Projektarbeit Klimamodellierung

Nikta Madjdi und Jana Löffelmann

Sommersemester 2019

Motivation

Mittels Global Circulation Models (GCMs) kann das Klima und die globale Zirkulation hinsichtlich verschieden Einflussfaktoren analysiert werden. Regional betrachtet, ist die Auflösung der GCMs jedoch zu grob, um die Auswirkung bestimmter Einflüsse auf lokaler Skala wiederzugeben. In dieser Arbeit wird die "Analog Method" (siehe z.B. Zorita and Von Storch [1999]) verwendet, um die großskalige Zirkulation, welche über die GCMs simuliert wird, mit historischen Beobachtungen zu vergleichen, wodurch ein Zusammenhang zwischen der large-scale (LASC) und der local-scale (LOSC) hergestellt wird.
Reanalysedaten werden hier als Basis der Untersuchungen verwendet und "Empirical Orthogonal Functions" dienen der Komprimierung der Daten. Die einzelnen Schritte, welche zur Beschaffung der Datensätze sowie für die Durchführung der Analog Method notwendig sind, werden im Folgenden genauer erläutert. Am Ende werden die mit dieser Methode gefundenen analogen Tage mithilfe des Spartacus-Datensatzes der ZAMG für die Maximumtemperatur in Österreich für ein exemplarisches Jahr validiert.
Die Programme inklusive kurzer Beschreibung, welche für die Ergebnisse dieser Arbeit verwendet wurden, sind in einem GitHub-Repositorium unter folgendem Link `https://github.com/jane320/Klima_EOF` zu finden und basieren auf der Programmiersprache *Python*.

Datensätze

Als Grundlage für die EOF-Analyse und alle weiteren Schritte dienen die Reanalysedaten des National Centers for Environmental Prediction (NCEP). Über den ftp-Server *ftp.cdc.noaa.gov* können die Daten für den Bodendruck, der relativen Feuchte sowie der spezifischen Feuchte für den vorhandenen Zeitraum von 1948 bis 2018 heruntergeladen werden. Der genau Pfad, bei dem die nc-Dateien zu finden sind, lautet /Projects/Datasets/ncep.reanalysis.dailyavgs. Die Parameter sind bereits Tagesmittelwerte und pro Jahr ist eine nc-Datei vorhanden. Somit enthält eine Datei für das entsprechende Jahr 365 bzw. 366 Tageswerte einer Größe für jeden Breiten- und Längengrad. Die Auflösung beträgt 2.5°.
Um den Datensatz zu komprimieren, wird in den Files für die relative und spezifische Feuchte die 700 hPa Fläche im vornherein ausgeschnitten. Somit beläuft sich die Datenmenge auf ca. 1.7 GB. Aus diesem Grund und der Einfachheit halber lässt sich der Datensatz direkt auf dem eigenen Computer speichern.
Weiterhin dient zur Validierung der Analog Method der Spartacus-Datensatz der Zentralanstalt für Meteorologie und Geodynamik (ZAMG). Dieser Datensatz besteht aus gegitterten Beobachtungen von Lufttemperatur und Niederschlagssumme auf täglicher Basis mit einer Auflösung von 1x1 km. Die Daten liegen seit 1961 für ganz Österreich vor und werden stetig mit neuen

täglichen Werten aktualisiert. Der Spartacus-Datensatz für die Jahre 1961 bis 2017 wurde uns freundlicherweise über eine Cloud zur Verfügung gestellt und lies sich daher einfach herunterladen und auf dem eigenen Computer abspeichern.

Analog Method

In dieser Projektarbeit interessiert man sich für das regionalskalige Klima, sprich für Europa bzw. Österreich, welches aus den GCMs abgeleitet werden soll. Daher ist ein Downskalingverfahren notwendig, welches sich in zwei Methoden unterteilen lässt, das dynamische und das statistische Downskaling.
Hier wird eine statistische Downskalingmethode angewendet, deren Ziel es ist, eine sogenannte Transferfunktion zwischen den LASC-Prädiktoren und den LOSC-Prädiktanden herzustellen und auf die Outputs der GCMs für den betrachteten Zeitraum anzuwenden.
Die "Analog Method" zählt zu den statistischen Methoden. Dabei wird die globale Zirkulation, welche mit den GCMs modelliert werden, mit historischen Messungen verglichen, wobei die ähnlichste Beobachtung als Analogon ausgewählt wird. Dafür sind lange Zeitreihen von Beobachtungen notwendig um korrespondierende Analoga zu finden [Zorita and Von Storch, 1999].
Für die Berechnung der Analoga werden dabei Vereinfachungen angewendet. Zum einen ist die räumliche Ausdehnung beschränkt auf die relevante Region, zum anderen wird Hintergrundrauschen durch die EOF-Analyse gefiltert, auf welche im nächsten Abschnitt näher eingegangen wird. Außerdem wird keine Prognose mit dieser Methode durchgeführt, sondern ein lokalskaliger Zustand wird einem dazu passenden großskaligen Zustand zugeordnet.
Das Ziel ist es, für jeden Tag im Jahr für die gesamte Zeitreihe Analoga zu finden. Um dies zu erreichen sind folgende Schritte notwendig:
Zuerst muss das sogenannte Preprocessing durchgeführt werden, damit die Daten für weitere Berechnungen weiterverwendbar sind. Als nächstes werden Anomalien für die betrachteten Tage berechnet sowie anschließend EOF-Analysen durchgeführt um schließlich die Analoga aufzufinden. Jeder der einzelnen Schritte wird in den folgenden Abschnitten detaillierter erläutert.

1. Preprocessing

Der Datensatz, welcher im netCDF-Format vorliegt, wird mittels *xarray* geöffnet. Dafür wird der Befehl `xarray.open_mfdataset` verwendet, der aus allen Jahren und den drei Variablen (Bodenluftdruck, relative und spezifische Feuchte) einen gebündelten Datensatz erstellt.
Da die EOF-Analyse nicht global durchgeführt werden soll, sondern sich nur auf einen bestimmten geografischen Bereich konzentriert, kann der Datensatz auf diesen Bereich reduziert werden, was am Ende Rechenzeit spart. Der Ausschnitt erstreckt sich von 10° W bis 25° E und von 32.5° N bis 67.5° N.
Das Problem, das an dieser Stelle auftritt, ist, dass die Längengrade ab dem Meridian nach Osten von 0° bis 360° definiert sind. Da die zu betrachtende Region westlich sowie östlich des Meridians liegt, kann diese nicht einfach aus dem Datensatz ausgeschnitten werden. Gelöst wird dieses Problem, indem zu allen Längen 180 addiert wird, das Array mittels dem Modulo-Operators % durch 360 geteilt und anschließend wieder 180 subtrahiert wird. Man erhält dadurch für die Längengrade ein Array, welches beim Meridian 0° aufweist, nach Osten bis 180° läuft und nach Westen bis −180°. Somit lässt sich der Bereich von −10° bis 25° einfach ausschneiden und der Datensatz kann auf die vorgegeben Region reduziert werden.
Weiterhin wird über `xarray.DataArray.squeeze` und `xarray.Dataset.drop` der die Dimension "levels" entfernt, die in den Daten der relativen und spezifischen Feuchte aufgrund der

zuvor erwähnten Reduzierung auf die 700 hPa-Fläche noch vorhanden ist und die Dimension 1 besitzt. Das gleiche passiert mit der Variable "time_bnds", welche in einigen späteren Files auftaucht. Am Ende des Preprocessings enthält der Datensatz die Dimensionen "lat", "lon" und "time" für die drei Variablen und die 71 Jahre.

2. Berechnung der Anomalien

Die Berechnung der Anomalien geschieht auf täglicher Basis. Dabei wird nicht nur der gewünschte Tag oder "Target Day" (TD) betrachtet sondern die jeweiligen Werte aus den 10 Tagen vor sowie nach dem TD miteinbezogen.
Zu Beginn muss der für jeden TD ± 10 Tage der Mittelwert über alle vorhandene Jahre und für jeden Gitterpunkt berechnet werden. Somit erhält man am Ende 365 Mittelwerte für jeden Gitterpunkt. Um den Pool an Tagen für jeden TD zu erstellen, wird der Befehl `pandas.DataFrame.rolling` verwendet mit einer Fenstergröße von 21 Tagen und dem TD in der Mitte des Fensters. Für jeden TD und dessen "Rolling Window" wird anschließend der Mittelwert und die Standardabweichung berechnet. Zu beachten ist dabei, dass die ersten und letzten zehn Tage des Datensatzes fehlende Werte in dem zugehörigen Rolling Window aufweisen. Diese Tage werden jedoch trotzdem in die Berechnung der Anomalien miteinbezogen. Weiterhin wird die Standardabweichung nicht für jeden Gitterpunkt einzeln sondern über die gesamte räumliche Dimension berechnet, da andernfalls die Variabilität der Werte geglättet wird. Die im nächsten Schritt errechneten Anomalien aller drei Größen müssen danach noch normiert werden, um aufgrund der unterschiedlichen Größenordnungen von Druck, relativer und spezifischer Feuchte, Problemen, die bei der Weiterverarbeitung der Daten auftreten können, vorzubeugen. Dies geschieht, indem die Anomalien durch die zuvor bestimmte Standardabweichung geteilt werden. Am Ende erhält man für jeden Tag des Jahres, jeden Gitterpunkt und alle Jahre die normierten Anomalien, welche im nächsten Abschnitt für die Berechnung der EOFs verwendet werden.

3. EOF-Analyse

Um die Dimension großer Datensätze zu reduzieren ohne dabei viele Informationen in den Daten zu verlieren, werden "Empirical Orthogonal Functions" (EOFs) verwendet. Die Grundidee dabei ist, ein kontinuierliches Feld in Raum und Zeit - zum Beispiel den Bodenluftdruck - zu zerlegen, um so mit möglichst wenigen Eigenvektoren die Varianz des eigentlichen Feldes zu beschreiben [Hannachi et al., 2007]. Für die Berechnung von EOFs eines sogenannten "Single Fields" - also eines Feldes, das nur die Variabilität einer skalaren Größe (z.B. Druck) beschreibt - wird zu Beginn das Anomaliefeld (siehe Abschnitt) sowie anschließend die Kovarianzmatrix gebildet [Bjomsson and Venegas, 1997]. Mit dieser Matrix wird das Eigenwertproblem gelöst, wodurch man die Eigenwerte und die dazugehörigen Eigenvektoren erhält. Die Eigenvektoren sind so gewählt, dass sie orthogonal zueinander sind und die EOFs repräsentieren. Die Eigenwerte sind ein Maß für die anteilige Größe der Varianz eines EOFs an der totalen Varianz der Kovarianzmatrix. Somit werden die EOFs nach der Größe der Eigenwerte sortiert und gewichtet, was bedeutet, dass das erste EOF in Verbindung mit dem größten Eigenwert und damit der größten Varianz gebracht wird.
Zusätzlich zu den EOFs werden die "Principal Components" (PCs) bestimmt, die die Projektionen des Feldes auf die EOFs darstellen. Gleichung 1 beschreibt den Zusammenhang zwischen dem Feld F, den PCs $\vec{a_j}$ und den EOFs:

$$F = \sum_{j=1}^{n} \vec{a_j}(EOF_j) \tag{1}$$

Die PCs sind unkorreliert und jeder Vektor $\vec{a_j}$ beschreibt die zeitliche Entwicklung für das j-te EOF. Die Anzahl der EOFs wird mit n festgelegt. Da meist die ersten n Eigenvektoren die Dynamik des Systems beinhalten und die kleineren Eigenwerte hauptsächlich durch Rauschen entstehen [Bjomsson and Venegas, 1997], reicht es aus, nur die ersten EOFs mit den größten Eigenwerten zu betrachten, wodurch die Dimension des Datensatzes stark reduziert wird.
Die Umsetzung der EOF-Analyse erfolgt hier mit den normalisierten Anomalien des Reanalysedatensatzes von 1979 bis 2017 und dem Package `eof.multivariate.standard.MultivariatEof`. Da in diesem Fall nicht nur ein Feld sondern multiple Felder (Bodenluftdruck, relative und spezifische Feuchte) betrachtet werden, dienen zur Analyse multivariate EOFs. Wie zuvor bei der Bestimmung der Anomalien, berechnen sich die EOFs ebenfalls für jeden Tag des Jahres, alle Gitterpunkte sowie über ein Rolling Window mit einem Fenster von 21 Tagen. Außerdem wird die Zeitdimension der "Day of Years" (DOY) mit der des Rolling Windows zusammengelegt und in "time" umbenannt, da das EOF-Package diese sonst nicht erkennt. Die EOFs werden daraufhin für die drei Variablen erzeugt, wodurch man den sogenannten `solver` erhält. Dieser enthält unter anderem die Werte der PCs sowie die Eigenwerte zu jedem EOF.

4. Analoga finden

Um nun für jeden DOY die Analoga zu finden, wird über eine Schleife jeder Tag ausgewählt. Eine weitere Schleife über die TD sucht dann nach dem Tag, welcher dem gewünschten DOY am ähnlichsten ist. Dafür wird zuerst über die erklärte Varianz, die 90% nicht überschreiten soll, die Anzahl an relevanten EOFs berechnet. Die EOFs sowie die PCs werden anschließend für den TD bestimmt. Weiterhin werden die Reanalysedaten mit `projectField` des TD in den EOF-Raum projiziert, um die Zeitkoeffizienten bzw. Pseudo-PCs zu erhalten. Dadurch kann jedes Feld als Linearkombination aus den EOFs und den dazugehörigen Pseudo-PCs reproduziert werden.
Über die Euklidische Norm zwischen den PCs und den Pseudo-PCs und die Anzahl der n EOFs, wird verglichen, wie ähnlich sich der DOY und der TD sind. Das Analogon erhält man, in dem man das Minimum der Norm bestimmt:

$$\sum_{k=1}^{n} \sqrt{(pc - pseudo\_pc(t))^2} \Rightarrow min \tag{2}$$

Da die ersten zehn Tage des betrachteten Zeitfensters erneut aufgrund des Rolling Windows eine geringere Dimension aufweisen als die nachfolgenden Tage, müssen die fehlende Werte ergänzt werden, damit auch diese in die Berechnung der Norm mit einfließen können. Außerdem wird der Wert des gesuchten Jahres ausgeschnitten, sodass das Analogon des DOY nicht mit dem DOY zusammenfällt.
Mit Hilfe der Funktion `numpy.nanargmin` wird dann der Index der minimalsten Norm gefunden, welche dem korrespondierenden Analogon für den DOY entspricht. Da die Berechnungen über "numpy arrays" durchgeführt wird, welche mit dem Befehl `.values` erzeugt werden, gehen die Informationen über den verwendeten Tag bzw. das Datum verloren. Um am Ende nicht nur eine Liste an Indizes für jeden Tag und dessen Analogon zu bekommen, wird die Funktion `get_Date_from_index` definiert, welche zu dem gewünschten Index das Datum erstellt.
Um nicht nur einen Analog-Tag zu bestimmen, kann durch ersetzen des Index der minimalsten Norm mit NaN die zweit-, dritt- etc. kleinste Norm und daher weitere analoge Tage gefunden werden. Hier werden insgesamt fünf Analoga pro DOY bestimmt.

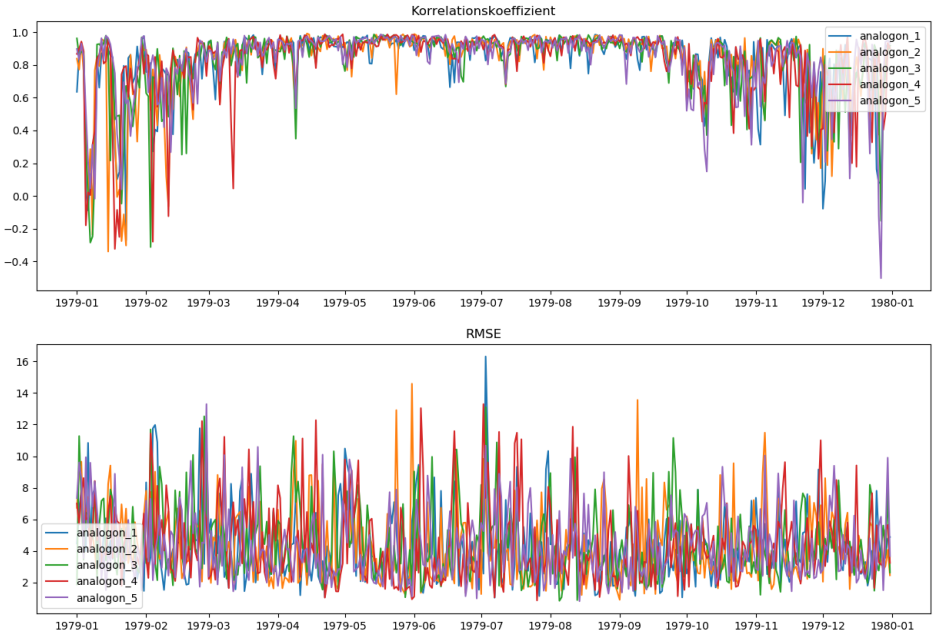

Abbildung 1: Korrelationskoeffizient und RMSE für das Jahr 1979.

Validierung

Zur Validierung wird nun der Spartacus-Datensatz herangezogen. Dabei wird der gewünschte Tag aus diesem Datensatz mit den dazu gefundenen analogen Tagen verglichen, indem sowohl der Korrelationskoeffizient als auch der Root Mean Squared Error (RMSE) berechnet werden. Aufgrund der langen Rechenzeit, welche der Funktion `numpy.corrcoef` verschuldet ist, wird nur das Jahr 1979 validiert.
In Abb. 1 ist der Korrelationskoeffizient und der RMSE für das Jahr 1979 für die Maximumtemperatur dargestellt. Dabei lässt sich erkennen, dass die Ergebnisse des Korrelationskoeffizienten im Frühling und Sommer am nächsten an 1 liegen im Vergleich zum Herbst und Winter, was für einen guten Zusammenhang zwischen dem TD und den Analogons spricht. Jedoch spiegelt sich dies nicht im RMSE wieder. Dieser ist nämlich ziemlich gleichmäßig über das Jahr verteilt. Es lässt sich vermuten, dass es in der kalten Jahreshälfte schwieriger ist, passende Analoga zu finden als in der warmen Jahreszeit und dass dies nicht durch weitere gefundene Analoga verbessert wird. Um eine genauere darüber Aussage zu treffen, müssten jedoch für weitere Jahre Analoga gefunden und analysiert werden. Bei näherer Betrachtung unterscheiden sich die verschiedenen Analoga im Winter deutlicher als im Sommer, was in Abb.2 und 3 erkennbar ist. Allerdings wird in dieser Darstellung nicht ersichtlich, welches Analogon den höchsten Korrelationskoeffizient bzw. den kleinsten RMSE über das gesamte Jahr aufweist.
Um nun sehen zu können, welche analogen Tage einem speziellen TD zugeordnet werden, wird die Maximumtemperatur exemplarisch für den TD am 11. Jänner und den 30. Juni 1979 und dessen dazugehörigen Analoga in Abb. 4 und 5 dargestellt.
Man erkennt, dass der erste analoge Tag in beiden ausgewählten Beispielen der Temperaturverteilung des TD sehr ähnlich ist. Außerdem fällt die Unterschiedlichkeit der Analoga untereinander auf, obwohl sie eigentlich alle den TD am besten beschreiben sollten. Im Winter sind

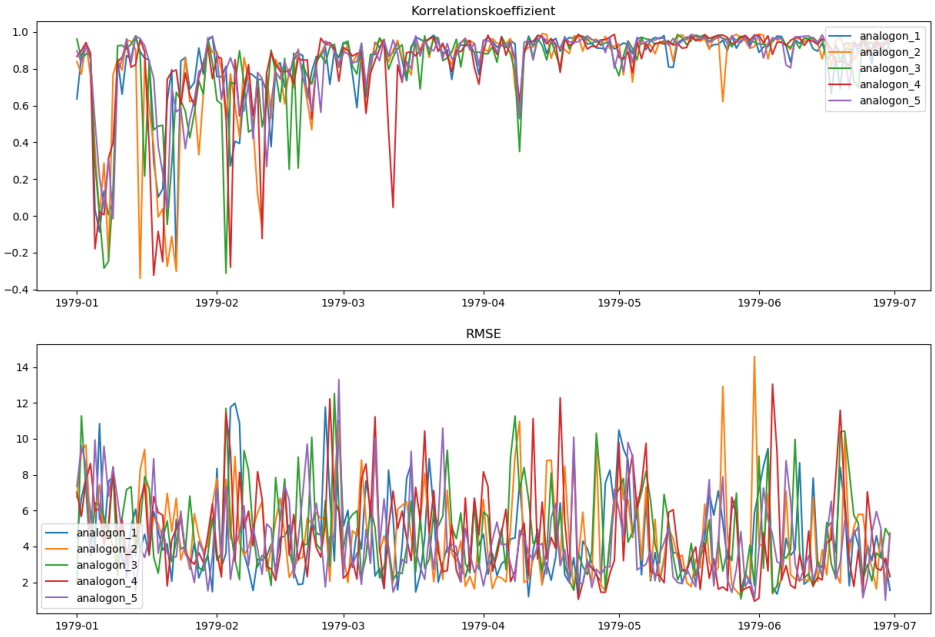

Abbildung 2: Korrelationskoeffizient und RMSE für 1. Jänner 1979 bis 30. Juni 1979.

Abbildung 3: Korrelationskoeffizient und RMSE für 1. Juli 1979 bis 31. Dezember 1979.

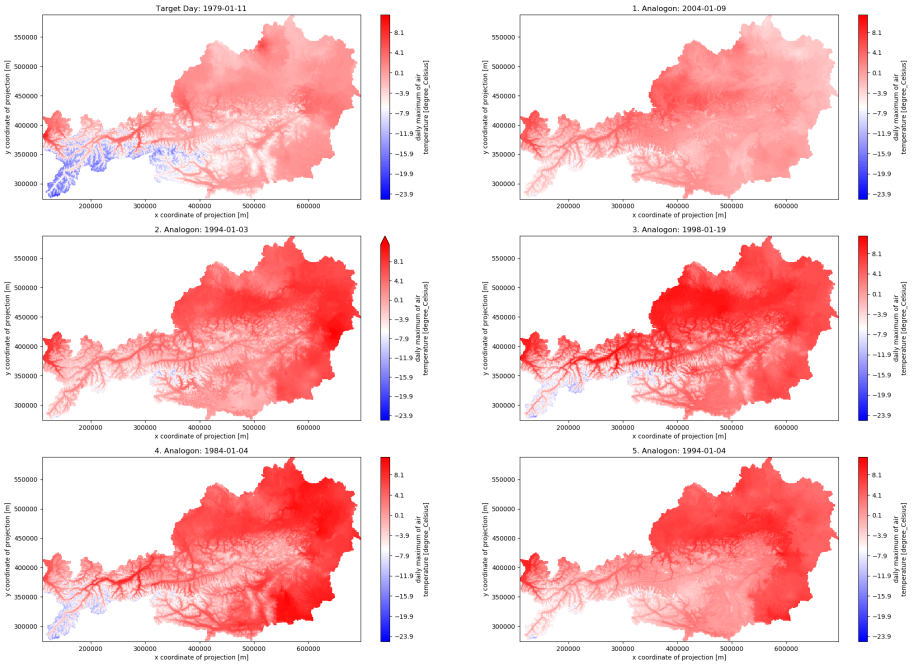

Abbildung 4: Target Day 11. Jänner 1979 (links oben) mit den ersten fünf Analoga.

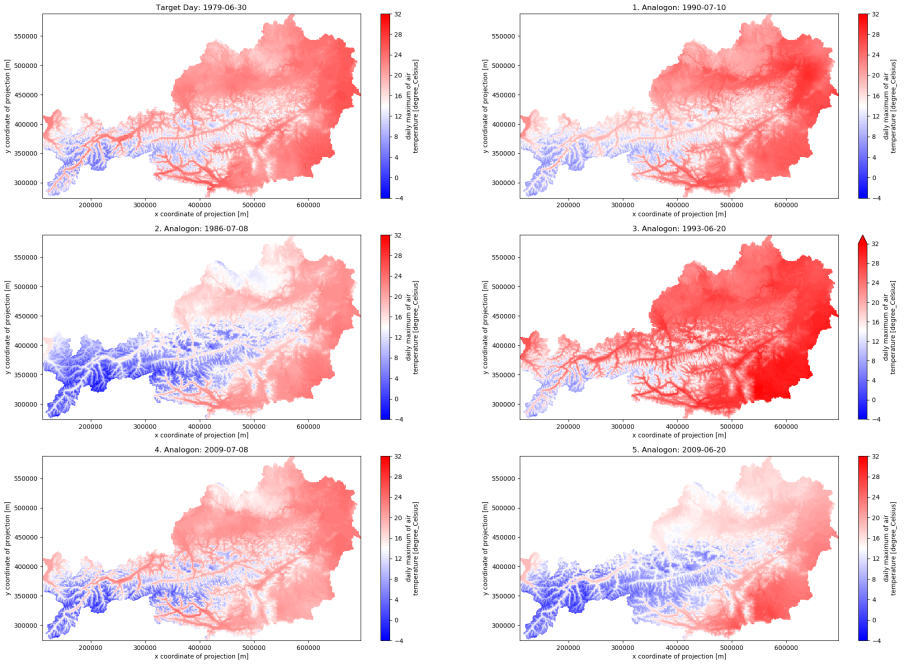

Abbildung 5: Target Day 30. Juni 1979 (links oben) mit den ersten fünf Analoga.

	Analog Nr.				
	1	2	3	4	5
Mittelwert	0.824	0.821	0.816	0.820	0.827
Median	0.904	0.904	0.899	0.894	0.898

Tabelle 1: Mittelwert und Median des Korrelationskoeffizienten für die Maximumtemperatur des Jahres 1979.

	Analog Nr.				
	1	2	3	4	5
Mittelwert	4.222	4.365	4.517	4.524	4.307
Median	3.693	3.888	3.936	3.887	3.946

Tabelle 2: Mittelwert und Median des RMSE für die Maximumtemperatur des Jahres 1979.

auch hier die Unterschiede markanter als im Sommer.
Wenn man nun den Mittelwert und Median der ersten, zweiten etc. Analoga miteinander vergleicht, korreliert - wie zu erwarten - der erste analoge Tag am besten mit dem TD und hat zusätzlich den geringsten Fehler von allen. Man sieht jedoch, dass die Unterschiede nur im Zehntel- oder Hundertstelbereich liegen (vgl. Tabelle 1 und 2).

Literatur

H Bjomsson and SA Venegas. A manual for eof and svdanalyses of climate data. *CCGCR Report. McGillUniversity, Montreal, Quebec, Canada*, pages 23–27, 1997.

A Hannachi, IT Jolliffe, and DB Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(9):1119–1152, 2007.

Eduardo Zorita and Hans Von Storch. The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of climate*, 12(8):2474–2489, 1999.