaljevanju zato pokažemo nekaj primerov multivariatne asimilacije meritev u200, v200 in Z200 v zmernih širinah.

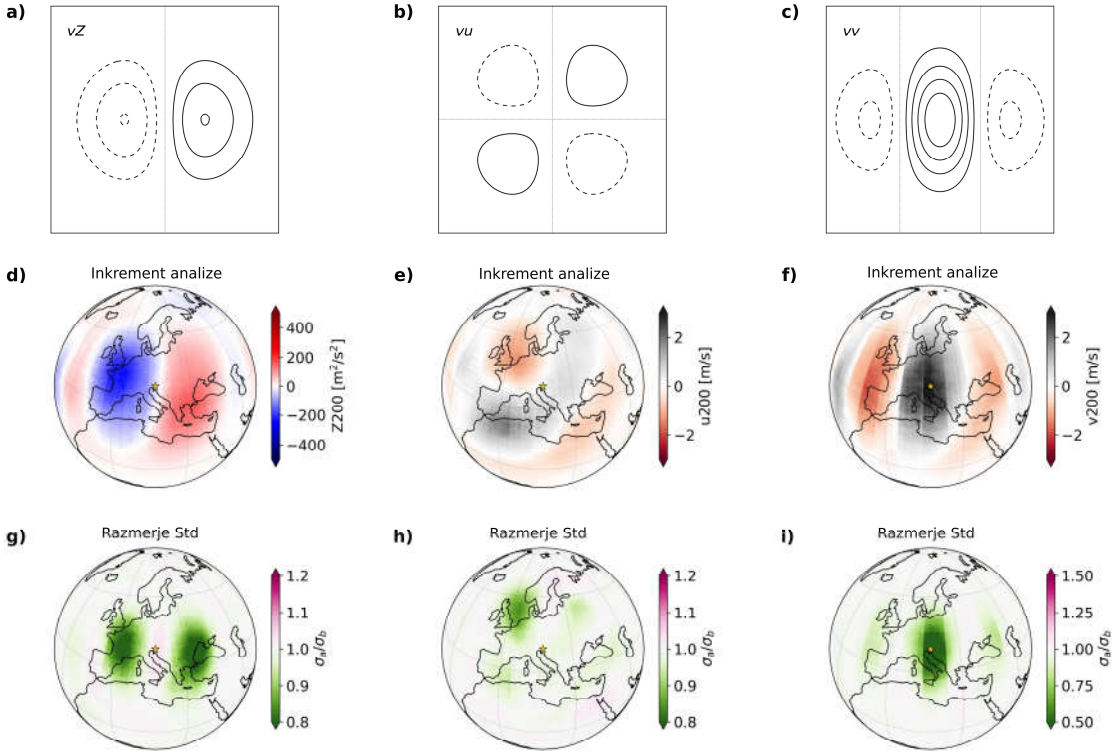
Najprej na sliki 9 prikažemo primer asimilacije meritve geopotenciala v tem sistemu. V panelih (a), (b) in (c) so so narisane korelacijske funkcije sklopljenih polj ob predpostavljenem geostrofskem ravnovesju in konstantnem standardnem odklonu ozadja (metodologija je razložena v poglavju 5.4.2 dela Kalnay (2023)). S korelacijskimi funkcijami si lahko med drugim pomagamo pri napovedovanju oblike inkrementa analize po asimilaciji posamezne meritve. Če je inkrement opazovanja Z200 pozitiven (kar velja za primer na tej sliki), potem pozitivna/negativna vrednost korelacijske funkcije v zgornji/spodnji polovici panela (b) napove pozitiven/negativen inkrement analize u200 severno/južno od lokacije meritve Z200. Točno tak odziv je prikazan v panelu (e), medtem ko panel (h) potrjuje, da je vpliv meritve omejen ravno na območji s pričakovanim teoretičnim inkrementom analize. Paneli (a), (d) in (g) potrjujejo tudi ustreznost odziva Z200, paneli (c), (f) in (i) pa ustreznost odziva v200. Rahlo popačenost odzivov v primerjavi s teoretičnimi lahko pripišemo (1) nehomogenosti standardnega odklona ozadja, ki jo predpostavi teorija, in (2) lastnosti nevronske mreže, da prilagodi odziv lokalnim značilnostim območja v bližini meritve. Podobno skladnost rezultatov vidimo tudi v primeru meritve u200 na sliki 10 in v200 na sliki 11.



Slika 9: Asimilacija meritve Z200 nad Ljubljano z $\delta \mathbf{y}^o = 300\,\mathrm{m}^2/\mathrm{s}^2$ in $\sigma_o = 100\,\mathrm{m}^2/\mathrm{s}^2$ z ozadjem za 15. 4. 2019. (a, b, c) Pričakovani odziv Z200, u200 ter v200 v primeru meritve Z200 v središču panela. Risane so izolinije korelacijske funkcije ob predpostavki geostrofskega inkrementa s konstantnim standardnim odklonom ozadja (konture pozitivne korelacije so risane s polnimi črtami, negativne s črtkanimi črtami, nična korelacija pa s tanko pikčasto črto, korak med izolinijami je 0.2). (d, e, f) Povprečni inkrementi analize Z200, u200, v200. (g, h, i) Razmerje med standardnim odklonom analize in standardnim odklonom ozadja za količine v enakem zaporedju kot pri (a, b, c) in (d, e, f). Lokacija opazovanja je označena z zlato zvezdo. Asimilacija je bila izvedena za 150 članov ansambla.



Slika 10: Kot slika 9, le da za meritev u200 z $\delta \mathbf{y}^o = 3\,\mathrm{m}^2/\mathrm{s}^2$ in $\sigma_o = 1\,\mathrm{m}^2/\mathrm{s}^2$



Slika 11: Kot slika 9, le da za meritev v200 z $\delta \mathbf{y}^o = 3\,\mathrm{m}^2/\mathrm{s}^2$ in $\sigma_o = 1\,\mathrm{m}^2/\mathrm{s}^2$.

Zaključki

V članku smo demonstrirali koncept variacijske asimilacije meritev ozračja z nevronskimi mrežami, tako da smo pokazali množico asimilacijskih eksperimentov z eno samo meritvijo različnih atmosferskih spremenljivk na različnih lokacijah.

Najprej smo s pomočjo podatkov reanalize ERA5 natrenirali kovolucijski variacijski avtokodirnik (VAE), ki smo ga uporabili za predstavitev polj temperature na 850 hPa (T850) v latentnem prostoru, tj. prostoru z močno zmanjšano dimenzijo (sliki 1 in 2). V tem prostoru smo definirali tudi 3D-Var cenilko (4), ki sestoji iz člena ozadja in člena opazovanj. Prvi meri meri razdaljo med vektorjem nekega stanja atmosfere in vektorjem prejšnje kratkoročne napovedi stanja atmosfere (ozadjem), ki sta oba definirana v latentnem prostoru, drugi pa meri razdaljo med dekodiranim vektorjem iskanega latentnega stanja in opazovanji. Z minimizacijo te cenilke smo dobili statistično optimalno oceno iskanega stanja ozračja, analizo, v latentnem prostoru.

Ključni prednosti VAE pred standardnim AE pri definiciji 3D-Var cenilke v latentnem prostoru sta njegova gladkost in gaussovskost. 3D-Var cenilko (4) smo definirali nekoliko drugače kot pretekle študije (Mack in sod., 2020; Amendola in sod., 2021). Člen opazovanj smo izvrednotili v prostoru meritev, tako da smo dekodirali latentni vektor v fizični prostor, na katerega deluje operator opazovanj na enak način kot pr pri klasični cenilki (3). S tem smo se izognili poddeterminiranemu problemu interpolacije meritev v fizični prostor in njegovi težko uporabni alternativi, tj. neposrednemu kodiranju meritev, ki zahteva ponovno treniranje nevronske mreže za vsako spremembo njihove postavitve. Naša metoda torej omogoča asimilacijo meritev ne glede na njihovo število, lokacijo ali natančnost, kar je nujna lastnost za numerično napovedovanje vremena.

Ker je bilo ozadje definirano v latentnem prostoru, smo v tem prostoru definirali tudi matriko kovarianc napak ozadja $\mathbf{B}_z$ (10), ki smo jo izmerili na enodnevnih razlikah kodiranih polj v časovnem odbodju 2015-2018. Pokazali smo, da je tako definirana $\mathbf{B}_z$ kvazi-diagonalna. Izvendiagonalni elementi so bili red velikosti manjši od diagonalnih, zato smo uporabili le slednje, kar je dodatno pospešilo minimizacijo cenilke (4). Kljub temu da je matrika $\mathbf{B}_z$ statična v latentnem prostoru, to ne velja za standardni odklon ozadja v fizičnem prostoru. Pokazali smo, da je slednji močno odvisen od dneva v letu, zaradi nelinearnosti dekodirnika pa tudi od trenutnega stanja atmosfere.

Z eksperimenti, v katerih smo za univariatni primer (T850) asimilirali posamezno meritev (slike 6, 7 in 8), smo pokazali, da matrika $\mathbf{B}_z$ opiše tako ravnovesja v zmernih širinah kot ravnovesja v tropih, slednjih pa trenutno sistemi za operativno napovedovanje vremena ne upoštevajo.

Avtokodirnik smo naučili tudi predstavitve vektorskega polja horizontalnega vetra na 200 hPa tlačni ploskvi in geopotenciala (u200, v200, Z200), pri čemer smo ustrezno povečali velikost latentnega vektorja. Pokazali smo, da matrika $\mathbf{B}_z$ uspešno razširi vpliv meritve tako v prostoru v okolici meritve kot tudi na druge količine (slike 9, 10 in 11). Na podlagi teh rezultatov menimo, da bo podoben princip deloval tudi za asimilacijo množice spremenljivk na različnih tlačnih nivojih.

V nadaljnjih raziskavah bomo, poleg razširitve na več tlačnih nivojev, poskusili dodati še količine, ki so v trenutnih prognostičnih modelih znane kot težko napovedljive, npr. intenziteta padavin. Tu se bomo oprli na nevronsko mrežo iz Perkan (2023), ki napoveduje časovni razvoj petih enonivojskih in petih večnivojskih spremenljivk. To mrežo bomo tudi uporabili za propagacijo naših polj v času, s čimer bomo lahko 3D-Var razširili na 4D-Var, ki lahko upošteva nesočasne meritve. Z ansambelsko napovedjo s 4D-Var bi lahko skonstruirali še napovedni del matrike $\mathbf{B}_z$, ki bi bil odvisen izključno od negotovosti trenutnega stanja atmosfere in ne več od klimatoloških značilnosti (Bonavita in sod., 2016).

S tem bi naš pristop zelo približali operativnemu napovedovanju vremena, kjer je matrika $\mathbf{B}$ sestavljena kot linearna kombinacija klimatološkega in napovednega dela (ECMWF, 2023).

Literatura

Amendola, M., Arcucci, R., Mottet, L., Casas, C. Q., Fan, S., Pain, C., Linden, P. in Guo, Y. K.. (2021). Data Assimilation in the Latent Space of a Convolutional Autoencoder. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12746 LNCS:373–386.

Bannister, R. N. (2008a). A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. Quarterly Journal of the Royal Meteorological Society, 134, 1951–1970.

Bannister, R. N. (2008b). A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. Quarterly Journal of the Royal Meteorological Society, 134, 1971–1996.

Bannister, R. N. (2021). Balance conditions in variational data assimilation for a high-resolution forecast model. Quarterly Journal of the Royal Meteorological Society, 147, 2917–2934.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X. in Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619(7970), 533–538.

Bjerknes, V. (1904). Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. Meteorologische Zeitschrift, 21, 1–7.

Bonavita, M., Hólm, E., Isaksen, L. in Fisher, M. (2016). The evolution of the ECMWF hybrid data assimilation system. Quarterly Journal of the Royal Meteorological Society, 142, 287–303.

Bonavita, M., Isaksen, L. in Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. Quarterly Journal of the Royal Meteorological Society, 138(667), 1540–1559.

Bonavita, M. (2023). On the limitations of data-driven weather forecasting models. https://doi.org/10.48550/arXiv.2309.08473

Bouallegue, Z. B. in sod. (2023). The rise of data-driven weather forecasting. https://doi.org/10.48550/arXiv.2307.10128

Brohan, P. (2022). Machine Learning for Data Asimilation (v1.0.0). https://github.com/philip-brohan/Proxy_20CR

Charney, J. G., Fjoertoft, R. in von Neuman, J. (1950). Numerical integration of the barotropic vorticity equation. Tellus, 2(4), 237–254.

Dueben, P. D. in Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Development, 11(10), 3999–4009.

ECMWF (2023). IFS Documentation CY48R1 - Part II: Data Assimilation. IFS Documentation CY48R1, no. 2. ECMWF, Reading, UK.

Fisher, M. (2003). Background error covariance modelling. Seminar on Recent Development in Data Assimilation, p. 45–63. ECMWF, Reading, UK.

Gill, A. E. (1980). Some simple solutions for heat-induced tropical circulation. Quarterly Journal of the Royal Meteorological Society, 106(449), 3759–3777.

Goodfellow, I., Bengio, Y. in Courville, A. (2016). Deep Learning. MIT Press.

Grooms, I. (2021). Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. Quarterly Journal of the Royal Meteorological Society, 147, 139–149.

Hakim G. J. in Masanam S. (2023). Dynamical tests of a deep-learning weather prediction model. https://doi.org/10.48550/arXiv.2309.10867

Hersbach, H., in sod. (2020). The ERA5 Global Reanalysis. Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049.

Kalnay, E. (2003) Atmospheric modeling, data assimilation, and predictability. Cambridge University Press.

Kingma, D. P. in Welling M. (2019). An introduction to variational autoencoders. Foundations and Trends in Machine Learning, 12, 307–392.

Kingma, D. P. in Welling M. (2022). Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114

Kosovelj, K., Kucharski, F., Molteni, F. in Žagar, N. (2019). Modal Decomposition of the Global Response to Tropical Heating Perturbations Resembling MJO. Journal of the Atmospheric Sciences, 76(5), 1457–1469.

Lahoz, W. A. in Schneider, P. (2014). Data assimilation: making sense of Earth Observation. Frontiers in Environmental Science, 2, 16.

Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112, 1177–1194.

Matsuno, T. (1966). Quasi-geostrophic motions in the equatorial area. Journal of the Meteorological Society of Japan, 44, 25–43.

Mack, J., Arcucci, R., Molina-Solana, M. in Guo, Y. K. (2020) Attention-based Convolutional Autoencoders for 3D-Variational Data Assimilation. Computer Methods in Applied Mechanics and Engineering, 372, 113291.

Melinc, B. in Zaplotnik, Ž. (2024). 3D-Var Data Assimilation using a Variational Autoencoder. https://doi.org/10.48550/arXiv.2308.16073

Perkan, U. (2023): Napovedovanje vremena s konvolucijskimi nevronskimi mrežami (Weather Forecasting with Convolutional Neural Networks). Master Thesis, Univerza v Ljubljani, FMF, 92 p. (in Slovenian).

Peyron, M., Fillion, A., Gürol, S., Marchais, V., Gratton, S., Boudier, P. in Goret, G. (2021). Latent space data assimilation by using deep learning. Quarterly Journal of the Royal Meteorological Society, 147(740), 3759–3777