Analog Methode

Martha Kogler, Vanessa Seitner

Sommersemester 2019

1 Motivation

Um Abschätzungen über das zukünftige Klima machen zu können, werden als Ausgangspunkt globale Klimamodelle (GCM) verwendet. Diese haben eine hohe räumliche Auflösung und müssen für lokalskalige Untersuchungen downgescaled werden. Dafür gibt es 2 Methoden:

1. Zum einen gibt es das dynamische Downscaling, bei der regionale Modelle am Rand die Informationen der GCMs übernehmen. Dynamische Downscalingmethoden benötigen jedoch eine hohe Rechenkapazität und weisen oft einen hohen Bias auf.

2. Daher gibt es zum anderen das empirisch statistische Downscaling. Das stützt sich auf Beobachtungen, die in der Vergangenheit einerseits auf der grobskaligen GCM-Skala und andererseits auf der regional/lokalen Skala gemacht wurden und verknüpft diese durch ‚Transferfunktionen'. Dafür gibt es verschiedene statistische Modelle.

Bei diesem Projekt soll eine empirisch statistische Methode, die Analog Methode, auf täglicher Basis für die Vergangenheit berechnet und im Anschluss mit einem Beobachtungsdatensatz validiert werden.

2 Daten Download

2.1 Verwendete Daten

Für dieses Projekt wurden die japanischen 55-Jahres Reanalysen (JRA-55) von der Japan Meteorological Agency verwendet. Dafür wurden vom 6-stündiger Datensatz von 31.12.1957 2100 UTC bis 01.01.2019 0000 UTC die Variablen Mean Sea Level Pressure ('mslp'), relative Feuchte auf 700 hPa ('rh') und die spezifische Feuchte auf 700 hPa ('sh') untersucht.
Die räumliche Auflösung der Daten beträgt 1.25 °.

2.2 Ausschnitt

plot_Ausschnitt.py

Zur Auswertung wurde ein Europaausschnitt gewählt, der die Koordinaten $\lambda = 10°W, 25°E$ und $\phi = 32.5°N, 67.5°N$ umfasst. Der gewählte Ausschnitt ist in Abbildung 1 farbig für den Parameter 'mslp' gekennzeichnet

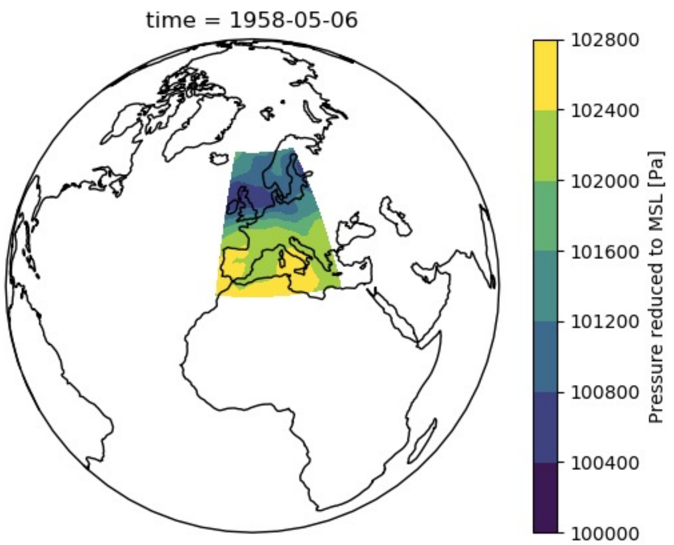

Abbildung 1: Gewählter Ausschnitt.

Hier war zu beachten, dass der Datensatz nur aufsteigende Longitudewerte zulässt. Um die 350° bis 25° aus dem Datensatz extrahieren zu können wurde daher folgende Formel zur Berechnung der erforderlichen Longitudewerte verwendet:

$$(\lambda + 180)\%360 - 180$$

, wobei % der Modulooperator ist, der den Rest einer Division ausgibt.

3 Preprocessing

anomalien_normiert.py

3.1 Mittelwert, Standardabweichung und Anomalien

Zuerst wurde aus den 6-stündigen Werten für jeden Parameter ein Tagesmittelwert erstellt.
Um das Auswahlfenster an Tagen zu erhöhen, wird nicht nur ein Tag berücksichtigt, sondern jeder Tag des Jahres mit einem Fenter von ± zehn Tagen versehen. Dies geschiet durch die *construct* Funktion von XArray in Python. Durch die Anwendung von *construct* wird dem DataArray eine Dimension hinzufügt, welche die Fenstergröße wiedergibt und in den weiteren Rechenschritten benutzt werden kann.
Die Berechnung von Mittelwert und Standardabweichung erfolgt über das Auswahlfenster von 21 Tagen. Daraus resultieren für ein Jahr 365 Mittelwerte für jeden Gitterpunkt.
Die Standardabweichung wird über das Auswahlfenster und zusätzlich über alle Gitterpunkte berechnet. Wird die Standardabweichung nicht für die gesamte räumliche Dimension berechnet, sondern für jeden einzelnen Gitterpunkt, so geht die Variabilität der einzelnen Gitterpunkte verloren. Diese Variabilität soll aber erhalten werden.
Um die Anomalien zu erhalten, wird vom aktuellen Tag der Mittelwert über das 21-tägige Auswahlfenster subtrahiert.

3.2 Normierte Anomlien

In weiterer Folge werden drei unterschiedliche Paramter mit unterschiedlichen Größenordnungen behandelt. Um numerische und physikalische Probleme - wie eine schlecht konditionierte Matrix oder ein Feld mit kleinere Größenordnung liefert keinen Einfluss - zu vermeiden werden die Anomalien normiert. Um die normierten Anomalien zu erhalten, werden die Anomalien durch die Standardabweichung des jeweiligen Tages dividiert.
In Abbildung 2 sind die normierten Anomalien für den Parameter 'mslp' für einen Beispielszeitraum dargestellt.

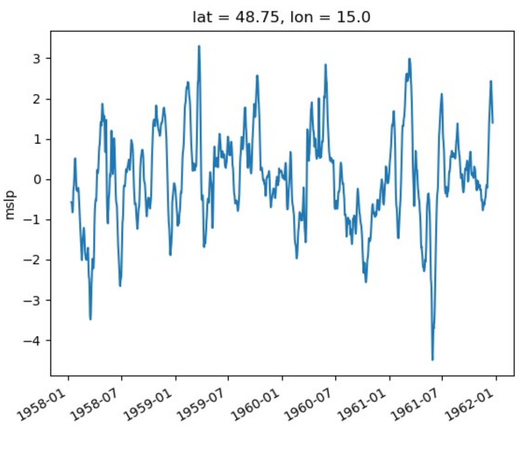

Abbildung 2: Normierte Anomalien für 48.75 ° Breite und 15.0 ° Länge.

4 Anwendung der Analog Methode

analog_methode.py

Für das Auffinden der Analoga werden die *Empirical Orthogonal Functions* (EOFs) herangezogen. Dadurch wird die Dimension der Datenmenge verringert, ohne dabei wichtige Informationen zu verlieren. Im Zuge der EOF Analyse erhält man die räumliche Information (EOF) und die dazugehörige zeitliche Information (Zeitkoeffizienten), die sogenannten *Principal Components* (PCs). Die Analoga werden dann im EOF- Raum gesucht, der von so vielen EOFs aufgespannt wird, wie nötig sind um 90% der Varianz zu erklären.

4.1 Vorbereitungen für die Analog Methode

Bevor mit der Durchführung der Analog Methode begonnen werden kann, werden einige Einstellungen getroffen. Es wird die Anzahl festgelegt, wie viele Analoga bestimmt werden sollen - in diesem Fall werden 4 Analoga gesucht. Es wird zudem ein Beispieljahr bestimmt, welches jedoch nicht zu verwechseln ist mit dem Jahr für welches Analoga gesucht werden. Das ausgewählte Jahr zu Beginn, dient lediglich zur Begrenzung der Schleife. Aus diesem Grund gibt es auch Probleme, falls als Beispieljahr ein Schaltjahr gewählt wird. Es war uns nicht möglich dieses Problem sauber zu lösen, weshalb wir dieses *Problem* ignorierten indem wir keines als Beispieljahr wählten.

4.2 Berechnung der EOFs

Zuerst wird das Zeitfenster festgelegt (21 Tage), für welches die EOFs berechnet werden sollen. In einem weiteren Schritt werden die Daten ausgeschnitten, die sich in diesem Zeitfenster befinden. Diesen Daten werden in einen *IrisCube* umgewandelt die EOF Berechnung kann durchgeführt werden.
Die Berechnung der *Empirical Orthogonal Functions* (EOF) erfolgt für jeden Tag jedes Jahres unter der Brücksichtigung des Auswahlfensters wie bei der Mittelwertbildung. Es wird für jeden Tag mit dem Zeitintervall von ± zehn Tagen eine multivariate EOF-Analyse durchgeführt. Dabei wird die vordefinierte *MultivariateEof* Python Funktion für Iris verwendet, welche nur für *IrisCubes* möglich ist. Die *MultivariateEof* Funktion gibt einen sogenannten *solver* zurück, welcher in weiterer Folge zur Erklärung der Varianz und der Berechnung der PCs verwendet wird.
Jede EOF erklärt einen gewissen Anteil der Gesamtvarianz. Es sollen so lange EOFs berechnet werden, bis eine Gesamtvarianz von 90 % erklärt wird. Dies geschieht über *solver.varianceFraction* in Verbindung mit einer Schleife, welche die Anzahl der benötigten EOFs zurückgibt. Somit kann mittels *solver.eofs* die notwendige Anzahl an EOFs berechnet werden.
Jeder EOF an einem Tag des Jahres ist eine PC in jedem Jahr zugeordnet. Diese bekommt man über *solver.pcs.*
Multipliziert man alle EOFs mit den zugehörigen PCs, so kann jeder Tag in der Vergangneheit reproduziert werden. Die resultierenden EOFs und PCs werden abgespeichert.
Es wird nun der aktuelle Tag ausgeschnitten, in einen *IrisCube* transformiert und für diesen werden die Pseudo-PCs berechnet. Diese Pseudo-PCs bekommt man durch Anwendung von *solver.projectField* auf den *IrisCube* des aktuellen Tages. Die berechneten PCs und Pseudo-PCs werden von einem *IrisCube* in einen *xarray.DataArray* rückverwandelt um die weiteren

Berechnung leichter durchzuführen. In Abbildung 3, Abbildung 4 und Abbilung 5 sind jeweils die ersten 3 EOFs der Parameter MSLP, RH und SH dargestellt.

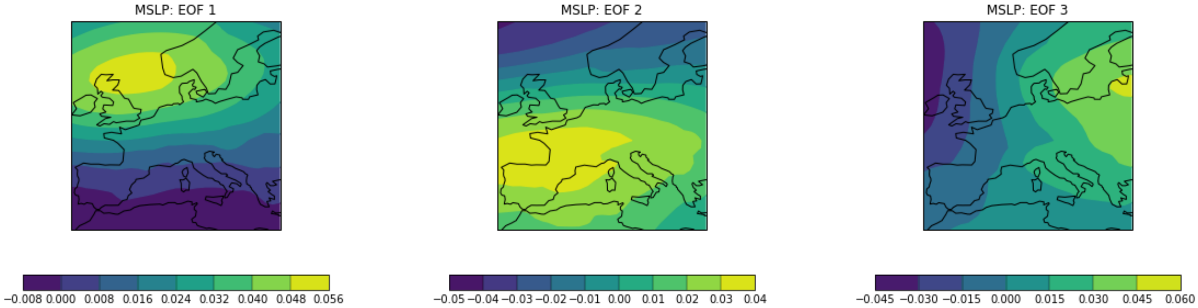

Abbildung 3: MSLP: Darstellung der ersten 3 EOFs

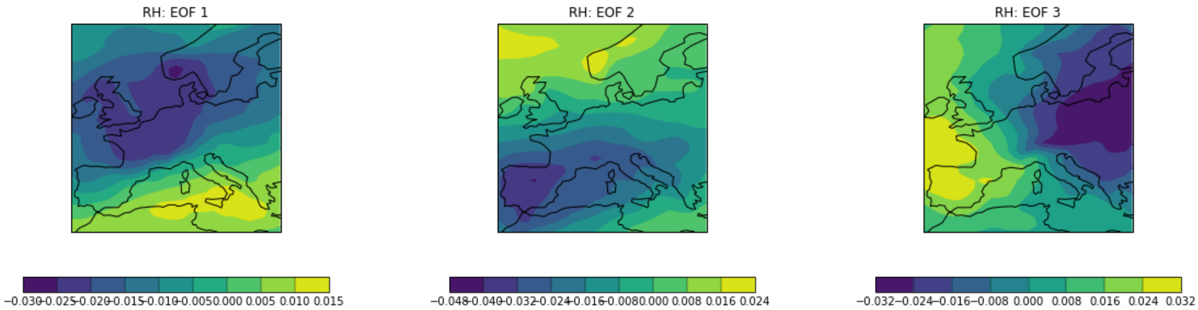

Abbildung 4: RH: Darstellung der ersten 3 EOFs

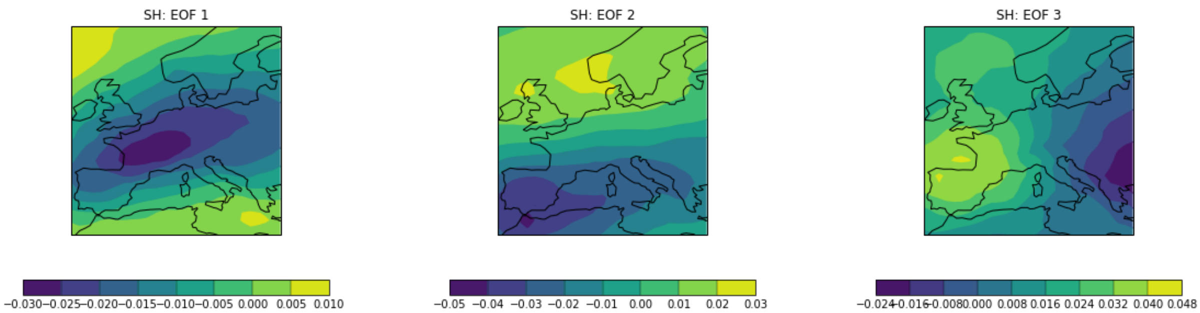

Abbildung 5: SH: Darstellung der ersten 3 EOFs

4.3 Berechnung der Norm

Für jeden Zeitpunkt der untersuchten Daten wird eine Norm zwischen den Pseudo-Pcs und den PCs berechnet. Bei diesem Projekt ergeben sich konkret aus 4 EOFs, 4 Pseudo-PCs. Diese 4 Pseudo-Pcs werden jeweils von den PCs abgezogen, quadriert und die Terme aufsummiert. Das ergibt für jeden Zeitpunkt in den PCs eine Norm (Dimension: Tag $\pm$ zehn Tage x Anzahl der Jahre).

4.4 Suchen des Minimums

Als nächster Schritt wird das Minimum der Norm gesucht. Da die Norm im selben Jahr am minimalsten ist, und wir ja Werte für dieses Jahr suchen wollen, wird das jeweilige betrachtete Jahr vor der Minimumssuche ausgeschlossen. Das Minimum stellt dann das 1.Analogon dar. Im Anschluss wird das nächste Minima gesucht, wobei hier zusätzlich das bereits gefundene 1. Analogon wieder ausgeschlossen wird. Das zweite Minimum ergibt dann das 2.Analogon usw. Das Verfahren wird für alle Normen berechnet und ergibt wie zu Beginn festgelegt 4 Analoga. Diese werden mit dem Datum des gesuchten Tages als Index abgespeichert.

5 Validierung

validierung.py, korrelationskoeffizient.py und korrelationskoeffizient_plot.py

5.1 Daten

Zur Validierung wurde der SPARTACUS Datensatz von 1961-2017 der ZAMG zur Verfügung gestellt. Das ist ein gegitterter Beobachtungsdatensatz für ganz Österreich mit einer räumlichen Auflösung von 1x1 km und einer zeitlichen Auflösung von 1 Tag. Der Datensatz umfasst die Parameter Minimum und Maximum der Lufttemperatur sowie die Niederschlagssumme. Zudem ist neben der räumlichen Verteilung auch ein Flächenmittel über die gesamte Region sowie Flächenminima und Flächenmaxima enthalten. Zur Vereinfachung wurde bei der Validierung nur das Flächenmittel berücksichtigt (d.h. jeweils ein Wert pro Tag für ganz Österreich).

Da der Spartacus-Datensatz nicht für die gesamte JRA-Zeitreihe vorhanden ist, wurden die Analoga für 1958-1960, sowie 2018 nicht bei der Validierung verwendet.

5.2 Vorgehensweise

Bei der Validierung soll die Analog-Methode in der Vergangenheit angewendet und getestet werden und der Fehler im Vergleich zum richtigen (beobachteten) Tag berechnet werden. Dabei wurden die vier berechneten Analoga für jeden Tag im Spartacus-Datensatz gesucht und ausgegeben. Danach wurde der Fehler im Vergleich zum wahren beobachteten Tag berechnet. Als Fehlermaß wurde der root-mean-square-error verwendet:

$$rmse = \sqrt{mean((x - x_{ref})^2)}$$

Er gibt an, wie stark eine Prognose im Durchschnitt von den beobachteten Daten abweicht. Je größer der RMSE ist, desto schlechter ist die Anpassung des Modells.

In Abbildung 6 ist der RMSE der ersten vier Analoga für nur einem Jahr (1961) für die Maximumtemperatur dargestellt.

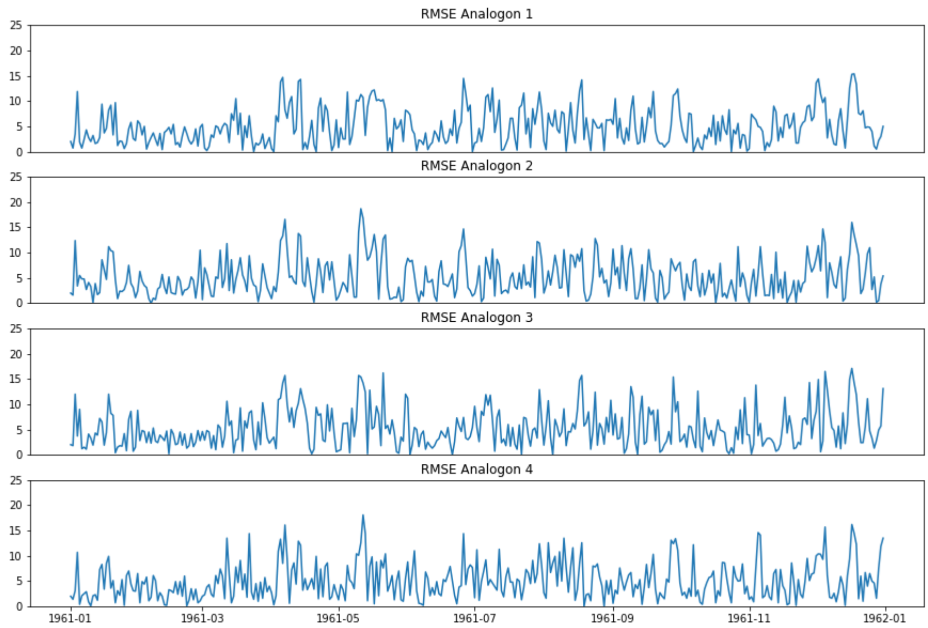

Abbildung 6: RMSE für Temperaturmaximum für 1961.

Da in Abbildung 6 keine qulitativen Aussagen über die gefundenen Analoga getroffen werden können, wird auch der Korrelationskoeffizient nach Pearson berechnet. Dies geschiegt durch *numpy.corrcoef*, wodruch der Korrelationskoeffizient mittels folgender Formel berechnet wird:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt(Var(X)Var(Y))}$$

, wobei X und Y die Werte der gefunden Tage und der echten Tage sind.
Der Korrelationskoeffizient weist Werte zwischen -1 und 1 auf. Wobei bei 1 die höchste Korrelation vorliegt. Werte um 1 deuten darauf hin, dass ein Zusammenhang zwischen den Parametern besteht.

In Abbildung 7 ist der Korrelationskoeffizient der ersten vier Analoga für ein Jahr (1961) für die Maximaltemperatur dargestellt.

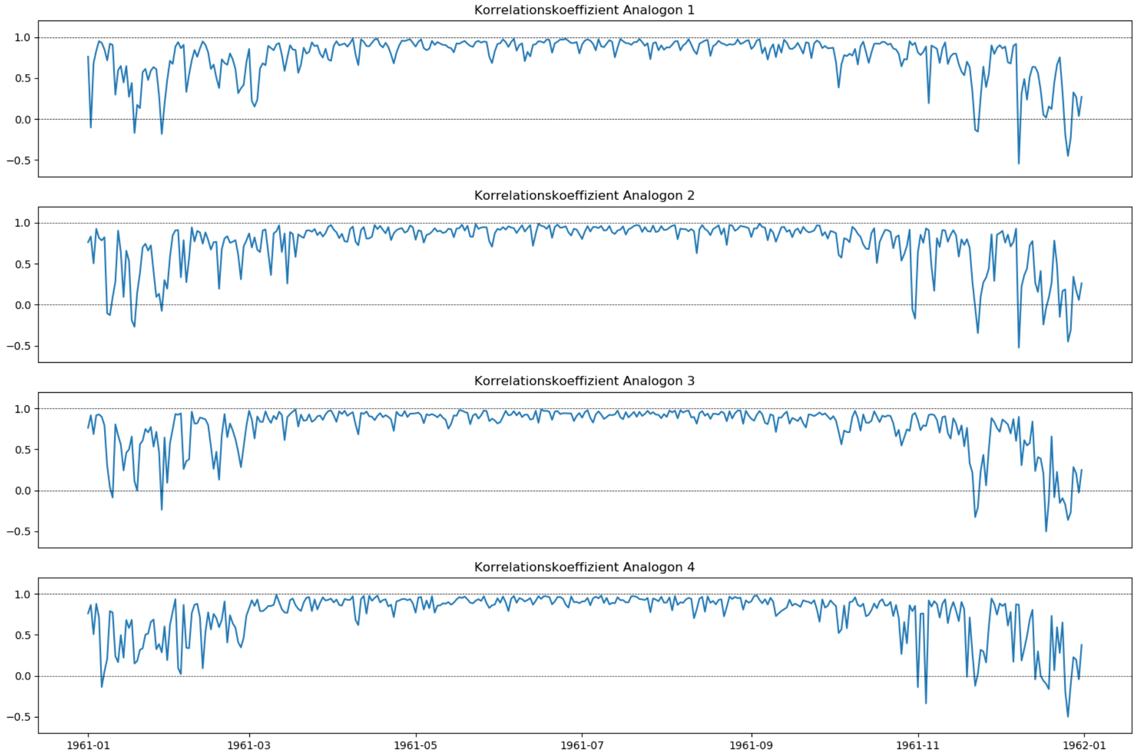

Abbildung 7: Korrelationskoeffizient für Temperaturmaximum für 1961.

In Abbildung 7 kann man eine sehr gute Korrelation von März bis in etwa Oktober erkennen. Im Winter ist die Korrelation nicht so gut gegeben. Dies lässt darauf schließen, dass eine Durchführung der Analogmethode mit diesen Einstellung gut im Sommer verwendet werden kann. Für eine Verwendung im Winter sollte man die Einstellungen verändern, wie zum Beispiel eine Erhöhung der durch die EOFs zu erklärenden Varianz.

Quellen

Japan Meteorological Agency/Japan. 2013, updated monthly. JRA-55: Japanese 55-year Reanalysis, Daily 3-Hourly and 6-Hourly Data. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D6HH6H41. Accessed 25.07.2019

https://www.zamg.ac.at/cms/de/forschung/klima/klimatografien/spartacus, 29.07.2019

https://de.statista.com/statistik/lexikon/definition/303/root$_m ean_s quare_e rror/$, 29.07.2019

https://www.crashkurs-statistik.de/der-korrelationskoeffizient-nach-pearson/, 30.07.2019