universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Untersuchung der Eigenschaften ausgewählter Verifikationsmaße anhand hochauflösender NWP Modellvorhersagen während MAP D-PHASE

Verfasserin

Sarah Umdasch

angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag.rer.nat.)

Wien, 2012

Studienkennzahl lt. Studienblatt	A 415
Studienrichtung lt. Studienblatt	Meteorologie
Betreuer:	O. Univ.-Prof. Dr. Reinhold Steinacker

Zusammenfassung

Die Kurz- und Kürzestfristvorhersagen meteorologischer Bodenparameter von sieben hochauflösenden numerischen Wettervorhersagemodellen für den Sommer und Herbst 2007 werden in dieser Arbeit mithilfe des Analysesystems Vienna Enhanced Resolution Analysis (VERA) verglichen und verifiziert.

Die Datenbasis stützt sich auf die internationalen Projekte COPS und MAP D-PHASE, aus denen sowohl ein sehr dichtes Beobachtungsnetz über Mitteleuropa als auch Vorhersagen aus den Modellketten verschiedener Wetterdienste zur Verfügung stehen. In dieser Arbeit wird eine Auswahl jener, teilweise ineinander eingebetteten Modelle behandelt: ALADIN-FR und AROME (Météo France), COSMO-7 und COSMO-2 (MeteoSchweiz), COSMO-EU und COSMO-DE (DWD) sowie CMC-GEM-L (EnvironmentCanada-CMC). Unterschiede zwischen den Modellfamilien, aber auch das verschiedene Verhalten von feinerem und gröberem Modell einer Kette können ausgemacht werden.

Als statistische Methoden werden „klassische“ (also nicht-räumliche, nicht-objektbasierte) Verifikationsansätze angewendet, die einen direkten Vergleich mehrerer Modelle erlauben. Zu diesem Zweck werden unter anderem *skill scores* ausgewertet, die die Verbesserung gegenüber einer Referenzprognose, wie einem langjährigen Klimamittel (MESOCLIM, eine 35-jährige mesoskalige Klimatologie über dem erweiterten Alpenraum) oder der Persistenz aufzeigen. Für *skill scores* binärer Ereignisse zeigt sich, dass eine Abwandlung ebendieser durch eine quantilsbasierte Schwellwert-Bestimmung speziell bei der Niederschlagsverifikation nützlich sein kann.

Herausforderungen an die Modelle stellt die komplexe Orographie im geographischen Gebiet der Untersuchungen, das einen Großteil der Alpen einschließt. Spezifischere Evaluierungen regionaler Unterschiede werden zusätzlich im kleineren „COPS-Gebiet“ von Vogesen, Rheintal und Schwarzwald durchgeführt. Die Ergebnisse zeigen zudem Unterschiede verschiedener Vorhersageaspekte im Tages- und Jahreszeitenverlauf, sowie für einzelne Kaltfrontereignisse auf.

Abstract

Seven high-resolution numerical weather prediction (NWP) models and their short-term forecasts of meteorological surface parameters in summer and fall 2007 are compared and verified against the analysis system VERA (Vienna Enhanced Resolution Analysis).

The data pool stems from the international projects COPS and MAP D-PHASE which provide both, a very dense observational network over Central Europe and forecasts obtained from different weather services' model chains. This work has a focus on a selection of these models, some of which are nested within each other: ALADIN-FR and AROME (Météo France), COSMO-7 and COSMO-2 (MeteoSwiss), COSMO-EU and COSMO-DE (DWD) as well as CMC-GEM-L (EnvironmentCanada-CMC). Differences between the model families, but also between the models of different resolution belonging to one chain could be identified.

In terms of statistical methods, "conventional" (that is to say, non-spatial, non-object-based) verification approaches which allow for a side-by-side comparison of several models are applied. For that purpose, among others, skill scores are calculated to evaluate the improvement over a reference forecast, such as long-term climatological means (MESOCLIM, a 35-years period mesoscale climatology over the greater Alpine region) or persistence. Regarding skill scores of binary events it becomes apparent that a modification of those using quantile-based threshold definitions can be useful, especially in precipitation verification.

Challenges for the models arise with the investigated geographical domain spanning most of the complex orography of the Alps. More specific evaluations of regional differences are carried out over the smaller COPS-domain covering the Vosges Mountains, the Rhine Valley and the Black Forest. Furthermore, results show diurnal and seasonal differences with respect to various features of the forecast as well as for individual frontal passages.

Inhaltsverzeichnis

1 Einleitung

Seit Beginn der numerischen Wettervorhersage (NWP, Numerical Weather Prediction) in den 1950er Jahren steht auch die Frage nach der adäquaten quantitativen Beurteilung der Qualität von produzierten Modellprognosen im Blickpunkt des Forschungs- und Anwendungsinteresses. Sowohl für die Weiterentwicklung der Prognosesysteme als auch für die Interpretation der ausgegebenen Vorhersagen, die zum Beispiel Entscheidungsprozesse in wetterabhängigen Situationen bestimmen, ist es wichtig, die Modelleigenschaften im Detail anhand vergangener Leistungen zu kennen.

Am Institut für Meteorologie und Geophysik der Universität Wien (IMGW) wird seit geraumer Zeit die Verifikation verschiedener Wettermodelle gegen das institutseigene Analysesystem VERA (Vienna Enhanced Resolution Analysis, Steinacker et al. 2000) vorangetrieben. Von verschiedenen Wetterdiensten wird oft ein Vergleich der Vorhersage mit der modelleigenen Analyse durchgeführt. Demgegenüber hat die Verwendung von VERA den Vorteil, dass durch die Modellunabhängigkeit keine systemeigenen Fehler in der Referenz das Ergebnis künstlich beeinflussen und daher eine direkte Nebeneinanderstellung verschiedener Vorhersagesysteme möglich ist.

Das Projekt VERITA (NWP model **VERI**fication over complex **T**errain with VER**A**, P20925, PI: M. Dorninger) intensivierte den Forschungsaufwand am IMGW in diese Richtung mit der Zusammenführung einer ausnehmend großen Datengrundlage aus internationalen Kollaborationen (vgl. Abschn. 2.1) und der Anwendung und (Weiter-)Entwicklung aktueller Methoden, wie objektbasierter Verifikation (Kiesenhofer 2011), Skalenzerlegung oder der Untersuchung des Einflusses des Analysefehlers auf das Verifikationsergebnis (Gorgas und Dorninger 2012b). Auch die vorliegende Arbeit ist Teil von VERITA und bringt zusätzlich zu jenen Auswertungen, welche teilweise eine schwer überschaubare Fülle von skalen-, orts- und fallabhängigen Resultaten erzeugen, traditionellere und somit oft zugänglichere Verifikationsmethoden zur Anwendung. Die Leistungen von sieben Modellen verschiedener nationaler Wetterdienste in der Kurzfristvorhersage mehrerer meteorologischer Bodenparameter sollen zusammengefasst und nebeneinander gestellt werden.

Obwohl Einfachheit und Übersicht in der Darstellung der Ergebnisse angestrebt werden, darf dies nicht auf Kosten eines differenzierten Blicks auf die unterschiedlichen Aspekte der Vorhersageeigenschaften, welcher nicht nur aus wissenschaftlicher Sicht, sondern auch für eine fundierte Anwendung von Modellprognosen unerlässlich ist, geschehen.

Von besonderem Interesse ist der Vergleich von unterschiedlich hoch aufgelösten Modellen aus einer Modellfamilie, deren Verschiedenheiten auch bei der Verifikation auf einem gemeinsamen Gitter in Erscheinung treten, wie sie in dieser Arbeit für drei Modellpaare durchgeführt wird. Nicht nur in der Vorhersage von hochreichender Konvektion, für welche nur die am höchsten aufgelösten der betrachteten Modelle ($< 3\,\mathrm{km}$ Gitterpunktsabstand) ohne Parametrisierungsschema auskommen, sondern auch in den übrigen meteorologischen Größen hat die horizontale Auflösung bestimmenden aber oft nicht problemlos zu bewertenden Einfluss auf das Ergebnis.

Eine besondere Herausforderung stellt das durch die Verfügbarkeit der Modellprognosen vorgegebene räumliche Gebiet für die Auswertungen dar, da es zu einem großen Teil aus dem orographisch komplexen Gelände der Alpen besteht (vgl. Abschn. 2.8), dessen zahlreiche, verschiedenartige und oftmals kleinskalige Auswirkungen auf die Atmosphäre von den Modellen, aber auch von einem Messnetz, schwer zu erfassen sind.

Einschränkungen im Anspruch an die Untersuchungen in dieser Arbeit ergeben sich bereits in den Grundlagen der Herangehensweise. So impliziert der gebräuchliche Ausdruck „Verifikation“ (von lat. *veritas* „Wahrheit“), dass die Modellprognosen mit einer bekannten, (bis auf einen vernachlässigbaren Messfehler) exakten Realität verglichen werden können. Da NWP Modelle jedoch keine Punktprognosen für die Orte der meteorologischen Messstationen, sondern Gitterpunktswerte ausgeben, die dessen Umgebung repräsentieren (vgl. Abschn. 3.7), wird zur Evaluierung eine modifizierte (da zumindest interpolierte) Version der Messwerte, eine Analyse, herangezogen. Trotz des Vorbehalts einer simulierten, möglicherweise fehlerbehafteten Wirklichkeit, werden im Folgenden die Ausdrücke „Verifikation“ und „Fehler“ wie in der Literatur gebräuchlich verwendet, auch wenn die Evaluierung gegen und die Abweichung von der verwendeten Analyse gemeint sind.

2 Die Datengrundlage und ihre Aufbereitung

2.1 Die Projekte COPS und MAP D-PHASE

Das „Forecast Demonstration"-Projekt (FDP) des WWRP (World Weather Research Programme) namens „Demonstration of Probabilistic Hydrological and Atmospheric Simulation of flood Events in the Alpine region" (D-PHASE) fand von Juni bis November 2007 statt (Rotach et al. 2009). Als Folgeprojekt des „Mesoscale Alpine Programme" (MAP, Bougeault et al. 2001) war es seine primäre Zielsetzung zu untersuchen, ob die bei MAP erhaltenen Forschungsergebnisse zu Fortschritten in der Vorhersage von extremen Wetterereignissen („*high-impact weather*") im Alpenraum geführt hatten. Es wurde ein ausnehmend großer Datensatz aus dem Output von mehr als 30 numerischen Wettervorhersagemodellen für den behandelten Zeitraum und das gewählte Gebiet über Mitteleuropa zusammengetragen. In einem zeitlichen und räumlichen Unterabschnitt des D-PHASE Projekts wurde zwischen Juni und August 2007 auch das WWRP „Research and Development"-Projekt (RDP) „Convective and Orographically induced Precipitation Study" (COPS, Wulfmeyer et al. 2011) durchgeführt. Die Zielsetzung von COPS beinhaltete, ein besseres Verständnis von konvektivem und orographisch induziertem Niederschlag zu erlangen. Für das gesamte Jahr 2007 wurden am IMGW meteorologische Beobachtungsdaten, die über die oft verwendeten, via GTS (Global Telecommunication System) transportierten Daten hinausgehen und aus unterschiedlichen operationellen Messnetzen vieler Länder Europas stammen, gesammelt und zum Joint D-PHASE/COPS (JDC) Datensatz vereint (Dorninger et al. 2009). Das JDC Netz enthält über tausend GTS Stationen und über 12 000 zusätzliche („non-GTS") Stationen, die in unterschiedlichem Ausmaß die gängigen meteorologischen Bodenparameter des GTS abdecken. Sowohl die Modellprognosen aus D-PHASE als auch der JDC Datensatz stehen im Archiv des World Data Center for Climate (WDCC) [1] zur Verfügung. Diese breite

[1] `http://cera-www.dkrz.de` (abgerufen am 08.03.2012).

Datenbasis bildet den Ausgangspunkt für vorliegende Arbeit sowie für zahlreiche publizierte Auswertungen, die teilweise über die ursprünglichen Hauptfragestellungen der zu Grunde liegenden Projekte hinaus gehen (z. B. Bauer et al. 2011, Eigenmann et al. 2011, Weusthoff et al. 2010 oder Zappa et al. 2008).

2.2 VERA: Vienna Enhanced Resolution Analysis

VERA (Vienna Enhanced Resolution Analysis, Steinacker et al. 2000) ist ein am IMGW entwickeltes Analysesystem, dessen Zweck es ist, die ungleichmäßig verteilten meteorologischen Bodenmessdaten im Flachland sowie in gebirgigen Regionen auf möglichst realistische Weise auf ein modellähnliches, regelmäßiges Gitter zu bringen. Die Interpolationsmethode von VERA basiert auf einem „*thin-plate-spline*“ Ansatz, bei dem die Krümmung und/oder der Gradient der skalaren Felder bzw. der kinematischen Größen der Vektorfelder minimiert wird.

Die großen Herausforderungen bei einem Analyse- (sowie bei einem Vorhersage-) System liegen in der Anwendbarkeit für komplexes Gelände, wo sich durch Phänomene wie dem verringerten Luftvolumen in Tälern gegenüber der Ebene, der Um- und Überströmungen von Gebirgen etc. sehr kleinskalige meteorologische Strukturen bilden, die nicht vom Messnetz aufgelöst werden können. Eine einfache horizontale Interpolation ist hier oft nicht sinnvoll. VERA bedient sich daher sogenannter „Fingerprints“ (Steinacker et al. 2006), um a priori Wissen über die Topographie und ihre typische physikalische Wirkung auf die Atmosphäre einzubinden. Charakteristische thermische und dynamische Druckgebilde im alpinen Gelände werden, sobald sie in den Daten erkannt werden, mit einer Gewichtung versehen und auf das vorhandene Feld aufgeprägt.

Um dem Einfließen von fehlerhaften Messwerten in die Analyse vorzubeugen, wird vor der Interpolation eine Qualitätskontrolle der Eingangsdaten durchgeführt, welche auf der Selbstkonsistenz der Daten basiert und sowohl systematische als auch einzelne grobe Fehler („*gross errors*“) eliminiert. In den hier verwendeten VERA Analysen wurde eine Vorgängerversion der bei Steinacker et al. (2011) beschriebenen Qualitätskontrolle VERA-QC angewendet, welche in ihren grundlegendsten Eigenschaften bereits mit der neueren Ausführung übereinstimmt.

Im Gegensatz zu den Modellfeldern ist das Gültigkeitsniveau der VERA Bodenanalysen nicht durch eine „reale“ (nur durch die Gitterauflösung abstrahierte) Topographie gegeben, sondern durch die sogenannte „Minimumtopographie“, welche die Höhe der alpinen Täler und Becken widerspiegelt. So liegt im betrachteten Gebiet,

das große Teile der Alpen umfasst, der höchste im verwendeten 8 km-Gitter aufgelöste Punkt der Minimumtopographie nur auf ca. 1512 m ü. NN (vgl. Abb. 2.4), während der reale höchste Gipfel der des Mont Blanc mit über 4800 m ü. NN ist. Der Unterschied zu den Topographien der sehr hochauflösenden Modelle kann an einzelnen Punkten bis über 2000 m ausmachen, was eine Anpassung besonders für die potentielle und äquivalentpotentielle Temperatur notwendig macht (vgl. Abschn. 2.5).

Ein für die Verifikation wichtiges Merkmal von VERA ist, dass es völlig unabhängig von Prognosemodellen betrieben wird und nicht wie manche andere Analyseverfahren zur ersten Abschätzung der Feldstrukturen („*first guess*“) eine Vorhersage heranzieht.

Die in dieser Arbeit verwendeten VERA Analysen sind mit einer Auflösung von 8 km über einem Gebiet von 1664 km × 1536 km mit dem Mittelpunkt der Tangentialebene bei 48.8° N, 7.0° O auf Basis des JDC Datensatzes bzw. nur der GTS Daten gerechnet. Es stehen Felder der potentiellen Temperatur θ, der äquivalentpotentiellen Temperatur θ_e, des auf Meeresniveau reduzierten Luftdruckes p_{msl}, der horizontalen Windkomponenten u und v, des ein-, drei- und sechsstündig akkumulierten Niederschlags RR und des Mischungsverhältnisses m zur Verfügung. Für alle angeführten meteorologischen Variablen bis auf das Mischungsverhältnis werden im Rahmen dieser Arbeit Auswertungen durchgeführt.

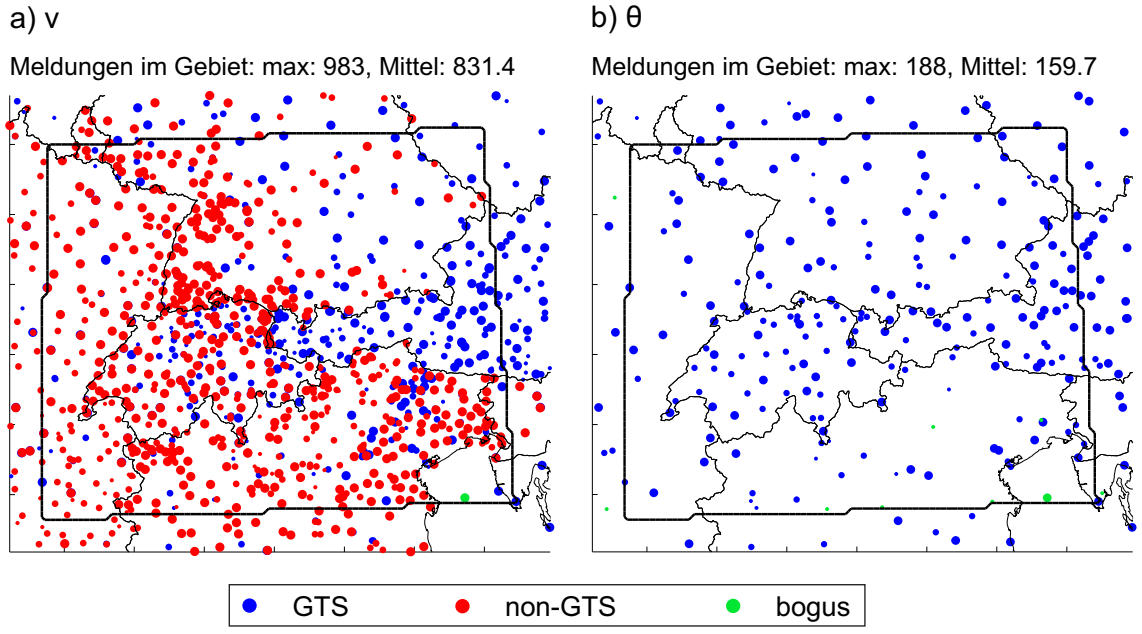

Abbildung 2.1: Die für die VERA Analysen verwendeten Stationsmeldungen für a) die meridionale Windkomponente und b) die potentielle Temperatur im Juni–August 2007, 03 UTC. Die Größe der Kreise deutet die Häufigkeit der Meldungen an: klein: Station meldete zu $< 50\,\%$ der Termine, mittel: 50–96.7 %, groß: $> 96.7\,\%$.

Obwohl die JDC Daten im einstündigen Abstand verfügbar sind, werden aufgrund einer signifikant niedrigeren Datendichte zu den Nebenterminen bevorzugt nur die dreistündlichen Analysen verwendet. Überdies wurden bei den Druckanalysen trotz Variation der Einstellungen der Qualitätskontrolle fallweise starke unrealistische Strukturen bei der Verwendung des extrem dichten JDC Datensatzes festgestellt. Aus diesem Grund werden für den Luftdruck und infolgedessen auch für die potentielle und äquivalentpotentielle Temperatur die Analysen auf Basis der GTS Stationen herangezogen. Es ist anzunehmen, bleibt jedoch zu zeigen, dass die Verwendung der mittlerweile vorliegenden neueren Version der VERA-QC und VERA bessere Ergebnisse bei den Druckanalysen auf Basis des dichten Datensatzes liefern würde. Abbildung 2.1 gibt Aufschluss über die Dichte der verwendeten Stationsmeldungen für v (JDC) und θ (nur GTS) im Sommer 2007 im betrachteten Gebiet.

2.3 Die verwendeten Modelldaten

Aus der großen Zahl der während MAP D-PHASE betriebenen hochauflösenden Prognosemodellen werden die sieben in Tabelle 2.1 zusammengefassten in die Auswertungen dieser Arbeit einbezogen. Die Auswahl beruht auf der Verfügbarkeit der Daten, denn ein adäquater Modellvergleich lässt sich nur bei Übereinstimmung von räumlichem Gebiet, Initialisierungs- und Laufzeit durchführen. Im Folgenden soll die Bezeichnung „Lauf“ für den Initialisierungszeitpunkt stehen, auch wenn im Allgemeinen der „Modelllauf“ eher den gesamten Berechnungsvorgang der Vorhersage beschreibt. Jedes der verwendeten Modelle wurde um 00 UTC initialisiert, alle COSMO Modelle stehen des Weiteren für den 12 UTC Lauf zur Verfügung. Die feineren COSMO Modelle wurden sogar dreistündlich initialisiert, jene Daten finden hier jedoch keine Verwendung. Die maximalen Laufzeiten variieren zwischen 18 und 71 Stunden, bewegen sich also im Kürzest- und Kurzfristbereich. Bei der Gegenüberstellung aller Modelle werden die längsten durchwegs vorhandenen, also die 18-stündigen Vorhersagen des 00 UTC Laufs herangezogen, die um 18 UTC gültig sind. Für viele andere Vergleiche, die z. B. längere Vorhersagedauern, andere Initialisationszeiten, etc. evaluieren, können nicht alle Modelle einbezogen werden. Auf die räumlichen Verfügbarkeit der Modelldaten wird in Abschn. 2.8 eingegangen.

Bei allen verwendeten Prognosesystemen handelt es sich um sogenannte „Limited-Area Models“ (LAMs), also Modelle, die nicht global, sondern über eine abgegrenzte Region betrieben werden. Die Bedingungen am Rand des Gebiets werden von einem übergeordneten, „antreibenden“ Modell vorgegeben. Bei den sehr hochauflösenden Modellen sind diese wieder LAMs, welche selbst wiederum in Globalmodelle einge-

bettet sind, man spricht auch von „*nesting*“ und verschiedenen „*nesting*“-Stufen. In der getroffenen Auswahl finden sich Teile aus vier Modellketten, bis auf das kanadische Modell sind immer jeweils zwei der LAMs ineinander genestet.

Alle verwendeten Prognosedaten stammen aus dem D-PHASE Datenarchiv des WDCC. Die im Folgenden zitierten Einträge in der Datenbank, verfasst von den jeweiligen zur Verfügung stellenden Institutionen, geben Aufschluss über die zum betrachteten Zeitpunkt verwendeten Modellspezifikationen. Weitere Details zu den eingesetzten physikalischen Parametrisierungen, numerischen Zeitschemata und Zeitschritten, Datenassimilationsmethoden und Gitterspezifikationen können für jedes der in den nächsten Unterabschnitten besprochenen Modelle in den angegebenen Quellen nachgelesen werden. Die zur Interpretation wichtigsten Unterschiede und Gemeinsamkeiten sind in Tabelle 2.1 zusammengefasst. Ein herausstechendes Merkmal bei allen Modellfamilien ist, dass die höher auflösenden Modelle ($< 3\,\mathrm{km}$ Gitterpunktsabstand) im Gegensatz zu den gröberen keine Parametrisierung für hochreichende Konvektion verwenden. Es wird erwartet, dass große konvektive Zellen durch ihre typische räumliche Skala von wenigen Kilometern vom Modell selbst erfasst werden können. Flache Konvektion muss für alle angeführten Modelle parametrisiert werden.

ALADIN-FR und AROME

Der französische nationale Wetterdienst betrieb während D-PHASE seine Version des Aire Limitée Adaption dynamique Développement INternational (ALADIN) Modells, ein hydrostatisch approximiertes LAM, das in Zusammenarbeit mit weiteren europäischen nationalen Wetterdiensten von Météo-France entwickelt wurde (Bubnova et al. 1993). Zum behandelten Zeitpunkt weist es eine horizontale Auflösung von 9.5 km und 46 vertikale Schichten auf (Bazile 2009). Übergeordnet ist dem ALADIN France, oder ALADIN-FR, das Globalmodell ARPEGE, mit dem es auch den dynamischen Kern teilt. Von den in dieser Arbeit behandelten Modellen ist ALADIN-FR das einzige hydrostatisch approximierte. In das ALADIN-FR ist wiederum das mit 2.5 km Gitterpunktsabstand noch feinere AROME (Application of Research to Operations at Mesoscale, Seity 2009) mit 41 vertikalen Niveaus genestet. AROME basiert auf dem dynamischen Kern einer nicht-hydrostatischen Version von ALADIN (ALADIN-NH) und den physikalischen Parametrisierungen des Forschungsmodells Meso-NH. 2007 befand sich AROME noch in Testphase und wurde auch während des D-PHASE Zeitraums noch weiterentwickelt. Operationelle Anwendung findet es seit Ende 2008. Weitere allgemeine sowie aktuelle Informationen

zu den von Météo-France betriebenen Modellen finden sich bei Fischer et al. (2005), Yessad (2011) und Bouttier (2010).

COSMO-EU und COSMO-DE

Das internationale COnsortium for Small-scale MOdelling (COSMO), das mehrere nationale Wetterdienste vereint, entwickelt und erhält ein nicht-hydrostatisches LAM für operationelle Anwendung und Forschungszwecke. Die Grundversion des COSMO Modells, welche damals den Namen Lokal-Modell (LM) trug, wurde vom Deutschen Wetterdienst (DWD) hervorgebracht. Aus dieser Basis entstanden das LME (LM Europa) und das LMK (LM Kürzestfrist), welche die seitlichen Randdaten zum Antrieb aus dem globalen Modell GME bezogen. Die Weiterentwicklung der beiden Modelle fand innerhalb von COSMO statt und im Jahr 2007 wurde im Sinne einer einheitlichen Namensgebung innerhalb des Konsortiums das LME in COSMO-EU umbenannt und das LMK in COSMO-DE. Das COSMO-EU hat eine horizontale Auflösung von 7 km und 40 horizontale Modellflächen (Schulz und Denhard 2009), was sich beim eingebetteten COSMO-DE auf 2.8 km Gitterpunktsabstand und 50 Schichten (Baldauf et al. 2009) erweitert. Weitere Unterschiede zwischen den beiden liegen im Zeitverfahren, der Datenassimilationsmethode und der physikalischen Parametrisierungen nicht nur für die hochreichende Konvektion, sondern auch für die Wolkenmikrophysik. Details zur Formulierung sind für das COSMO-EU bei Schulz und Schättler (2009) und für das COSMO-DE bei Baldauf et al. (2011) zu finden.

COSMO-7 und COSMO-2

Auch MeteoSchweiz betreibt LAMs aus der COSMO Familie: das COSMO-7 und das COSMO-2, welche den beiden COSMO Modellen des DWD sehr ähnlich sind. Der wesentliche Unterschied liegt im verwendeten Globalmodell, welches in diesem Fall nicht das GME, sondern das vom European Centre for Medium-Range Weather Forecast (ECMWF) entwickelte Integrated Forecast System (IFS) ist. Auch bezüglich der Gitterstrukturen und Modellgebiete liegen wieder Unterschiede vor: So werden COSMO-7 und COSMO-2 beide auf 60 vertikalen Schichten gerechnet, der horizontale Gitterpunktsabstand liegt im COSMO-7 bei ungefähr 7 km (Ament und Arpagaus 2009b), im COSMO-2 bei 2.2 km (Ament und Arpagaus 2009a). COSMO-2 war zum betrachteten Zeitraum noch nicht im operationellen Einsatz. Einzelheiten zum Aufbau der COSMO Modelle im Allgemeinen können in den verschiedenen Teilen der Modellbeschreibung von Doms (2011), Doms et al. (2011) sowie Schraff und Hess

(2011) nachgelesen werden.

CMC-GEM-L

Der kanadische Wetterdienst Canadian Meteorological Centre (CMC), Teil von Environment Canada (EC), betrieb währen D-PHASE zwei Global Environment Multiscale (GEM) Modelle, welche die Namenszusätze „L“ für „*low resolution*“ und „H“ für „*high resolution*“ tragen. In den folgenden Modellvergleich kann jedoch nur das gröbere CMC-GEM-L (McTaggart-Cowan 2009) einbezogen werden, da das feinere CMC-GEM-H täglich nur um 06 UTC gestartet wurde und nicht wie die anderen Modelle für den 00 UTC Lauf verfügbar ist. Das CMC-GEM-L hat mit ungefähr 15 km die gröbste horizontale Auflösung der hier behandelten Modelle und wurde auf 58 vertikalen Modellflächen gerechnet. Die Randbedingungen werden vom eigenen globalen Modell CMC-GEM meso-global vorgegeben. Details zum Modellaufbau finden sich bei Mailhot et al. (2006), Côté et al. (1998) und Milbrandt et al. (2008).

Modell	horiz. Auflösung	Prognose-zeitraum	Institution	antreibendes Modell	hochr. Konvektion
ALADIN-FR	9.5 km	30 h	Météo France	ARPEGE	paramterisiert
AROME	2.5 km	30 h	Météo France	ALADIN-FR	nicht param.
CMC-GEM-L	15.0 km	24 h	EC, CMC	CMC-GEM meso-global	parametrisiert
COSMO-2	2.2 km	24 h	MeteoSchweiz	COSMO-7	nicht param.
COSMO-7	7.0 km	72 h	MeteoSchweiz	ECMWF	parametrisiert
COSMO-DE	2.8 km	18 h	DWD	COSMO-EU	nicht param.
COSMO-EU	7.0 km	72 h	DWD	GME	parametrisiert

Tabelle 2.1: Die getroffene Auswahl aus den während D-PHASE betriebenen hochauflösenden Modellen und einige ihrer Eigenschaften. Angaben der horizontalen Auflösung in km sind teilweise nur mittlere Werte, da die Modelle oft lat/lon-Gitter verwenden.

2.4 MESOCLIM, ein mesoskaliger Klimadatensatz über 35 Jahre

Als zusätzliche klimatologische Referenz sollen in dieser Arbeit die im Rahmen des Projekts „MESOscale Alpine CLIMatology“ (MESOCLIM, P18296) am IMGW für den Zeitraum von 1971 bis 2005 gesammelten hoch aufgelösten synoptischen Daten dienen. Der Datensatz enthält Wettermeldungen von über 4000 Land- und Seestationen in Europa und Umgebung (34° N bis 63° N und 11° W bis 42° O) und wurde aus

mehreren Quellen zusammengeführt. So flossen sowohl die ERA-40 (ECMWF 40 Year Re-Analysis) Reanalysedaten als auch Stationsdaten des MARS (Meteorological Archival and Retrieval System, ECMWF) für das gesamte Gebiet sowie von DWD, MeteoSchweiz und ZAMG (Zentralanstalt für Meteorologie und Geodynamik) für jeweils Deutschland, die Schweiz und Österreich ein.

Aus den gesammelten Daten wurden mithilfe von VERA Analysefelder für ein Gebiet von mehr als 3000 km × 3000 km und einer Auflösung von 16 km erstellt (Mittelpunkt des Gebiets: 47.26° N, 11.39° O). Die Felder aller VERA Parameter mit Ausnahme des Niederschlags stehen also für dieses Gitter und diese Zeitspanne dreistündlich zur Verfügung. Die Niederschlagsmenge ist über sechs Stunden akkumuliert und liegt daher auch nur im Sechs-Stunden-Intervall vor. Zudem sind für den Niederschlag in den ersten neun Jahren von 1971–1979 vergleichsweise nur sehr wenige Meldungen vorhanden, die sich auf ein kleines Gebiet um Österreich und Deutschland beschränken. Für den Niederschlag werden daher im Folgenden die Analysen jenes Zeitraums nicht verwendet, die Klimaperiode wird auf die verbleibenden 26 Jahre (1980-2005) verkürzt.

2.5 Aufbereitung der Modelldaten

Von den bereits erwähnten zu verifizierenden Parametern werden die horizontalen Windkomponenten und die Niederschlagsmenge sowohl von den Modellen als auch von VERA direkt ausgegeben. Die potentielle und die äquivalentpotentielle Temperatur sind hingegen nicht Teil eines gewöhnlichen Modell-Outputs und müssen daher erst aus Temperatur, Luftdruck und dem jeweils zur Verfügung stehenden Feuchtemaß berechnet werden.

Die **potentielle Temperatur** θ (2.1) ist jene Temperatur, die ein Luftpaket hätte, wenn es trockenisentrop von seinem aktuellen Drucklevel p auf das Referenzniveau $p_0 = 1000\,\mathrm{hPa}$ komprimiert oder expandiert würde.

$$\theta = T\left(\frac{p_0}{p}\right)^{\kappa} \quad \text{mit} \quad \kappa = \frac{\mathrm{R}_L}{\mathrm{c}_p} = \frac{287\,\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{K}^{-1}}{1005\,\mathrm{J}\cdot\mathrm{kg}^{-1}\cdot\mathrm{K}^{-1}} \approx \frac{2}{7} \tag{2.1}$$

$\mathrm{R_L}$ bezeichnet die Gaskonstante für trockene Luft und c_p die spezifische Wärmekapazität bei konstantem Druck.

Bei der **äquivalentpotentiellen Temperatur** θ_e (2.2) werden zusätzlich die Phasenumwandlungen Verdunstung und Kondensation zugelassen.

$$\theta_e = \theta \exp\left(\frac{\mathrm{L}q}{\mathrm{c}_p T}\right) \quad \text{mit} \quad q = 0.622\frac{e}{p - 0.377e} \tag{2.2}$$

Dabei ist e der Dampfdruck (Partialdruck des Wasserdampfes), L die Phasenumwandlungswärme bei Kondensation oder Verdunstung von Wasser, q die spezifische Feuchte und T wieder die aktuelle Lufttemperatur.

Für den Luftdruck wird zwar auch von den Modellen nicht nur der Wert am Modellgitterpunkt, sondern auch der auf Meeresniveau reduzierte übermittelt, jedoch erfolgt die Reduktion je nach Wetterdienst auf unterschiedliche Arten. Wie in Gorgas (2006) gezeigt, ist es jedoch sinnvoll, bei einem Modellvergleich eine einheitliche **Druckreduktion** durchzuführen, da ansonsten tendenziell größere Fehler auftreten. Daher werden alle Druckwerte unter Verwendung einer Standardmethode nach Reuter et al. (2001), welche auf der barometrischen Höhenformel und der Annahme einer polytropen Atmosphäre basiert und eine Schwerekorrektur enthält (vgl. auch Gorgas 2006), auf Meeresniveau reduziert.

Die **Niederschlagsmenge** wird zur Vereinheitlichung mit den Klimadaten aus den einstündigen Werten über sechs Stunden aufsummiert.

Außerdem müssen die **Windkomponenten** bei denjenigen Modellen, die ein rotiertes sphärisches Gitter mit verschobenem Nordpol verwenden (alle COSMO Modelle), auf das geographische Koordinatensystem transformiert werden (Kiesenhofer 2011).

Die Gitter aller Modelle sowie das von VERA sind exakt durch die geographischen Koordinaten und die Höhe über dem Meeresspiegel für jeden Gitterpunkt gegeben. Um die Modelldaten auf das 8 km VERA-Gitter zu bringen, wird sowohl eine horizontale **Cressman-Interpolation** wie bei Gorgas (2006) angewendet als auch eine **Höhenkorrektur** der potentiellen und äquivalentpotentiellen Temperatur, die den Unterschieden zwischen Modelltopographie und VERA Minimumtopographie Rechnung trägt.

Für die Höhenkorrektur werden aus den zweimal täglich (um 00 und 12 UTC) verfügbaren Radiosondenmessungen [2] von sieben Stationen im betrachteten Gebiet (Payerne, Idar-Oberstein, Stuttgart, Kümmersbruck, München, Udine, Mailand) die mittleren Höhenverläufe von θ und θ_e für jedes Monat des Herbsts und Sommers 2007 berechnet. Durch lineare Regression zwischen 500 und 3400 m werden die mittleren Gradienten γ_θ und γ_{θ_e} ermittelt (Abb. 2.2), welche in Folge zur Korrektur gemäß

[2] Die Radiosondendaten wurden aus dem Online-Archiv der University of Wyoming bezogen: `http://weather.uwyo.edu/upperair/sounding.html` (abgerufen am 14.03.2012).

(2.3) und (2.4) angewendet werden.

$$\theta_{\text{korr}} = \theta - (H_{\text{Modell}} - H_{\text{VERA}}) \cdot \gamma_\theta \tag{2.3}$$

$$\theta_{e,\text{korr}} = \theta_e - (H_{\text{Modell}} - H_{\text{VERA}}) \cdot \gamma_{\theta_e} \tag{2.4}$$

Die mittlere Schichtung im Sommer 2007 ist, wie in Abb. 2.2 erkennbar, gleichzeitig trockenstabil (θ steigt mit der Höhe) und feuchtlabil (θ_e sinkt mit der Höhe), auch „bedingt labil“ genannt. Auch im September ist der θ_e-Gradient noch schwach negativ. Die Stabilität (feucht und trocken) nimmt mit jedem Monat zu und im Oktober und November herrschen im Mittel absolut stabile Verhältnisse. Die verwendeten Stationen der Radiosondenaufstiege liegen nördlich und südlich der Alpen über das „limitierte D-PHASE Gebiet“ (vgl. Abschn. 2.8) verteilt und der für die Regression verwendete Höhenbereich entspricht der Spannweite der Täler und Gipfel der Modelltopographien. Daher können die berechneten Gradienten als ausreichend repräsentativ für die Punkte, in denen ein signifikanter Unterschied zwischen Modelltopographien und VERA Minimumtopographie herrscht, angesehen werden. Eine noch genauere aber auch aufwändigere Methode würde für jeden Orts- und Zeitpunkt die nächstgelegene verfügbare Radionsondenmessung zur Korrektur verwenden.

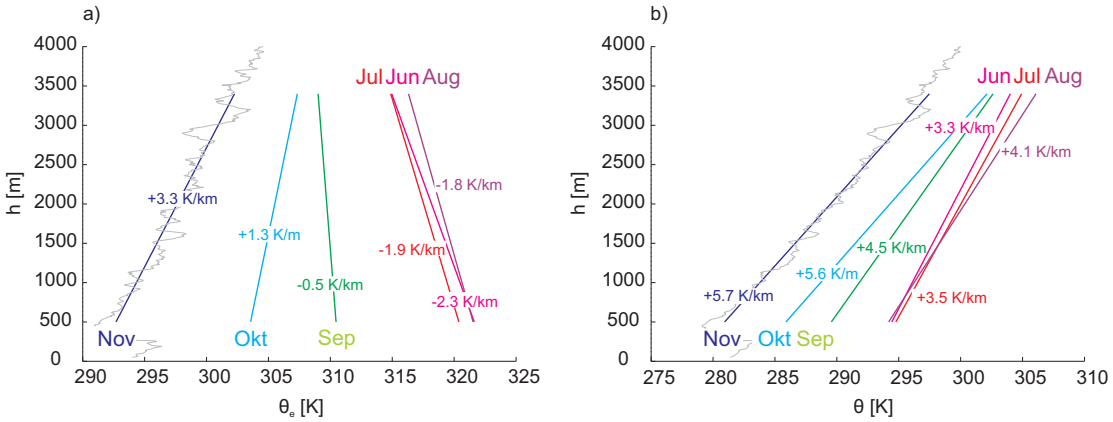

Abbildung 2.2: Linear approximierte, monatlich gemittelte Höhenverläufe der a) äquivalentpotentiellen und b) potentiellen Temperatur aus den verfügbaren 00 und 12 UTC Radiosondenmessungen des jeweiligen Monats (2007) der Stationen Payerne (06610), Idar-Oberstein (10618), Stuttgart (10739), Kümmersbruck (10771), München (10868), Udine (16044) und Mailand (16080). Die angegebenen Gradienten $\gamma_{\theta(e)}$ werden zur Höhenkorrektur verwendet. Zusätzlich eingezeichnet (grau) ist der mittlere Höhenverlauf im November.

Für den Wind wäre eine Höhenkorrektur nach den Radiosondenmessungen nicht sinnvoll, da Messungen in der Höhe die Zustände in der freien Atmosphäre wiedergeben, während sich der Modellvergleich nur auf Bodennähe (10 m) bezieht. Außerdem spielen lokale Einflüsse eine viel größere Rolle als bei der (äquivalent-) potentiellen

Temperatur. Es wird daher keine Korrektur durchgeführt, es kann einzig festgelegt werden, dass Werte ab einer gewissen, willkürlich festlegbaren Höhendifferenz ($|H_{\text{Modell}} - H_{\text{VERA}}|$, z. B. 300 m) ausgeschlossen werden.

2.6 Aufbereitung der MESOCLIM Daten

Für den in dieser Arbeit benötigten Zweck, eine klimatologische Referenz zu bilden, werden aus den MESOCLIM Daten langjährige Mittelwerte für jeden VERA-Gitterpunkt und jeden (stündlichen oder dreistündlichen) Zeitpunkt des Halbjahres benötigt. Aus den vorhandenen Zeitreihen der VERA Parameter am 16 km-Gitter werden dazu gleitende Mittel, die jeweils auch die Werte von ±14 Tagen miteinbeziehen, gebildet. Wichtig dabei ist, dass immer nur dieselben Tageszeiten zusammengefasst werden. Ein Mittelwert ergibt sich also aus 29 Tagen × 35 Jahren = 1015 Werten der Zeitreihe und steht zwar für einen genauen Termin des Jahres, enthält jedoch ähnlich einem Monatsmittel die Information aus 29 Tagen. Für den Niederschlag wurden wie erwähnt nur 26 Jahre verwendet.

Für einen ausgewählten Gitterpunkt ungefähr 16 km östlich von Innsbruck ist in Abb. 2.3 a bzw. b der erhaltene mittlere Jahres- und Tagesgang des reduzierten Luftdrucks abgebildet. Als grobe Kontrolle der Daten kann festgestellt werden, dass der Jahresverlauf sehr gut mit dem der Monatsmittel aus den NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) Reanalysen [3] für einen nahe gelegenen Gitterpunkt (c) übereinstimmt. Ebenso weist der Tagesverlauf die bei Knabl (2004) beschriebenen typischen Charakteristika des Bodendruckverlaufs im Alpenraum (d) auf. Auch die Mittelungen der anderen meteorologischen Parameter wurden ähnlich überblicksweise auf ihre Sinnhaftigkeit überprüft.

Die so erhaltenen gemittelten Gitterpunktszeitserien werden im nächsten Schritt wieder zu Feldern zusammengesetzt und mittels Cressman Interpolation auf das 8 km-VERA Gitter gebracht. Davor wird außerdem für Zusatzbetrachtungen zeitlich mit kubischen Splines auf einen einstündigen Zeitabstand interpoliert.

[3] Die NCEP/NCAR Reanalysedaten (vgl. Kalnay et al. 1996) werden von NOAA/OAR/ESRL PSD, Boulder, Colorado, USA auf ihrer Website `http://www.esrl.noaa.gov/psd/` (abgerufen am 08.03.2012) frei zur Verfügung gestellt.

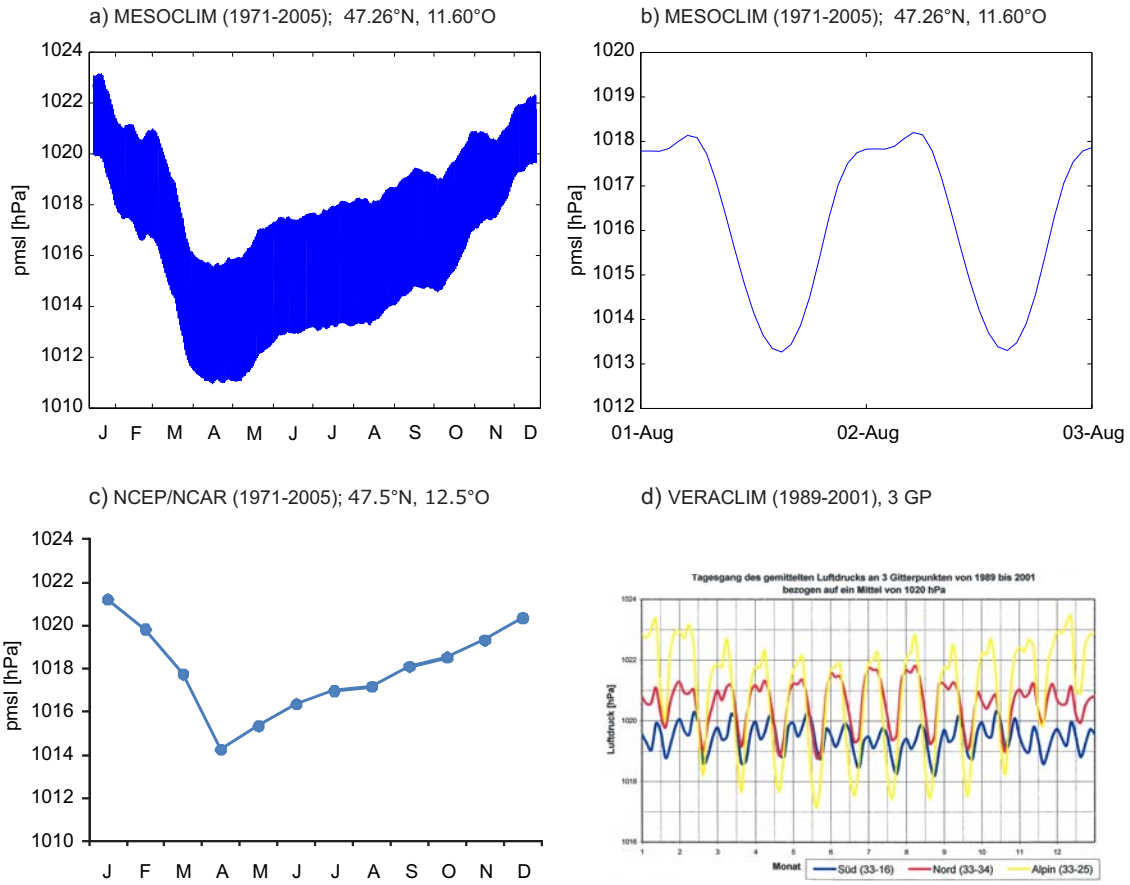

Abbildung 2.3: Klimamittelwerte des reduzierten Drucks als Reanalyse-Gitterpunktszeitreihen. a) mittlerer Jahresgang für 47.26° N, 11.60° O (nahe Innsbruck) aus MESOCLIM; b) Tagesschwankungen für dasselbe Beispiel; c) Monatsmittel aus den NCEP/NCAR Reanalysen für den nächstliegenden Gitterpunkt (47.5° N, 12.5° O); d) monatliche mittlere Tagesgänge aus VERACLIM für drei Gitterpunkte (südl. der Alpen/nördl./inneralpin, Abszisse: je 00–24 UTC für jedes Monat), Quelle (d): Knabl (2004).

2.7 Berechnung der Windgrößen

Wie in Abschn. 3.5 genauer ausgeführt, werden zur Auswertung für den Wind in dieser Arbeit nicht die horizontalen Komponenten u und v direkt herangezogen, sondern Windgeschwindigkeit *ff* und Windrichtung *dd*. Diese müssen für Modell-, VERA- und MESOCLIM-Felder gleichermaßen berechnet werden.

Es werden die gängigen meteorologischen Richtungskonventionen verwendet:

- $u > 0, v = 0$ entspricht Westwind ($dd = 270°$),
- $u = 0, v > 0$ entspricht Südwind ($dd = 180°$).

Die Berechnung von Windgeschwindigkeit und Windrichtung nach diesen Konven-

tionen ergibt sich wie folgt:

$$ff = |\vec{v}| = \sqrt{u^2 + v^2} \tag{2.5}$$

$$dd = (\arccos(\frac{v}{ff}) + \pi) \cdot \frac{180}{\pi} \quad \text{für } u \geq 0, ff \neq 0 \tag{2.6}$$

$$dd = (-\arccos(\frac{v}{ff}) + \pi) \cdot \frac{180}{\pi} \quad \text{für } u < 0, ff \neq 0 \tag{2.7}$$

Da eine skalare Mittelung von *dd* nicht sinnvoll möglich ist, wurde zur Berechnung der Klimawerte von *dd* komponentenweise gemittelt, für *ff* jedoch direkt skalar (vgl. Abschn. 3.5).

2.8 Zeitliche und räumliche Aufgliederung

Das räumliche Gebiet, das in MAP-D-PHASE behandelt wurde, reicht von 43° N bis 50° N und von 2° O bis 18° O und umspannt die Alpen großräumig. Die Ausschnitte, über welche die Modelle der verschiedenen Wetterdienste während des Sommers und Herbsts 2007 tatsächlich gerechnet wurden, weichen jedoch unglücklicherweise in unterschiedlichen Regionen davon ab. Die Überlappungsfläche aller betrachteten Modelle, also die größte zur Gegenüberstellung verwendbare Domäne, ist in Abb. 2.4 zu sehen und soll im Weiteren **„limitiertes D-PHASE Gebiet“** genannt werden. Es ist ungefähr 600 km × 500 km groß, enthält 5435 VERA-Gitterpunkte und erfasst immer noch den größten Teil des Alpenbogens sowie der nördlich und südlich benachbarten Gebiete.

Auf eine sehr viel kleinere Fläche konzentriert sich das **Gebiet von COPS**, welches für diese Arbeit wie in Abb. 2.4 erkennbar, an der Topographie orientiert nachgezeichnet wurde. Es enthält im Westen den Gebirgszug der Vogesen, auf den nach Osten hin das Rheintal und die weitere Erhebung des Schwarzwaldes folgen. Zum Zweck der genaueren Untersuchung der orographischen Besonderheiten wurde eine zusätzliche Unterteilung des COPS-Gebiets getroffen. Die Subdomänen „Vogesen“ und „Schwarzwald“ grenzen im Rheintal aneinander, während das „Rhein“-Gebiet ungefähr von Kamm zu Kamm reicht. Die drei teilweise überlappenden Unterregionen sind mit je ca. 300 Gitterpunkten ungefähr gleich groß und daher in Auswertungen vergleichbar.

Zur ersten zeitlichen Stratifizierung des Datensatzes bietet es sich an, die einzelnen

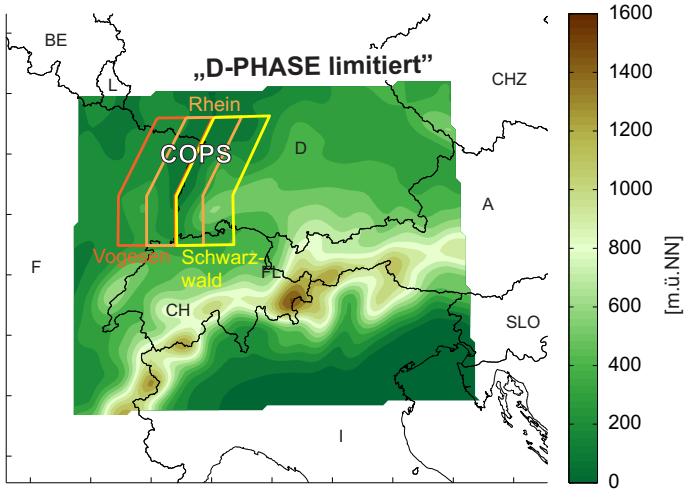

Abbildung 2.4: VERA Minimumtopographie und die Lage des „limitierten D-PHASE Gebiets" (Überlappungsbereich der Modelldomänen) sowie der COPS-Domäne und deren Unterteilung in Vogesen, Rhein und Schwarzwald.

Monate zusammenzufassen und so Unterschiede im **Halbjahresverlauf** herauszuarbeiten. Einen gröberen Überblick sollen Vergleiche zwischen den **Jahreszeiten** Sommer (Juni–August) und Herbst (September–November) und Bewertungen des gesamten sechsmonatigen Zeitraums geben. Bei allen diesen Unterteilungen wird immer nur eine Tageszeit und eine Vorhersagedauer (wie erwähnt bevorzugt 18 UTC und +18 h) betrachtet, um die Variationen im Tagesverlauf auszuklammern.

Die **tageszeitlichen Unterschiede** in den Modellprognosen werden beim Vergleich der (drei-)stündlichen Vorhersagwerte des 00 UTC Laufs untersucht. Eine Einschränkung ergibt sich dadurch, dass die ersten sechs Stunden einer Prognose im Allgemeinen nicht als vertrauenswürdig eingeschätzt werden, da das Modell ungefähr diese Zeit zum „Einschwingen" benötigt, bis in sich konsistente Felder erzeugt werden („*spin-up*" Effekt). Sinnvolle Aussagen für die Morgenstunden zwischen 00 und 06 UTC können daher bei Ausschluss anderer Initalisationszeiten nur getroffen werden, wenn Vorhersagezeiten bis 30 Stunden herangezogen werden (vorhanden für ALADIN-FR, AROME, COSMO-7 und COSMO-EU). Für den Abend nach 18 UTC stehen aus dem 00 UTC Lauf keine Vorhersagen des COSMO-2 zur Verfügung.

Ein gängiger Gegenstand der Verifikation ist, das Verhalten eines Modells mit **steigender Prognosedauer** zu untersuchen. Diese Betrachtungen sind allerdings hauptsächlich bei mittelfristigen Prognosen über mehrere Tage sinnvoll und stehen daher nicht im Hauptaugenmerk dieser Arbeit. Von den vorliegenden Modellen kommen für Auswertungen dieser Art nur COSMO-7 und COSMO-EU in Frage, welche mit 72 Stunden die größten Vorhersagedauern haben.

Eine weitere Möglichkeit um das Modellverhalten zu prüfen, ist das Herausfiltern

signifikanter Wetterlagen wie **Frontdurchgänge**. Die Qualität der Modelle unter diesen „schwierigen“ Verhältnissen von relativ starker Varianz und schnellen Änderungen in den meisten meteorologischen Parametern ist von besonderem Interesse. Im betrachteten Zeitraum wurden 14 Frontdurchgänge im „limitierten D-PHASE Gebiet“ identifiziert (Tabelle 2.2). Zur Bestimmung wurden nicht nur die VERA Bodenanalysen, sondern auch die GFS (Global Forecast System) Prognosen von θ_e in 850 hPa herangezogen. [4] Die Zeiträume der Ereignisse wurden näherungsweise vom Beginn des Einflusses der Front auf das Gebiet bis zu dessen Ende definiert und haben eine Dauer von 24–36 h. Bei dreistündlicher Auflösung ergibt das je 9–13 und insgesamt 139 Termine. Ein etwaiger Einfluss des Tagesgangs auf die Prognosegüte spielt durch das überlagerte Signal des Frontdurchgangs eine untergeordnete Rolle und sollte aufgrund des ähnlichen gewählten Zeitraums von einem bis eineinhalb Tagen für die Ereignisse annähernd gleich sein. Die Fälle von 20. und 22. Juni sowie die von 2. und 3. Juli sind zwar jeweils demselben Frontensystem zuzuordnen, um die Dauer der einzelnen Ereignisse vergleichbar zu halten, wurden diese Termine jedoch geteilt.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Beg	15.06. 09:00	20.06. 15:00	22.06. 00:00	02.07. 00:00	03.07. 12:00	07.08. 18:00	16.08. 00:00	03.09. 09:00	18.09. 00:00	24.09. 21:00	05.10. 12:00	17.10. 18:00	09.11. 00:00	24.11. 03:00
Dur	24 h	30 h	27 h	24 h	27 h	36 h	30 h	24 h	24 h	24 h	24 h	24 h	24 h	33 h

Tabelle 2.2: Beginn und Dauer der Kaltfrontdurchgänge während des Sommers und Herbsts 2007 im „limitierten D-PHASE Gebiet“.

Da die Frontdurchgänge alle Tageszeiten umspannen, können aufgrund der vorher besprochenen Einschränkungen für die Berechnung eines Verifikationsmaßes über ein oder mehrere Ereignisse zum einen bei Verwendung ausschließlich des 00 UTC Laufs nur die Modelle mit einer Prognosedauer von mindestens 30 Stunden (ALADIN-FR, AROME, COSMO-7 und COSMO-EU) verglichen werden. Oder aber es werden zwei Initialisationszeiten (00 und 12 UTC) und davon jeweils die Vorhersagen über 9–18 Stunden einbezogen, wie es für alle COSMO Modelle möglich ist.

[4] `http://www.wetter3.de/Archiv` (abgerufen am 08.03.2012).

3 Methoden der Verifikation

Ein zu verifizierender Datensatz kann als Menge von Wertepaaren aus Vorhersagen y und Beobachtungen o betrachtet und durch eine zweidimensionale Häufigkeitsverteilung $p(y, o)$ beschrieben werden. Die Qualität von Vorhersagen ist gemäß des verteilungsorientierten Ansatzes von Murphy und Winkler (1987) gegeben durch die gesamte zeitunabhängige Information, die in dieser bivariaten Verteilung von Prognose- und Beobachtungswerten enthalten ist. Murphy und Winkler (1987) formulieren ein Gerüst der Vorhersageverifikation, das auf der Faktorisierung dieser bivariaten Häufigkeitsverteilung in die marginalen Verteilungen von Beobachtung und Vorhersage $p(o)$ und $p(y)$ sowie auf der bedingten Verteilung der Beobachtungen bei gegebenen Vorhersagen $p(o|y)$ und der bedingten Verteilung der Vorhersagen bei gegebenen Beobachtungen $p(y|o)$ basiert. In dieser Arbeit besteht der zu evaluierende aufbereitete Datensatz aus numerischen Modellprognosen und Analysen auf einem Gitter. Die in diesem Kapitel beschriebenen Methoden sind jedoch auch in allgemeineren Fällen anwendbar, daher werden die Begriffe „Analyse“ und „Beobachtung“ hier äquivalent verwendet.

Unter der Dimensionalität D eines Verifikationsproblems versteht man die Anzahl der Verifikationsgrößen (oder bedingten Wahrscheinlichkeiten), die man braucht, um $p(y, o)$ vollständig zu rekonstruieren. D ist definiert durch die Summe der Freiheitsgrade, also durch die Zahl der unterschiedlichen Kombinationen von Vorhersage- und Beobachtungswerten weniger eins (Murphy 1991). Typische meteorologische Verifikationssituationen sind sehr hochdimensional, wie schon überblicksmäßig an der Häufigkeitsverteilung Abb. 3.2 a zu erkennen ist, wo jeder Balken eine Kombination (y_i, o_j) markiert und, obwohl die Genauigkeit in der graphischen Darstellung bereits auf ungefähr 0.5 °C reduziert wurde, die Zahl der unterschiedlichen Wertepaare D in der Größenordnung von 10^3 liegt. Im Vergleich zur Größe des Datensatzes ($N = 157\,615$) sind die absoluten Häufigkeiten der einzelnen (y_i, o_j)-Paare sehr klein (< 900), die einzelnen Freiheitsgrade also nicht ausreichend bestimmt.

Es ist offensichtlich, dass in solchen typischen Fällen die vollständige Beschreibung der Vorhersagequalität im Sinne einer vollständigen Beschreibung der bivariaten Häufigkeitsverteilung in der Praxis für die Verifikation statistisch schwer möglich

(Unterbesetzung) und durch die hohe Anzahl an benötigten Maßen wenig nützlich ist. Gewisse Einschränkungen des Anspruchs auf die umfassende Bewertung der Vorhersagegüte sind also notwendig.

In vorliegender Arbeit soll die Prognosequalität vordergründig durch möglichst simple, zusammenfassende Methodiken beschrieben werden, um einen einfach zugänglichen Vergleich verschiedener Modelle zu schaffen. Dies muss jedoch im Bewusstsein der dadurch begangenen Vereinfachungen der beschriebenen Komplexität und Dimensionalität des Problems geschehen.

Man unterscheidet grundsätzlich zwischen kategorialen Prädiktanden, die nur eine endliche Anzahl von möglichen Realisierungen haben, und kontinuierlichen Prädiktanden, für die es innerhalb eines Intervalls unendlich viele mögliche Realisierungen gibt. In der Praxis liegen jedoch auch kontinuierliche Variablen diskret vor, da sie von Messgeräten und Vorhersagesystemen immer nur bis zu einem gewissen Maß an Genauigkeit, also einer fixen Anzahl an Dezimalstellen, gemessen bzw. vorhergesagt werden können. In dieser Arbeit werden durchwegs theoretisch kontinuierliche Variablen wie Temperatur, Wind oder Niederschlag behandelt. Eine mögliche Herangehensweise, das Verifikationsproblem zu vereinfachen, ist die Verringerung seiner Dimensionalität, indem eine weitere Kategorisierung der Variablen in eine geringere Anzahl an Klassen vorgenommen wird. Für viele Anwendungen ist jedoch auch die Annahme der Kontinuierlichkeit sinnvoll, vor allem, wenn eher allgemeinere Aussagen über die Übereinstimmung zwischen Beobachtungen und Vorhersagen getroffen werden sollen, anstatt sich auf die Über- oder Unterschreitung eines Schwellwertes zu konzentrieren.

Die in den folgenden Abschnitten beschriebenen Verifikationsmaße sind einzelne Skalare, die für einen Ausschnitt eines Datensatzes von Wertepaaren (y_k, o_k) mit $k = 1, ..., N$ Orts- und Zeitpunkten berechnet werden können. Die Größen werden zum einen für einzelne Zeitpunkte über unterschiedliche räumliche Gebiete evaluiert, zum anderen können auch längere Zeiträume zusammengefasst werden. Im letzteren Fall ist jedoch zu beachten, dass bei nicht-additiven Größen (d.h. alle, bei denen nichtlineare Operationen vorkommen) nicht einfach der arithmetische Mittelwert über die einzelnen Termine gebildet werden kann, sondern das Maß für den gesamten Datenausschnitt berechnet werden muss. Eine Ausnahme, bei der die Orts- und Zeitpunkte nicht ohne Unterscheidung zusammengefasst werden, ist die Anomaliekorrelation (vgl. Abschn. 3.3).

In dieser Arbeit werden nur Verifikationsmaße, die Analyse und Vorhersage an ein und demselben Zeit- und Ortspunkt vergleichen, eingesetzt. Methoden, die einer

möglichen räumlichen Verschiebung Rechnung tragen, wie objektbasierte Verifikation oder „*Fuzzy*“ Ansätze, welche vor allem für den Niederschlag vielversprechende Herangehensweisen darstellen, werden anhand derselben Datengrundlage in der Diplomarbeit von Kiesenhofer (2011) behandelt.

Idealerweise müssten die zur Evaluierung verwendeten Daten statistisch unabhängig sein. Diese Anforderung ist jedoch hier wie in den meisten meteorologischen Untersuchungen nicht vollständig erfüllbar, da die räumlichen und zeitlichen Abstände zwischen den Wertepaaren zu klein sind. Es sollte beachtet werden, dass dadurch die Varianz der Werte unterschätzt werden könnte.

Formal werden im Folgenden die in der Literatur gebräuchlichen englischen Bezeichnungen mehrerer Maßzahlen und themenbezogener Begriffe beibehalten, da in vielen Fällen keine einheitliche deutsche Nomenklatur existiert und so Unklarheiten vermieden werden sollen.

3.1 Grundlegende Verifikationsmaße

3.1.1 Kontinuierliche Prädiktanden

Häufig wird als erster Schritt in der Verifikation von Wetterprognosen der systematischen Fehler oder Bias (3.1) berechnet, der als die Differenz zwischen dem Mittelwert der Prognosen μ_y und dem der Beobachtungen μ_o definiert ist. Dieser gibt die mittlere Über- oder Unterschätzung der Beobachtungen durch das Modell an und ist im Idealfall Null. Keine Information ist jedoch gegeben über den Fehler in unterschiedlichen Bereichen der Analysewerte. So kann z. B. ein geringer Temperatur Bias durch konstant gute Prognosen zustande kommen, oder etwa durch Unterschätzungen der niedrigen Temperaturen und Überschätzungen der hohen Temperaturen, die sich im Mittel ausgleichen. Die volle Information aus der bivariaten Häufigkeitsverteilung würde sich erst durch die bedingten Biase bei allen möglichen Prognose- und Analysewerten erschließen. Dennoch gibt auch schon das einzelne Maß des Bias Einblick in die Vorhersagequalität.

$$\mathrm{B} = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k) = \mu_y - \mu_o \tag{3.1}$$

Weiteren Aufschluss über die mittlere Übereinstimmung der Beobachtung-Vorhersage Wertepaare geben der mittlere absolute Fehler (MAE, Gl. 3.2), der mittlere quadratische Fehler (MSE, Gl. 3.3), sowie die Wurzel aus dem mittleren quadrati-

schen Fehler (RMSE). Der MSE und der RMSE gewichten durch das Quadrieren größere Abweichungen stärker als kleinere, was einerseits wünschenswert ist, andererseits die Maße sensitiv auf Ausreißer macht – eine Eigenschaft, die der MAE nicht teilt. Die besten Werte erreichen MSE und RMSE, wenn die Mittelwerte von Vorhersage und Analyse übereinstimmen, für den MAE ist die Übereinstimmung der Mediane ausschlaggebend (Jolliffe und Stephenson 2003). Der RMSE hat gegenüber dem MSE den Vorteil, dass er dieselbe physikalische Einheit wie die zu evaluierende Variable besitzt.

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N}|y_k - o_k| \tag{3.2}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k)^2 \tag{3.3}$$

Der lineare Korrelationskoeffizient oder Pearson-Korrelationskoeffizient (3.4) zwischen Analyse und Prognose $r_{y,o}$ gibt über die Übereinstimmung unabhängig vom Bias Auskunft. Er ist gegeben durch das Verhältnis der Kovarianz zum Produkt der Standardabweichungen von Analyse und Prognose (σ_o und σ_y) und basiert auf der Annahme einer näherungsweisen Normalverteilung der Variablen. Perfekte lineare Korrelation wird durch $r_{y,o} = 1$ wiedergegeben.

$$r_{o,y} = \frac{\frac{1}{N}\sum_{k=1}^{N}(y_k - \mu_y)(o_k - \mu_o)}{\sigma_y \cdot \sigma_o} \tag{3.4}$$

$$\sigma_o = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(o_k - \mu_o)^2} \tag{3.5}$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \mu_y)^2} \tag{3.6}$$

3.1.2 Aufspaltung des MSE und Taylor Diagramm

Der MSE kann mithilfe einfacher Umformungen zerlegt werden:

$$\mathrm{MSE} - \mathrm{B}^2 = \sigma_y^2 + \sigma_o^2 - 2\sigma_y\sigma_o r_{o,y} \tag{3.7}$$

Die linke Seite von (3.7) wird *Bias-korrigierter MSE* genannt, die Wurzel daraus ist der *Bias-korrigierte RMSE* ($RMSE_{bc}$).

$$\mathrm{RMSE}_{bc}^2 = \sigma_o^2 + \sigma_y^2 - 2\sigma_o\sigma_y r_{o,y} \tag{3.8}$$

In diesem Ausdruck kann eine Äquivalenz zum Kosinussatz (3.9) für ein Dreieck mit den Seitenlängen a, b und c und dem vom a und b eingeschlossenen Winkel ϕ erkannt werden.

$$c^2 = a^2 + b^2 - 2ab\cos\phi \tag{3.9}$$

Die drei Maßzahlen σ, $RMSE_{bc}$ und $r_{o,y}$ einer Prognose lassen sich also als Dreieck darstellen. Taylor (2001) erkannte die Möglichkeit, dadurch mehrere Attribute verschiedener Vorhersagesysteme gleichzeitig in einem Diagramm (Abb. 3.1) sichtbar zu machen. Die Standardabweichungen von Analyse und Prognose sind durch die Entfernung vom Nullpunkt gegeben, wobei der Analysepunkt auf der Abszisse festgemacht ist. Der Winkel zwischen den so aufgespannten Geraden ist durch den Arkuskosinus des Korrelationskoeffizienten gegeben, welcher radial abzulesen ist. Der Abstand zwischen Analyse und Prognose gibt den Bias-korrigierten RMSE. In der Praxis wird oft mit der Standardabweichung der Beobachtung normiert, sodass der Analysepunkt immer bei 1 liegt. Der Korrelationskoeffizient ist invariant gegenüber dieser Modifikation, der normierte $RMSE_{bc}/\sigma_o$ ist hingegen bei Gegenüberstellung verschiedener Vorhersagesituationen (unterschiedlicher Analysen) nicht mehr direkt ohne Berücksichtigung der Analyse-Varianz vergleichbar.

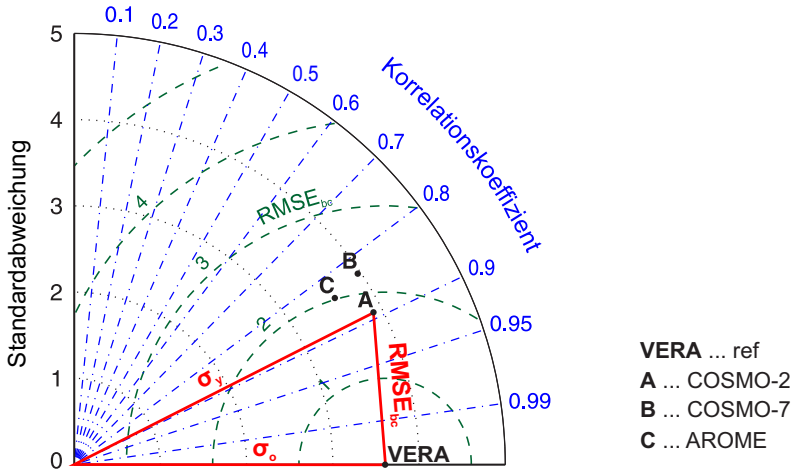

Abbildung 3.1: Taylor Diagramm, Darstellungsweise am Beispiel der 18 h Vorhersagen (00 UTC Lauf) von p_{msl} 04.07.–06.07.2007 im limitierten D-PHASE Gebiet.

3.1.3 Reduktion auf binäre Prognosen

Grundsätzlich kann jede Vorhersage eines kontinuierlichen Prädiktanden auf Vorhersagen beliebig weniger Kategorien reduziert werden, indem Schwellwerte festgelegt

werden. Der Extremfall einer kategorialen Prognose ist eine binäre oder dichotome Prognose mit nur zwei möglichen Realisierungen für y und o wie 1 oder 0 bzw. „ja“ oder „nein“. Abbildung 3.2 zeigt, wie eine „kontinuierliche“ Prognose der potentiellen Temperatur in eine dichotome umgewandelt werden kann. Liegt der Wert der vorherzusagenden Variablen über dem Schwellwert, so wird das Ereignis als zutreffend gewertet, ansonsten als nicht-zutreffend. Jedes Vorhersage-Beobachtungspaar addiert, je nach Übereinstimmung der beiden, einen Zähler zu einem Eintrag der 2x2-Kontingenztabelle (Tabelle 3.1).

		Beobachtungen (o)		
		ja	nein	gesamt
Vorhersagen (y)	ja	*a (hits)*	*b (false alarms)*	$a+b$
	nein	*c (misses)*	*d (correct rejections)*	$c+d$
	gesamt	$a+c$	$b+d$	N

Tabelle 3.1: 2x2-Kontingenztabelle.

Diese Zelleinträge bilden die Basis vieler häufig verwendeter Verifikationsmaße, speziell auch der in Abschn. 3.4 behandelten *skill scores*. Die bivariate Häufigkeitsverteilung ist im dichotomen Fall durch die Kontingenztabelle beschrieben, die Zelleinträge geben die bedingten Häufigkeiten an, z. B. $a/N = p(o_{ja}|y_{ja})$, die Randsummen bestimmen die marginalen Häufigkeiten: $(a+b)/N = p(y_{ja})$. Die Dimensionalität des binären Problems ist $2 \cdot 2 - 1 = 3$, das heißt, die gesamte Kontingenztabelle als Spiegel der 2D-Häufigkeitsverteilung kann durch drei (gut gewählte) Maßzahlen rekonstruiert werden (vgl. Ende des Abschnitts).

Eine solche Stratifizierung eines Datensatzes in nur zwei Kategorien bedeutet eine Vereinfachung und Zusammenfassung des Sachverhalts und ist wie besprochen immer mit einem Verlust von Information verbunden. Um diesem entgegen zu wirken und einer gewissen Breite in der Verteilung der Parameter Rechnung zu tragen, ist es möglich, mehrere Kategorien zu verwenden, oder aber mehrere 2x2-Kontingenztabellen bei unterschiedlichen Schwellwerten zu betrachten. In dieser Arbeit wird nur die zweite Methode gewählt, da für das binäre Problem eine größere Auswahl an oft auch einfacher zu interpretierenden Maßen (im Folgenden auch verkürzend „2x2 Fehlermaße“ oder „2x2 *scores*“, etc. genannt) als für das multikategoriale Problem zur Verfügung steht.

Eine Kategorisierung wird oft durchgeführt für die Parameter Niederschlagsmenge und Wind. Der Hauptgrund dafür ist, dass bei *RR* und *ff* die Überschreitung gewisser Schwellwerte meteorologisch besonders interessant und für viele Nutzer von

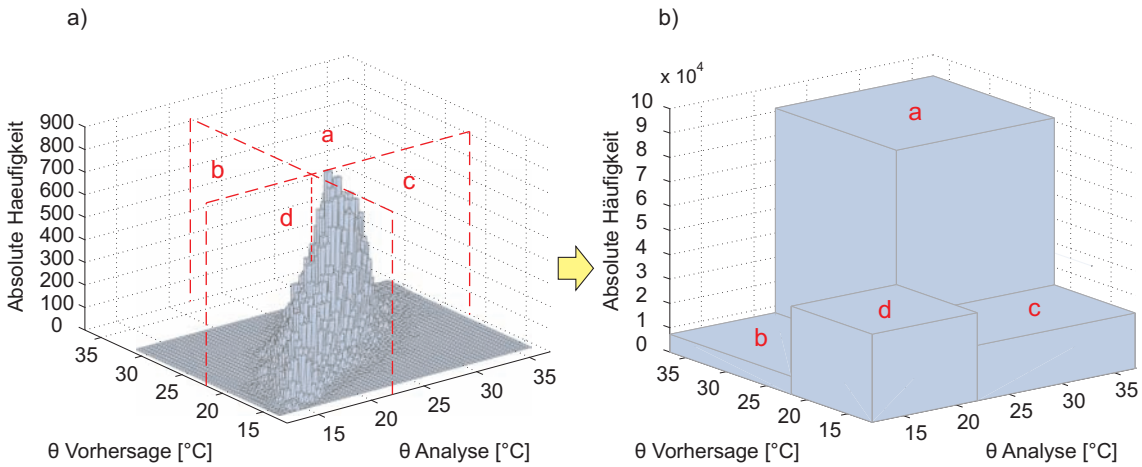

Abbildung 3.2: Bivariate Häufigkeitsverteilung der Analyse- und (18 h COSMO-7) Prognosewerte für θ, Jun. 2007, limitiertes D-PHASE Gebiet. a) zeigt eine annähernd kontinuierliche Prognose (Genauigkeit ca. 0.5 °C), die Reduktion auf eine dichotome (Schwellwert 22 °C) ist durch die roten gestrichelten Linien angedeutet; b) zeigt das Ergebnis der Reduktion, die absoluten Häufigkeiten sind die Einträge der 2x2-Kontingenztabelle, markiert durch a,b,c und d in der Notation von Tabelle 3.1.

Vorhersagen wichtig ist. Bei der Windrichtung bietet sich eine Einteilung aufgrund des beschränkten Wertebereichs an. Ein weiterer Punkt ist, dass die Verteilungen von Niederschlag und Windgeschwindigkeit durch ihre einseitige Begrenzung stark von der Normalverteilung abweichen und dadurch für sie der Einsatz vieler gebräuchlicher Maßzahlen kontinuierlicher Prädiktanden problematisch ist.

Auch für dichotome Vorhersagen ist ein Bias definiert, der sogenannte *Frequency Bias* (3.10). Anders als der Bias für kontinuierliche Prognosen, gibt er das Verhältnis der Anzahl der vorhergesagten Ereignisse zu der der beobachteten an. Er zeigt also auf, ob ein Modell ein Ereignis insgesamt zu oft (*Frequency Bias* > 1, „*overforecasting*“) oder zu selten (*Frequency Bias* < 1, „*underforecasting*“) vorhersagt. Im Fließtext soll zur besseren Lesbarkeit die Bezeichnung „Bias“ auch für den *Frequency Bias* beibehalten werden. Wenn immer es sich um zweiwertige Vorhersagesituationen handelt, ist der *Frequency Bias* gemeint.

$$\textit{Frequency Bias} = \frac{a+b}{a+c} \tag{3.10}$$

Die relative Häufigkeit eines Ereignisses wird auch *Base Rate* BR genannt und ist gegeben durch

$$\mathrm{BR} = p(o_{ja}) = \frac{a+c}{N}\,. \tag{3.11}$$

Nicht selten finden die bedingten Häufigkeiten einer „ja-“ Vorhersage bei Auswertungen Verwendung: Die *Probability of Detection* (POD, Gl. 3.12, auch *Hit Rate* genannt) gibt den Anteil der aufgetretenen Ereignisse, die auch korrekt vorhergesagt wurden, wieder und die *Probability of False Detection* (POFD, Gl. 3.13, oder *False Alarm Rate*) die relative Häufigkeit einer positiven Vorhersage des Ereignisses, wenn es nicht aufgetreten ist.

$$\mathrm{POD} = p(y_{ja}|o_{ja}) = \frac{a}{a+c} \tag{3.12}$$

$$\mathrm{POFD} = p(y_{ja}|o_{nein}) = \frac{b}{b+d} \tag{3.13}$$

Wie oben erwähnt, existieren für die bivariate Häufigkeitsverteilung $p(y, o)$ zwei Faktorisierungen, herleitbar aus dem Multiplikationssatz der Wahrscheinlichkeitsrechnung. Eine davon ist die *likelihood-base rate* (LBR) Faktorisierung:

$$p(y, o) = p(y|o)p(o) \tag{3.14}$$

Im binären Fall entspricht der erste Term (die *likelihoods*) und der POFD, der zweite Term der Base Rate. Aus diesen drei Maßzahlen (POD, POFD und BR) kann also die Kontingenztabelle vollständig beschrieben werden. Daher können auch alle auf der Kontingenztabelle basierenden Maße, insbesondere die in Abschn. 3.4 beschriebenen *skill scores*, als Funktionen der drei ausgedrückt werden. Weitere in der Literatur vorgeschlagene Möglichkeiten solcher Dreiergruppen an Fehlermaßen, die die gesamte Information von $p(y, o)$ im binären Fall enthalten, sind Bias, POD und POFD oder Bias, PSS (3.27) und ORSS (3.34) (Stephenson 2000).

3.2 Skill scores - Allgemeines

Unter dem „*skill*“ einer Vorhersage versteht man ihre relative Qualität im Vergleich zu einer Referenzprognose (Wilks 1995), wie z. B. einer Klima-, Persistenz- oder Zufallsvorhersage. Quantitativ kann der *skill* einer Vorhersage, also ihre Verbesserung gegenüber der Referenz, durch verschiedene *skill scores* ermittelt werden.

Die allgemeine Formulierung eines *skill scores* lautet

$$\mathrm{SS}_{(\mathrm{ref})} = \frac{A - A_{\mathrm{ref}}}{A_{\mathrm{perf}} - A_{\mathrm{ref}}}\,. \tag{3.15}$$

A	... ein wählbares Fehlermaß (*measure of accuracy*) bzw. sein Wert für die zu bewertende Prognose
A_{perf}	... Wert von A für eine perfekte Prognose, meist 1
A_{ref}	... Wert von A für die Referenzprognose

Manchmal wird zusätzlich mit 100 % multipliziert, um den Prozentsatz der Verbesserung gegenüber der Referenz zu erhalten. Für eine perfekte Prognose ist $A = A_{\text{perf}}$ und der *skill score* erreicht seinen Maximalwert von 1. Prognosen, die gemäß des Fehlermaßes A besser sind als die Referenzprognose, ergeben einen positiven *skill score*, schlechtere einen negativen. Keine Verbesserung, also das „Null-*skill*" Niveau, wird durch $\text{SS} = 0$ angezeigt. Wie weit ein *skill score* ins Negative gehen kann, hängt vom gewählten Fehlermaß A ab, ist jedoch in der Praxis von geringer Bedeutung. Nicht alle *skill scores* folgen der allgemeinen Formulierung (3.15).

Referenzprognosen

In dieser Arbeit finden die drei gebräuchlichsten Referenzprognosen Verwendung:

- *Persistenz*: Als Vorhersage wird der beobachtete Wert der Variablen eines vorhergehenden Zeitpunktes (im Allgemeinen des „jüngsten" verfügbaren Termins vor der Initialisierung) verwendet. Eine Persistenzprognose liefert bei Kurzzeitprognosen von bis zu 2 Tagen für viele meteorologischen Parameter oftmals bessere Ergebnisse als eine Klimareferenzprognose, ist also schwerer von den Modellen zu übertreffen (vgl. Abschn. 4.6.2).
- *Klima*: Bei einer Klimareferenzprognose wird der klimatologische Mittelwert der Variablen als Vergleichsvorhersage verwendet. In dieser Arbeit wird jener aus dem 35-jährigen MESOCLIM Datensatz des IMGW für jeden Gitterpunkt und für jede Tages- und Jahreszeit berechnet (vgl. Abschn. 2.6). Der Vorteil der Verwendung des Klimas als Referenz ist, dass es von den Modellvorhersagen und den Analysen unabhängig ist.[1]
- *Zufall*: Das zu übertreffende „Null-*skill*" Niveau kann auch von einer Zufallsprognose definiert werden. Diese ist jedoch nicht völlig beliebig, sondern muss dieselbe marginale Verteilung von Analyse- und Prognosewerten wie die Ausgangsvorhersage haben. Die Verwendung einer Zufallsprognose als Referenz ist

[1] Es sollte angemerkt werden, dass die Modellunabhängigkeit von MESOCLIM und VERA nicht vollständig gegeben ist, da über dem Meer auch „Bogus"-Stationen, also Modelldaten, verwendet werden. Im hier betrachteten räumlichen Gebiet ist ihr Einfluss jedoch vernachlässigbar klein.

vor allem für *skill scores* binärer Prädiktanden verbreitet, für welche sich eine solche aus der Kontingenztabelle konstruieren lässt. Die Wahrscheinlichkeit eines zufälligen *hits* ergibt sich beispielsweise aus der Wahrscheinlichkeit einer „ja"-Prognose multipliziert mit der Wahrscheinlichkeit einer „nein"-Prognose. Die Anzahl der zufälligen *hits*, *false alarms*, etc. sind also gegeben durch

$$a_{\text{rand}} = \frac{(a+b)(a+c)}{N}, \quad b_{\text{rand}} = \frac{(a+b)(b+d)}{N}, \ldots \tag{3.16}$$

Aus diesen Einträgen der „Zufalls-Kontingenztabelle" können binäre Fehlermaße für eine Zufallsvorhersage gebildet werden. Einige der in Abschn. 3.4 definierten *skill scores* lassen sich aus diesen Faktoren direkt ableiten.

Als Referenz wird eine Vorhersage verwendet, die weniger „Können" und Aufwand als eine Modellprognose voraussetzt und von der daher auch erwartet wird, dass sie schlechtere Ergebnisse als das Modell liefert. Überdies sollte die Referenzprognose für „leicht vorherzusagende" meteorologische Situationen schwerer zu übertreffen sein als für „schwierige", sodass akkurate Prognosen von den *skill scores* bei „einfachen" Fällen weniger gut bewertet werden als bei „schwierigen". Die Verwendung des Klimas als Referenz impliziert also, dass man annimmt, das Modell käme mit Verhältnissen, die nahe am Klimamittel liegen besser zurecht, als mit „Extremen". Bei der Persistenz wird ein Wetterumschwung als schwieriger vorherzusagen eingeschätzt als ein Weiterbestehen der meteorologischen Situation, beim Zufall werden Gegebenheiten, die von einer zufälligen Prognose gut erfasst werden, als einfach eingestuft. All dies sind zulässige Annahmen, die Auswahl der passenden Referenzprognose muss situationsspezifisch getroffen werden (vgl. Abschn. 4.6). Wichtig ist, dass beim Vergleich verschiedener Modelle oder Zeitperioden anhand von Skill Scores eine einheitliche Referenz sowie ein einheitliches räumliches Gebiet verwendet wird.

3.3 Skill scores für kontinuierliche Prädiktanden

MSE-, RMSE- und MAE- *skill scores*: Aus den Standard-Verifikationsmaßen MAE (3.2), MSE (3.3) und RMSE, welche in Abschn. 3.1.1 genauer beschrieben sind, können *skill scores* gemäß der allgemeinen Formulierung (3.15) gebildet werden. Für die Verwendung im dimensionslosen *skill score* spielt die physikalische Einheit des zugrunde liegenden Maßes keine Rolle, am häufigsten wird der MSE herangezogen. Der erhaltene MSE-*skill score* (3.17) wird auch „Reduction of Variance" (RV) ge-

nannt und gibt die Verbesserung des mittleren quadratischen Fehlers der Vorhersage gegenüber dem einer Klimareferenzprognose an.

$$\mathrm{RV} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE_{clim}}} \tag{3.17}$$

Die RMSE- und MAE- *skill scores* sind äquivalent zu (3.17) formuliert. Auch die Wurzel der RV wird manchmal als Verifikationsmaß verwendet und als „Multipler Korrelationskoeffizient“ bezeichnet. Der RMSE-*skill score* und der multiple Korrelationskoeffizient haben einen ähnlichen Informationsgehalt wie die RV und spielen daher in dieser Arbeit beim Modellvergleich eine untergeordnete Rolle. Anstatt des MSE der Klimareferenzprognose kann auch der der Persistenzprognose verwendet werden. Da diese Abwandlung jedoch weniger gebräuchlich ist, soll sie im Folgenden durch die Bezeichnung $\mathrm{RV_{pers}}$ gekennzeichnet werden, während RV immer für den am Klima referenzierten *skill score* stehen soll.

Anomaliekorrelation (AC): Für die AC gibt es zwei verschiedene Formen: die zentrierte (3.18) und die unzentrierte (3.19). Als Anomalien werden in diesem Zusammenhang die Abweichungen des Vorhersage- bzw. Analysewerts vom klimatologischen Wert c an einem Zeit- und Ortspunkt k bezeichnet: $y'_k = y_k - c_k$ und $o'_k = o_k - c_k$.

Die zentrierte AC entspricht dem linearen Korrelationskoeffizienten nach Pearson (3.4) für die Anomalien der betrachteten Variablen.

$$\mathrm{AC}_c = r_{y',o'} = \frac{\sum_{k=1}^{N}(y'_k - \bar{y'})(o'_k - \bar{o'})}{[\sum_{k=1}^{N}(y'_k - \bar{y'})^2 \cdot \sum_{k=1}^{N}(o'_k - \bar{o'})^2]^{1/2}} \tag{3.18}$$

Die Terme $\bar{y'}$ und $\bar{o'}$ bezeichnen die Mittel über die Prognose- bzw. Analyse-Anomalienfelder zu einem Zeitpunkt. Wenn also die $\mathrm{AC_c}$ für längere Zeiträume berechnet wird, muss zuerst über die Orts-, dann über die Zeitpunkte summiert werden. Bei der unzentrierten AC-Formulierung werden diese Feldmittel der Anomalien hingegen nicht von den Faktoren abgezogen.

$$\mathrm{AC}_u = \frac{\sum_{k=1}^{N} y'_k o'_k}{[\sum_{k=1}^{N}(y'_k)^2 \cdot \sum_{k=1}^{N}(o'_k)^2]^{1/2}} \tag{3.19}$$

Die $\mathrm{AC_c}$ wird nur dann gleich der $\mathrm{AC_u}$, wenn sowohl die Analyse-Anomalien, als auch die Vorhersage-Anomalien über die betrachtete Region gemittelt gleich Null sind. Dies kann für globale Felder zutreffen, ist jedoch im Allgemeinen und vor allem bei relativ kleinen räumlichen Gebieten nicht gegeben. Im operationellen Gebrauch wird

eher die $\mathrm{AC_c}$ verwendet, für theoretische Studien auch teilweise die $\mathrm{AC_u}$. Auch weil sie durch die Äquivalenz mit dem Pearson-Korrelationskoeffizenten unter Umständen intuitiver zugänglich ist, soll in den folgenden Auswertungen nur mehr die zentrierte Form (3.18) verwendet werden. Mit der Bezeichnung AC ist demgemäß im Folgenden immer die $\mathrm{AC_c}$ gemeint.

Theoretisch möglich, aber kaum in der Literatur zu finden, ist eine AC mit Persistenz als Referenz, bei der die Anomalien durch $y'_k = y^t_k - y^{t-1}_k$ und $o'_k = o^t_k - o^{t-1}_k$ mit dem hochgestellten Index als Zeitindex gegeben sind und welche im Weiteren als $\mathrm{AC_{pers}}$ bezeichnet wird.

In manchen Arbeiten wird die Anomaliekorrelation auch „*pattern correlation*“ genannt, da sie die Übereinstimmung der Muster der Anomalienfelder von Vorhersage und Analyse wiedergeben soll. Sie misst im Unterschied zu RV, MSE, etc. nicht den Magnituden- sondern den Phasen-Fehler. Systematische Fehler (bedingte und unbedingte Biase) haben also keine Auswirkung auf den Wert der AC, daher wird die AC auch oft als ein Maß für den „potentiellen *skill*“ und nicht für den wahren *skill* bezeichnet. Es ist jedoch festzuhalten, dass die AC trotz ihres verbreiteten Beinamens keine Aussage über die vorherrschenden räumlichen Muster in den Feldern trifft. Um Unterschiede in Bezug auf gegebene räumliche Strukturen und Skalen zu erfassen, müssen andere, oft komplexere Verifikationsmethoden angewendet werden.

Eine mathematische Beziehung zwischen $\mathrm{AC} = r_{o',y'}$ und RV lässt sich herstellen, wenn noch einmal die Zerlegung des MSE aus (3.7) betrachtet wird. Der MSE ist für die Anomalien gleich wie für die Absolutwerte von Vorhersagen und Beobachtungen und so lässt sich (3.7) direkt umschreiben zu

$$\mathrm{MSE} = \mathrm{B}^2 + \sigma^2_{y'} + \sigma^2_{o'} - 2\sigma_{y'}\sigma_{o'}r_{o',y'} \,. \tag{3.20}$$

Für eine Klimavorhersage vereinfacht sich diese Aufspaltung des MSE:

$$\mathrm{MSE_{clim}} = \mu^2_{o'} + \sigma^2_{o'} \tag{3.21}$$

Einsetzen von (3.20) und (3.21) in (3.17) führt zu dem gewünschten Zusammenhang zwischen RV und AC. Mit den folgenden Annahmen reduziert sich das Ergebnis zu einer einfachen Approximation (3.22): a) das Mittel über die Analyse-Anomalie ist nahe Null, b) der Bias ist vernachlässigbar und c) die Varianz des Vorhersagefeldes ist ungefähr gleich der des Analysefeldes.

$$\mathrm{RV} \approx 2\mathrm{AC} - 1 \tag{3.22}$$

Unter den angegebenen Voraussetzungen ist eine AC von 0.5 gleichbedeutend mit dem „Null-*skill*“ Level der RV, die oft in der Literatur für synoptisch nützliche Vorhersagen als untere Grenze angenommene AC = 0.6 mit einer RV von 0.2. In der Praxis werden jedoch häufig eine oder mehrere der Bedingungen a)–c) verletzt. Im Speziellen wird bei Verwendung eines unabhängigen Klimas oft das Mittel über die Analyse-Anomalien $o_k - c_k$ über einen relativ kurzen Zeitraum nicht nahe Null sein. Auch der Bias kann häufig nicht vernachlässigt werden. Daher ist für die meisten Auswertungen in dieser Arbeit die Approximation (3.22) nicht anwendbar und die Maße AC und RV sind nicht redundant.

3.4 Skill scores für binäre Prognosen

In der Literatur findet sich eine Fülle von unterschiedlichen Definitionen von *skill scores* für binäre Prognosen, da dieses Thema in der Verifikation ausgehend von der sogenannten „Finley Affäre“ (Murphy 1996) schon ab dem Ende des 19. Jahrhunderts diskutiert wurde. Finley (1884) hatte zur Evaluierung seiner binären Tornado-Vorhersagen mit der Unterscheidung „Tornado“ oder „kein Tornado“ ausschließlich den Prozentsatz der korrekten Prognosen (Percent Correct oder PC, auch Hit Rate genannt) berechnet.

$$\mathrm{PC} = \frac{a+d}{N} \tag{3.23}$$

Für seine täglichen Prognosen resultierte ein hoher PC von 96,6 %. Schnell wurde jedoch die Kritik laut, dass ein sogar noch höherer Wert erzielt worden wäre, wenn einfach immer „kein Tornado“ prognostiziert worden wäre. Gilbert (1884), Peirce (1884) und Doolittle (1885) publizierten innerhalb kurzer Zeit Vorschläge für geeignetere Maße der Vorhersagegüte. Im Laufe der Geschichte wurden die damals formulierten Maßzahlen häufig wieder aufgegriffen und auch mehrmals mit neuen Namen besetzt, sodass heute unterschiedliche Nomenklaturen in Verwendung sind, was mitunter zu Verwirrung führen kann. Da es keine verbindliche Festlegung der Benennungen gibt, kann auch in dieser Arbeit nur eine subjektive Auswahl der zu verwendenden Namen getroffen werden, es sollen jedoch auch die weiteren in der Literatur gebräuchlichen Bezeichnungen erwähnt werden.

Heidke Skill Score (HSS): Der HSS (weniger bekannt als *Cohen's kappa*) wurde ursprünglich eingeführt von Doolittle (1885), der als direkte Reaktion auf Finleys Tornado-Vorhersage den Percent Correct (3.23) in Bezug auf dessen Wert für eine Zufallsprognose stellte. Der HSS folgt demnach der allgemeinen Formulierung eines *skill scores* (3.15).

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{PC}_{\mathrm{rand}}}{\mathrm{PC}_{\mathrm{perf}} - \mathrm{PC}_{\mathrm{rand}}} \tag{3.24}$$

Ausgedrückt mit den Einträgen der Kontingenztabelle und denen einer Zufallsprognose (3.16) ergibt das

$$\mathrm{HSS} = \frac{(a+d)/N - [(a+b)(a+c)+(b+d)(c+d)]/N^2}{1-[(a+b)(a+c)+(b+d)(c+d)]/N^2}\,, \tag{3.25}$$

$$\mathrm{HSS} = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}\,. \tag{3.26}$$

Peirce Skill Score (PSS): Der PSS (auch genannt *True Skill Statistics* oder *Hansen-Kuiper's Discriminant*) ist ähnlich formuliert wie der HSS, jedoch steht im Nenner nicht der PC einer allgemeinen Zufallsprognose, sondern der einer Zufallsprognose ohne Bias.

$$\mathrm{PSS} = \frac{(a+d)/N - [((a+b)(a+c)+(b+d)(c+d)]/N^2}{1-[(a+c)^2+(b+d)^2]/N^2} \tag{3.27}$$

Anschaulicher ist der Zugang zum PSS durch seine Äquivalenz zur Differenz zwischen POD und POFD:

$$\mathrm{PSS} = \mathrm{POD} - \mathrm{POFD} = \frac{a}{a+c} - \frac{b}{b+d} = \frac{ad-bc}{(a+c)(b+d)} \tag{3.28}$$

Beim der PSS muss beachtet werden, dass wenn ein Zelleintrag der Kontingenztabelle sehr groß ist, der andere in derselben Spalte fast vernachlässigt wird.

Equitabe Threat Score (ETS): Der ETS (auch bekannt als *Gilbert Skill Score*) folgt der allgemeinen Form eines *skill scores* mit dem Threat Score TS (oder Critical Success Index) als zugrunde liegendem Fehlermaß:

$$\mathrm{ETS} = \frac{\mathrm{TS} - \mathrm{TS}_{\mathrm{rand}}}{\mathrm{TS}_{\mathrm{perf}} - \mathrm{TS}_{\mathrm{rand}}} \tag{3.29}$$

$$mit \quad \mathrm{TS} = \frac{a}{a+b+c}, \quad \mathrm{TS}_{\mathrm{perf}} = 1 \tag{3.30}$$

Ausgedrückt durch die Zelleinträge der Kontingenztabelle ergibt dies:

$$\mathrm{ETS} = \frac{a - a_{\mathrm{ref}}}{a - a_{\mathrm{ref}} + b + c}\,, \tag{3.31}$$

wobei a_{ref} die Anzahl der *hits* einer Zufallsprognose ist (vgl. 3.16):

$$a_{\mathrm{ref}} = \frac{(a+b)(a+c)}{N} \tag{3.32}$$

Odds Ratio Skill Score (ORSS): Der ORSS (oder *Yule's Q*) basiert auf dem Odds Ratio OR, einem Verifikationsmaß, das vor allem in der medizinischen Statistik häufige Verwendung findet. Unter *Odds* versteht man in diesem Zusammenhang die Wahrscheinlichkeit eines Ereignisses ausgedrückt durch eine Quote, welche definiert ist durch das Verhältnis der Wahrscheinlichkeit p zur ihrer komplementären Wahrscheinlichkeit $(1 - p)$. So hat zum Beispiel ein Ereignis mit einer Wahrscheinlichkeit von 0.8 einen *Odds*-Wert von $0.8/(1 - 0.8) = 4$, oder landläufig gesprochen die Quote 4 zu 1. OR beschreibt nun das Verhältnis der bedingten *Odds* eines *hits* zu den bedingten *Odds* eines *false alarms*. Ausdrückbar ist dies durch POD und POFD und vereinfacht ergibt sich OR als Verhältnis des Produkts der korrekten Vorhersagen zum Produkt der inkorrekten Vorhersagen:

$$\mathrm{OR} = \frac{\mathrm{POD}/(1 - \mathrm{POD})}{\mathrm{POFD}/(1 - \mathrm{POFD})} = \frac{ad}{bc} \tag{3.33}$$

Ein Odds Ratio von eins wird erhalten, wenn Prognosen und Beobachtungen statistisch unabhängig sind (Zufallsprognose), je besser die Prognose, desto höher der Wert von OR. Da das Odds Ratio für eine perfekte Prognose nicht definiert ist ($Nenner = 0$), kann nicht die allgemeine Formel zur Ableitung eines *skill scores* (3.15) verwendet werden. Dennoch kann mithilfe des Referenzniveaus von $\mathrm{OR}_{\mathrm{rand}} = 1$ eine Transformation auf einen *skill score* mit Wertebereich [-1,1] durchgeführt werden:

$$\mathrm{ORSS} = \frac{\mathrm{OR} - 1}{\mathrm{OR} + 1} = \frac{ad - bc}{ad + bc} \tag{3.34}$$

Der ORSS basiert nur auf den bedingten, nicht den marginalen Verteilungen. Durch seine Definition ergibt er oftmals sehr hohe *skill* Werte, die in ihrer Magnitude nicht mit den Ergebnissen anderer *skill scores* verglichen werden können. Sobald ein Element der Kontingenztabelle Null ist, werden künstlich die Extremwerte des ORSS erzeugt, die Ergebnisse sind nicht mehr interpretierbar. Problematisches Verhalten tritt auch auf, wenn ein Zelleintrag sehr klein ist.

Extreme Dependency Score (EDS): Der EDS wurde als erstes von Coles et al. (1999) definiert und von Stephenson et al. (2008) als Verifikationsmaß für seltene Ereignisse vorgeschlagen, da er durch seine logarithmische Abhängigkeit selbst für eine verschwindende Base Rate informative Werte liefert. Unter dem Logarithmus (log) wird hier und im Weiteren immer der natürliche Logarithmus verstanden.

$$\mathrm{EDS} = \frac{2\log((a+c)/N)}{\log(a/N)} - 1 \tag{3.35}$$

Der EDS hängt jedoch nicht explizit von b und d ab, er kann also beispielsweise für sehr unterschiedliche Prognosen denselben Wert liefern, solange a und c sowie die Summe von $b+d$ gleich bleiben. Alleine dadurch ist ersichtlich, dass der EDS alleine keinen ausreichenden Vergleich von Prognosequalität leisten kann.

Stephenson et al. (2008) beschreiben zur Veranschaulichung des EDS, in welcher Form die POD für $\mathrm{BR} \to 0$ für verschiedene Vorhersagesysteme (bei annähernd konstantem Bias) gegen Null konvergiert. Für Zufallsprognosen ist diese Konvergenz linear, für perfekte theoretische Prognosen bliebe POD für alle Base Rates konstant. Vorhersagen, deren PODs langsamer als Zufallsprognosen gegen Null konvergieren, ergeben positive EDS Werte.

Symmetric Extreme Dependency Score (SEDS): Hogan et al. (2009) führen eine transponiert symmetrische (vgl. Abschn. 3.4.1) Modifikation des EDS ein:

$$\mathrm{SEDS} = \frac{\log((a+b)/N) + \log((a+c)/N)}{\log(a/N)} = \frac{\log(a_{rand}/a)}{\log(a/N)} \tag{3.36}$$

mit a_{rand} aus (3.16). Der SEDS berücksichtigt somit auch die *misses* und *correct rejections*, behält aber dennoch das vorteilhafte Verhalten des EDS im Falle einer verschwindenden Base Rate bei, kann also sowohl für häufige als auch für sehr seltene Ereignisse verwendet werden.

Skill anhand der ROC: Die ROC (Relative Operating Characteristics) Kurve ist eine Anwendung der Signal Detection Theoroy (SDT), die sich mit der Fähigkeit befasst, zwischen zwei Möglichkeiten wie „Signal“ und „Signal+Rauschen“ zu unterscheiden (*discrimination*). Eine ROC-Ebene spannt die Werte von POD und POFD auf, welche für ein Vorhersagesystem durch die Variation des Schwellwerts kontinuierlich verändert werden können. Von der Form der so erstellten ROC-Kurve, bzw.

der Fläche unter ihr (AUC), kann auf die Prognosequalität rückgeschlossen werden: Eine ROC Kurve entlang der 45° Linie (AUC = 0.5) bedeutet, dass POD und POFD gleich sind, was einem Zufallsprozess entspricht. Diese Linie markiert also das „Null-*skill*“ Niveau. Für perfekte Prognosen verläuft die ROC-Kurve am äußeren linken und oberen Rand des Diagramms (AUC = 1). Empirische ROC-Kurven verlaufen konvex mit Anfangs- und Endpunkt bei [0,0] bzw. [1,1] und lassen sich aus vorgegebenen Normalverteilungen von Ereignis und nicht-Ereignis erstellen. Reale meteorologische Vorhersagesysteme produzieren eine ähnliche Form (Jolliffe und Stephenson 2003). Die Fläche unter der Kurve AUC kann als Maß für den *skill* verwendet werden. Es treten jedoch Schwierigkeiten in der praktischen Berechnung auf, wenn der ROC-Graph wie oben beschrieben durch Variation der Schwellwerte erstellt wird. Dann liegt jeder dieser abgewandelten Vorhersagen eine andere Base Rate zugrunde und es sollte deshalb nicht über diese integriert (gemittelt) werden (Jenkner et al. 2008).

Wie bereits erwähnt, können alle besprochenen 2x2 *skill scores* als Funktionen der drei Faktoren der *likelihood-base rate* Faktorisierung POD, POFD und BR ausgedrückt werden. ORSS und PSS sind alleine durch POD und POFD ausdrückbar und können daher als Kurvenscharen in der ROC-Ebene dargestellt werden (Abb. 3.3). Für die anderen *skill scores* müssen fixe Base Rates angenommen werden. Die resultierenden Graphen sind für HSS, ETS, EDS und PSS Geraden mit variierenden Steigungen und Schnittpunkten mit der y-Achse. Alle hier angeführten 2x2 *skill scores* außer dem EDS (und den Klima *skill scores* aus Abschn. 3.4.4) ergeben Null sobald POD = POFD, teilen also zumindest das „Null-*skill*“ Niveau mit der ROC. Die Kurvenscharen des SEDS verlaufen zwar konvex, schneiden jedoch die Ordinate an verschiedenen Punkten und sind nicht symmetrisch. Nur für den ORSS ergeben sich Kurven positiven *skills*, die der Form der empirischen ROC entsprechen, wie in Abb. 3.3 zu erkennen ist. Der ORSS kann also als Maß des *skills* im Sinne der ROC bzw. SDT angesehen werden.

Zusammenhänge zwischen den binären *skill scores*: Da die in diesem Abschnitt besprochenen *skill scores* nur von den vier Einträgen der Kontingenztabelle abhängen ist es offensichtlich, dass sie alle in Beziehung zueinander stehen. Sowohl HSS, PSS als auch ORSS sind proportional zum Term $(ad - bc)$. Besonders einfach ist der Zusammenhang zwischen dem HSS und dem ETS:

$$\mathrm{HSS} = \frac{2\mathrm{ETS}}{1 + \mathrm{ETS}} \tag{3.37}$$

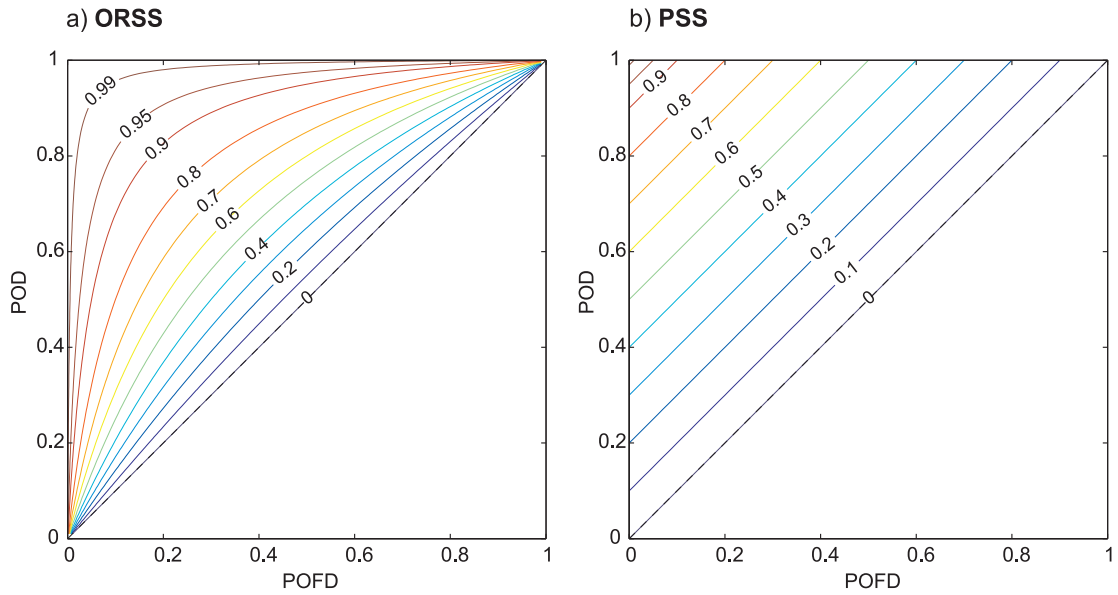

Abbildung 3.3: Positive Isoplethen von a) Odds Ratio Skill Score und b) Peirce Skill Score auf der ROC-Ebene.

3.4.1 Eigenschaften der Fehlermaße für dichotome Prädiktanden

Angesichts der Fülle an existierenden Möglichkeiten, eine binäre Vorhersagesituation zu verifizieren, drängt sich die Frage auf, welche von ihnen die geeignetste ist um Prognosegüte zu bewerten und zu reihen. Bedauerlicherweise lässt sich kein Fehlermaß als das allgemein „beste“ identifizieren. Jedes beschreibt für gewöhnlich unterschiedliche Aspekte des Verifikationsproblems. Die **Metaverifikation** beschäftigt sich damit, wünschenswerte Attribute einer Maßzahl zu definieren und zuzuordnen. In der Literatur herrscht teilweise große Uneinigkeit, einerseits schon in der Definition dieser Eigenschaften, vor allem jedoch in ihrer Bewertung. Bei Jolliffe und Stephenson (2003) werden die Attribute *consistency*, *equitability* und *regularity* als die hauptsächlichen Auswahlkriterien, die direkt relevant für Fehlermaße nicht-probabilistischer dichotomer Prognosen sind, angesehen. Diese und weitere Eigenschaften von *skill scores* werden im Folgenden genauer besprochen und sind bei möglicher Zuordnung für die verwendeten *skill scores* in Tabelle 3.2 zusammengefasst.

***Consistency*:** Ein Fehlermaß wird dann als konsistent mit einer Direktive (oder Vorschrift) für die Vorhersage bezeichnet, wenn es genau bei Erfüllung dieser maximiert wird. Zum Beispiel kann gezeigt werden, dass der PSS die höchsten Werte erreicht, wenn die vorhergesagte relative Häufigkeit der Base Rate entspricht. Das

heißt, der PSS ist konsistent mit der Vorschrift, eine „ja“-Vorhersage auszustellen, wenn die (erwartete) vorhergesagte Wahrscheinlichkeit die Base Rate übersteigt. Auch für die anderen *skill scores* lassen sich jene vorhergesagten Wahrscheinlichkeiten, bei denen sie maximiert werden – die *optimal threshold probabilities* –, finden (Jolliffe und Stephenson 2003). Diese weisen jedoch in den meisten Fällen Abhängigkeiten vom Maß selbst auf und lassen sich deshalb nicht so anschaulich auf damit konsistente Vorhersage-Vorschriften umleiten.

***Equitability*:** Es gab und gibt in der Literatur unterschiedliche Auffassungen von *equitability*. Als Referenz zur Darlegung dieses Konzepts werden meist Gandin und Murphy (1992) zitiert, die zwei Anforderungen an ein *equitable* Maß stellen:

1. Alle zufälligen Vorhersagesysteme, einschließlich konstanter Prognosen, werden gleich bewertet, typischerweise mit Null.
2. *Equitable* Maßzahlen müssen als gewichtete Linearkombinationen der Einträge der Kontingenztabelle formulierbar sein, wobei die Gewichte außer von der Base Rate nicht von der Kontingenztabelle abhängen.

Gemäß dieser strengen Definition wird der PSS als einziges Maß als *equitable* klassifiziert. Für die meisten Betrachtungen wird jedoch die erste Bedingung als Schlüssel zur *equitability* angesehen und die zweite, welche hauptsächlich der mathematischen Vereinfachung dient, verworfen. So werden auch nicht-lineare Maßzahlen eingeschlossen, die andere wünschenswerte Eigenschaften aufweisen, wie die bessere Verwendbarkeit für seltene Ereignisse.

Anforderung 1 ergibt sich durch den gerechtfertigten Anspruch an ein Verifikationsmaß, dass es die Leistung unterschiedlicher Vorhersagesysteme richtig reihen kann. Man kann zeigen, dass wenn Bedingung 1 nicht erfüllt ist, in manchen Fällen eine Zufallsvorhersage besser klassifiziert wird, als eine, die *skill* aufweist (vgl. Hogan et al. 2010). Alle besprochenen 2x2 *skill scores* erfüllen Bedingung 1, wenn als Zufallsprognose die aus der Kontingenztabelle laut (3.16) definierbare verstanden wird. Daher werden oft, z. B. bei Jolliffe und Stephenson (2003), sowohl PSS und HSS als auch ORSS und ETS als *equitable* bezeichnet. Bei Hogan et al. (2010) wird jedoch klargestellt, dass genau genommen Bedingung 1 für alle möglichen Zufallsprognosen gelten muss, was für ETS und ORSS nur bei unendlich großen Stichproben der Fall ist. Es wird daher zwischen Maßen, die *truely equitable* sind, wie PSS und HSS, und solchen, die *asymptotically equitable* sind, wie ETS und ORSS, differenziert.

***Regularity*:** Das Konzept der *regularity* kommt aus der *Signal Detection Theory* und bezieht sich auf die Form der ROC Kurve. Ein regulärer ROC Graph ist vollständig, geht durch [0,0] und [1,1] und schneidet an keinem anderen Punkt die Achsen. Eine restriktivere Definition fordert zusätzlich die Konvexität der Kurve. Jolliffe und Stephenson (2003) argumentieren, dass sowohl empirische ROC Kurven als auch solche realer Vorhersagesysteme, die auf einem konstanten *skill* Niveau operieren, eine reguläre Form haben und dass daher die *regularity* eine vernünftige Anforderung an ein Maß des *skills* ist. In dieser Argumentation steckt jedoch durch die Annahme, die Form von realen ROC Kurven konstanten *skills* zu kennen, schon eine bestimmte Definition von *skill*. Es kann hinterfragt werden, ob es gerechtfertigt ist, diese dann zu verwenden, um auf die allgemeine Güte aller *skill scores* rückzuschließen. Wie oben beschrieben, ist von den angeführten Größen nur der ORSS regulär, mit Ausnahme der äußersten Extremwerte von [0,0] und [1,1], die bei ORSS nicht definiert sind (Abb. 3.3).

Invarianz gegenüber *hedging* und transponierte Symmetrie: Eine weiterer Gesichtspunkt, der in der Metaverifikation diskutiert wird, ist die Problematik von *hedging*, also die bewusste Lenkung einer Prognose in eine gewisse Richtung entgegen der „wahren Einschätzung“ der Lage, mit der Absicht, eine bessere Bewertung durch ein bestimmtes Verifikationsmaß zu erreichen. Es wäre vorteilhaft, ein Fehlermaß definieren zu können, das vollkommen insensitiv auf *hedging* ist. In der Praxis ist dies jedoch schwer zu erreichen, da sich alle *skill*-Maße prinzipiell für eine bestimmte Vorhersage-Wahrscheinlichkeit optimieren lassen. Durch die einfachste Art von *hedging* verbesserbar sind Maßzahlen, die *misses* und *false alarms* unterschiedlich bewerten. Wenn die *misses* stärker bestraft werden, ermutigt das Maß eher dazu, das Ereignis seltener als erwartet vorherzusagen (*underforecasting*), im umgekehrten Fall zu *overforecasting*. Daher kann **transponierte Symmetrie**, also die gleiche Gewichtung der *misses* und *false alarms*, in vielen Fällen als wünschenswerte Eigenschaft eines Verifikationsmaßes angesehen werden und wird von den meisten der hier behandelten *skill scores* erfüllt. Eine Ausnahme ist der PSS, bei dem jedoch die unterschiedliche Bewertung der *misses* und *false alarms* von den *hits* und *correct rejections* abhängt. Ist $a > d$, so wird *overforecasting* stärker bestraft, im umgekehrten Fall *underforecasting*. Zum Beispiel verschlechtert im Falle von sehr seltenen Ereignissen, bei denen in sinnvollen Prognosen $d >> a$ ist, ein zu seltenes Vorhersagen den PSS stärker als ein zu häufiges Vorhersagen. So verhindert der PSS auch ohne transponierte Symmetrie durch seine Formulierung die einfachste Form des *hedging*. Es soll angemerkt werden, dass es auch Situationen gibt, in denen es Sinn macht, die beiden Falschvorhersagen unterschiedlich zu bewerten, z. B. wenn

ökonomische Schäden durch eine nicht-Vorhersage eines aufgetretenen Ereignisses höher eingeschätzt werden als durch eine Falschvorhersage eines nicht aufgetretenen Ereignisses.

Manchmal wird unter *hedging* auch das gezielte Vorhersagen eines Ereignisses mit seiner klimatologischen Häufigkeit, also auf längere Sicht das Anstreben eines ausgeglichenen Bias, verstanden. Der Einfluss dieser Art von *hedging* auf die Maßzahlen ist jedoch von vielen Faktoren abhängig (vgl. Abschn. 3.4.2) und kann sogar zu schlechteren Werten führen als für die ursprünglichen Prognosen erreicht wurden, wie zum Beispiel bei HSS, PSS, ETS und ORSS für Finley's Tornadoprognosen (vgl. Stephenson 2000). Von den Angeführten zeigt der ORSS die geringste Veränderung durch diese Art von *hedging*.

Sinnhaftigkeit für seltene Ereignisse (BR $\to 0$): Für sehr seltene Ereignisse weisen ETS, HSS, PSS und ORSS ein problematisches Verhalten auf. Stephenson et al. (2008) beschreiben im Detail, wie die einzelnen Maße gegen einen nicht-informativen Grenzwert streben, wenn die Base Rate gegen Null geht. ETS, HSS und PSS nähern sich in diesem Fall dem Nullniveau an, während ORSS gegen seinen Maximalwert von eins strebt (oder im Fall von Vorhersagen mit negativem *skill* gegen -1). Der ETS wird für eine verschwindende Base Rate zum TS, der PSS zur POD und der HSS zu $2(1 + \mathrm{TS}^{-1})^{-1}$, wobei diese Beziehungen nur bei positivem *skill* gelten. Für die Verifikation sehr seltener Ereignisse empfehlen Stephenson et al. (2008) daher die Verwendung des EDS, der für $\mathrm{BR} \to 0$ gegen für verschiedene Vorhersagesysteme unterscheidbare, endliche Werte strebt. Hogan et al. (2009) legen dar, dass der EDS weder *(asymptotisch) equitable* noch *transponiert symmetrisch* ist und definiert als Alternative den SEDS, der diese Eigenschaften erfüllt, während er dasselbe Grenzverhalten bei $\mathrm{BR} \to 0$ aufweist wie der EDS.

Linearität: Hogan et al. (2009) definieren lineare Maße so, dass sie durch eine Änderung eines *miss* zu einem *hit* bei gleich bleibender relativer Häufigkeit von Beobachtung und Vorhersage (z. B. durch das Addieren von eins zu a und d bei gleichzeitigem Subtrahieren von eins von b und c) immer um den gleichen Wert verbessert werden, unabhängig von den Ausgangsbeträgen von a–d. Diese Definition steht im Einklang mit der von Gandin und Murphy (1992) als zweite Bedingung für *equitibility* beschriebenen, außer dass die Gewichte der Linearkombination in der weniger strengen Definition von Hogan et al. (2009) auch von $p(y_{ja})$ abhängen dürfen. Der ORSS ist beispielsweise stark nichtlinear und steigt bei der oben beschriebenen Änderung der Kontingenztabelle in den niedrigeren Wertebereichen schneller

als lineare *scores* (z. B. HSS oder PSS), bei höheren Ausgangswerten langsamer.

Komplementäre Symmetrie: Ein komplementär symmetrisches Verifikationsmaß bewertet das Auftreten des Ereignisses gleich wie dessen nicht-Auftreten. Die Zeilen und Spalten der Kontingenztabelle können also vertauscht werden, ohne dass sich der erreichte Wert ändert. Die Beurteilung des Überschreitens eines Schwellwerts liefert für eine solche Größe dasselbe Ergebnis wie die des Unterschreitens des Schwellwerts.

Eigenschaft	HSS	PSS	ETS	ORSS	EDS	SEDS
True equitibility	✓	✓	-	-	-	-
Asymptotical equitibility	✓	✓	✓	✓	-	✓
Regularity (ROC)	-	-	-	✓	-	-
Transponierte Symmetrie ($b \leftrightarrow c$)	✓	-	✓	✓	-	✓
Sinnhaftigkeit bei BR $\rightarrow 0$	-	-	-	-	✓	✓
Linearität	✓	✓	-	-	-	-
Komplementäre Symmetrie ($y \leftrightarrow o$)	✓	✓	-	✓	-	-
Kann für Vorhersagen mit Bias perfekt werden	-	-	-	✓	✓	-

Tabelle 3.2: Eigenschaften verschiedener 2x2 *skill scores*, in Anlehnung an Hogan et al. (2009). Häkchen bedeuten das Erfüllen der Eigenschaft, Minuszeichen das nicht-Erfüllen.

3.4.2 Bias-Abhängigkeit und Adjustierungen

Wie in 3.4.1 erwähnt, weisen alle besprochenen 2x2 *skill scores* eine Abhängigkeit vom systematischen Fehler der Vorhersage auf. Speziell für den Threat Score und den ETS kann gezeigt werden, dass Vorhersagen mit einem höheren Bias tendenziell einen höheren *skill* erreichen (Mason 1989; Hamill 1999). Der Einfluss des Bias auf den ETS ist jedoch keineswegs linear, sondern hängt von anderen Faktoren wie der Base Rate ab und kann meist nicht direkt quantifiziert werden. Deshalb ist die Verwendung des ETS zum Vergleich unterschiedlicher Prognosesysteme nur dann zulässig, wenn diese einen ähnlichen Bias aufweisen (Hamill 1999). Um diesem Problem Abhilfe zu schaffen, ist es sinnvoll, den Bias getrennt von der restlichen, darüber hinausgehenden Übereinstimmung zu betrachten, welche auch als potentieller *skill* oder Phasen-Übereinstimmung bezeichnet werden kann. Für auf kontinuierlichen

Variablen beruhende, durch Schwellwerte auf dichotome Ereignisse reduzierte Vorhersagen gibt es verschiedene Ansätze, den Bias hierzu aus den unterschiedlichen *skill scores* zu eliminieren (Casati et al. 2004; Jenkner et al. 2008; Mesinger 2008). Diese Methoden beruhen auf einer Rekalibirierung der Prognose zu einer Bias-freien Kontingenztabelle, also einer Gleichstellung der vorhergesagten und der beobachteten relativen Häufigkeit des Ereignisses.

Bei Casati et al. (2004) wird zu diesem Zweck vorgeschlagen, den Schwellwert, der ein beobachtetes Ereignis definiert, auf dem festgelegten Wert zu belassen, denjenigen für eine „ja"-Prognose aber so zu verschieben, dass die marginalen Häufigkeiten übereinstimmen. Ausdrückbar ist diese nichtlineare Transformation durch die folgende Rakalibirierungsfunktion:

$$y' = F_o^{-1}(F_y(y)), \tag{3.38}$$

wobei o die Analyse, y die Vorhersage, y' die rekalibrierte Vorhersage und F_o sowie F_y die empirischen kumulativen Verteilungsfunktionen von jeweils Analyse und Vorhersage sind. In dieser Arbeit wird die Methode aus (3.38) zum Test näherungsweise angewandt, das Hauptaugenmerk liegt jedoch auf dem im nächsten Abschnitt (3.4.3) beschriebenen Ansatz.

3.4.3 Quantilsbasierte scores nach Jenkner et al. (2008)

Eine andere Herangehensweise, die den Einfluss des Bias aus dem *skill score* entfernt, ist es, den Schwellwert, der ein Ereignis definiert, über die Quantile der Verteilung des zu verifizierenden Datensatzes festzulegen (Jenkner et al. 2008). So variieren zwar die Absolutwerte der Schwellen sowohl zwischen Beobachtung und Vorhersage als auch zwischen unterschiedlichen zu verifizierenden Situationen, nicht jedoch ihre Quantilswerte. Die Überschreitungshäufigkeit wird für Analyse und Vorhersage gleich festgelegt $p \equiv p(o_{ja}) = p(y_{ja})$, also $b = c$. Für die so modifizierte Kontingenztabelle lässt sich der PSS vereinfachen zu

$$\mathrm{PSS}_{qb}(p) = 1 - \frac{c}{(p - p^2)N}\,. \tag{3.39}$$

Jenkner et al. (2008) empfehlen die Verwendung des $\mathrm{PSS}_{\mathrm{qb}}$ gemeinsam mit der Quantilsdifferenz QD zwischen Modell und Analyse, welche ein Maß für den Bias ist:

$$\mathrm{QD}(p) = q_y(p) - q_o(p), \tag{3.40}$$

wobei die $q_{y,o}(p)$ die Quantilswerte, also diejenigen Werte bezeichnen, die in Vorhersage und Analyse mit der Häufigkeit p überschritten werden. Es ist außerdem sinnvoll, die QD(p) mit dem arithmetischen Mittelwert von $q_y(p)$ und $q_o(p)$ zu normieren. Die so berechenbare standardisierte Quantilsdifferenz QD'(p) ist im Gegensatz zur absoluten QD dimensionslos und auf das Intervall von $[-2, +2]$ beschränkt.

Der große Vorteil dieser Methode ist, dass die zugrunde liegenden relativen Häufigkeiten p immer bekannt sind, und es dadurch möglich ist, durch ein gewichtetes Mittel ein integrales Gesamtmaß über alle Schwellwerte zu bilden. Für die QD' lautet dieses

$$\overline{\mathrm{QD}'} = \frac{1}{\int w(p)dp} \int_0^1 w(p)|\mathrm{QD}'(p)|dp\,, \tag{3.41}$$

$$w(p) = \frac{q_o(p) + q_y(p)}{2}\,. \tag{3.42}$$

Ähnlich formuliert ist der integrale $\mathrm{PSS_{qb}}$, nur wird hier für die Gewichtungsfunktion $w(p)$ nicht der arithmetische, sondern der geometrische Mittelwert von $q_o(p)$ und $q_y(p)$ verwendet:

$$\overline{\mathrm{PSS}_{qb}} = \frac{1}{\int w(p)dp} \int_0^1 w(p)\mathrm{PSS}_{qb}(p)dp \tag{3.43}$$

$$w(p) = \sqrt{q_o(p)q_y(p)} \tag{3.44}$$

Ein Nachteil der Verwendung dieser Maße ist, dass Quantile nicht so intuitiv zugänglich sind wie Absolutwerte und für die praktische Anwendung oft eine Verifikation der Überschreitung eines bestimmten Schwellwerts (z. B. $\mathrm{RR} > 5\,\mathrm{mm}$) gefordert wird.

Bei der Auswertung für den Niederschlag muss des Weiteren darauf geachtet werden, dass oft erst relativ hohe Quantile Regenmengen von mehr als 0.1 mm, also messbaren Niederschlag, beziffern (siehe Abb. 4.9). Aussagen über die Überschreitung geringerer Werte sind jedoch wenig sinnvoll und daher werden im Folgenden nur diejenigen p betrachtet, bei denen q_o und (der Großteil der) $q_y \geq 0.1\,\mathrm{mm}$ sind.

Eine Einschränkung für die Anwendung aller Bias-Adjustierungen ist, dass der Datensatz groß genug sein muss, da bei nicht ausreichender Population der Kontingenztabelle der Bias oft nicht vollständig eliminiert werden kann. Anders gesagt, es können in diesen Fällen keine zwei Werte gefunden werden, die jeweils von exakt

gleich vielen Prognosen wie Analysen übertroffen werden. Daher werden die quantilsbasierten und Bias-adjustierten *scores* in dieser Arbeit nur für Zeitspannen von mindestens einem Monat berechnet.

3.4.4 Einbindung des Klimas in die 2x2 skill scores

Skill scores dichotomer Prädiktanden verwenden als Referenz für gewöhnlich eine aus der Kontingenztabelle produzierbare, von den Verteilungen abhängige Zufallsprognose (3.16) und beziehen keine Klimawerte mit ein.

Hamill und Juras (2006) geben jedoch einen Hinweis darauf, dass die in der Formulierung einiger *skill scores* enthaltene Wahrscheinlichkeit des Ereignisses aus der Stichprobe (die Base Rate) durch seine klimatologische Wahrscheinlichkeit ersetzt werden kann und schlagen damit eine Einbindung eines langjährigen Klimas in 2x2 *skill scores* vor. Möglich ist diese Art der Modifizierung für den ETS und den HSS, indem in ihren Definitionen die Base Rate durch die klimatologische Wahrscheinlichkeit $p_{clim}(o_{ja})$ ersetzt wird, nicht aber für PSS und ORSS, da diese nicht von der Base Rate abhängen.

Eine andere Möglichkeit, das langjährige Klima in den ETS und den HSS einzubauen ist es, dieses anstatt der Zufallsprognose als Referenz zu verwenden. Dazu wird eine Kontingenztabelle erstellt, bei der anstatt des Modelloutputs die klimatologischen Mittel als Prognosen angenommen werden. Aus deren Einträgen a_{clim}, b_{clim}, c_{clim}, d_{clim} kann das jeweilige dem *skill score* zugrunde liegende Fehlermaß für eine Klimareferenzprognose berechnet und in die allgemeine Formel eingesetzt werden. Der auf diese Art auf das Klima bezogene ETS lautet also

$$\text{ETS}_{\text{clim}} = \frac{\text{TS} - \text{TS}_{\text{clim}}}{1 - \text{TS}_{\text{clim}}} \tag{3.45}$$

mit dem Threat Score für eine Klima-Prognose

$$\text{TS}_{\text{clim}} = \frac{a_{\text{clim}}}{a_{\text{clim}} + b_{\text{clim}} + c_{\text{clim}}}\,. \tag{3.46}$$

Analog kann der HSS_{clim} durch PC und PC_{clim} definiert werden:

$$\text{HSS}_{\text{clim}} = \frac{(a+d)/N - [(a_{\text{clim}} + d_{\text{clim}})]/N}{1 - [(a_{\text{clim}} + d_{\text{clim}})]/N} \tag{3.47}$$

Diese Modifizierungen von ETS und HSS erscheinen auf den ersten Blick intuitiv und naheliegend, jedoch werden durch sie die Eigenschaften der Maße grundlegend

verändert. Einerseits wird durch die Umformulierungen (Klima-) Information hinzugenommen, andererseits wird Information über die Übereinstimmung zwischen Prognose und Beobachtung weggenommen. Werden die Klimareferenzprognosen als fest angenommen, so ist der $\mathrm{ETS}_{\mathrm{clim}}$ bzw. der $\mathrm{HSS}_{\mathrm{clim}}$ eine nichtlineare Abbildung des TS bzw. PC und weist daher in Bezug auf die in Tabelle 3.2 angeführten Merkmale eher das Verhalten der zugrunde liegenden Maßzahl auf. Im Gegensatz zu ihren klassischen Formulierungen sind beide nicht mehr (asymptotisch) *equitable*, beim $\mathrm{HSS}_{\mathrm{clim}}$ entfällt die Linearität und der $\mathrm{ETS}_{\mathrm{clim}}$ kann auch für Prognosen mit Bias perfekt werden. Trotz des Verlusts der *equitabillity* können der $\mathrm{ETS}_{\mathrm{clim}}$ und der $\mathrm{HSS}_{\mathrm{clim}}$ aus (3.45) und (3.47), interpretierbar als Verbesserung des Threat Score bzw. des Percent Correct eines Modells gegenüber der Klimareferenzprognose, für die Verifikation von Nutzen sein und sollen auch in dieser Arbeit evaluiert werden. Die von Hamill und Juras (2006) vorgeschlagenen Formulierungen werden nicht weiter ausgewertet.

3.5 Auswertungsmöglichkeiten für den Wind

Der horizontale Wind ist eine vektorielle Größe $\vec{v} = (u, v)$ und kann daher nicht mit denselben Verifikationsmethoden ausgewertet werden wie die skalaren Variablen Druck, Temperatur, etc.

Eine besonders einfache Möglichkeit wäre es, die beiden horizontalen Windkomponenten als skalare, unbeschränkte, kontinuierliche Prädiktanden, getrennt voneinander zu verifizieren. Eine solche Aufspaltung ist jedoch physikalisch und für die Interpretation nur bedingt sinnvoll, da u und v hochgradig voneinander abhängig sind und für Anwender oft eine Information über Windgeschwindigkeit *ff* und Windrichtung *dd* anschaulicher ist.

Aus diesem Grund werden in dieser Arbeit *ff* und *dd* einzeln ausgewertet. Die Windgeschwindigkeit lässt sich als skalare, positiv definite Größe mit den 2x2 Maßen für verschiedene Schwellwerte evaluieren, auch die meisten *scores* für kontinuierliche Prädiktanden sind anwendbar. Eine Einschränkung ist durch die relativ starke Abweichung von der Normalverteilung durch die Beschränkung nach unten gegeben. Üblicherweise werden für *dd* und *ff* nur Winde mit einer Geschwindigkeit von $\geq 1\,\mathrm{m/s}$ ausgewertet, da die Messgenauigkeit ungefähr in diesem Bereich liegt.

Etwas vorsichtiger muss man an die Auswertung der Windrichtung herangehen. Durch ihre vektoriellen Eigenschaften ist eine skalare Mittelung auf dem Wertebereich von $[0°, 360°]$ nicht sinnvoll möglich.Ersichtlich wird dies am einfachen Beispiel

von $\overline{(350°, 10°)} = 180°$ und $\overline{(0°, 20°)} = 10°$ (wo der Überstrich für die arithmetische Mittelung steht), bei dem die zu mittelnden Richtungen nur um 10° nach rechts verschoben sind, die beiden Ergebnisse jedoch um 170° nach links. Daher wird für die Windrichtung eine komponentenweise Mittelung herangezogen, das heißt u und v werden getrennt gemittelt und dann auf *dd* zurück gerechnet. Die Bildung von Differenzen für die Windrichtung muss so definiert werden, dass nur Ergebnisse eines Absolutwerts $\leq 180°$ möglich sind, also immer der spitze Winkel zwischen den beiden Windvektoren verwendet wird. Das Vorzeichen gibt die Drehrichtung an: positive Werte bedeuten eine Rechtsablenkung des ersten Terms im Vergleich zum Subtrahenden, negative Werte eine Linksablenkung. Werden diese Richtlinien zur Berechnung von Mittelwerten und Differenzen eingehalten, so können die Verifikationsmaße für kontinuierliche Prädiktanden auch für die Windgeschwindigkeit ausgewertet werden.

Überdies bietet sich die Verwendung von 2x2 *scores* mit einer Einteilung in die acht Hauptwindrichtungen an. Die „ja"-Einträge der Kontingenztabelle werden dabei nicht nur durch die Überschreitung eines Schwellwertes, sondern durch ein bestimmtes Werteintervall gegeben, z. B. 247.5°–292.5° für Westwind.

Es bleibt anzumerken, dass die hier verwendete Vorgehensweise, die Windrichtung und Windgeschwindigkeit getrennt zu behandeln, insbesondere zu mitteln nicht in jedem Fall die exakteste Strategie ist, da die beiden Größen nicht unabhängig voneinander sind. Zum Beispiel werden bei Gruber (2005) Verifikationsmaße für den gesamten Windvektor besprochen sowie eine Vektormittelung angegeben, bei der ein Korrelationsterm zwischen Richtung und Geschwindigkeit eingeht und so die sehr kleinen Windgeschwindigkeiten nicht im Vorhinein weggelassen werden müssen.

3.6 Unsicherheit eines Verifikationsmaßes

Werden zur Bewertung einer Vorhersage absolute Werte von Verifikationsmaßen herangezogen, so bleibt die Einschätzung der Vorhersagesituation eigentlich unvollständig, da keine Aussage über die Unsicherheit, Vertrauenswürdigkeit oder statistische Signifikanz der Werte gegeben ist. Zur Quantifizierung der Unsicherheit von Verifikationsmaßen wird in den meisten Anwendungen aus der Literatur, wie auch in dieser Arbeit, allein der Stichprobenfehler aus der Statistik berücksichtigt. Eine Ausnahme bilden Gorgas und Dorninger (2012a) und Gorgas und Dorninger (2012b), die Wege aufzeigen, die Unsicherheit der verwendeten Analysedaten zu ermitteln (u. a. durch Analyse Ensembles) und die daraus resultierende Unsicherheit

der Verifikationsergebnisse zu berechnen.

Kehrt man wieder zur ausschließlich (frequentistischen) statistischen Sichtweise zurück, so kann man einen Verifikationsdatensatz als nur eine von vielen möglichen Stichproben einer Population (Grundgesamtheit) ansehen, das berechnete Verifikationsmaß $\widehat{\Theta}$ somit als Schätzwert des „wahren“, unbekannten aber festen, zugrunde liegenden Werts Θ (Jolliffe 2007). Werden zur Berechnung nicht genügend statistisch unabhängige Daten herangezogen, so ist der erhaltene Wert nicht aussagekräftig, ein gutes Ergebnis kann alleine aufgrund von zufälligen Schwankungen erreicht werden. Der so entstehende Fehler wird oft als Stichprobenfehler (*sampling uncertainty*) bezeichnet.

Die Unsicherheit einer Größe kann durch Konfidenzintervalle für ein gewähltes Konfidenzniveau $(1 - \alpha)$, meist 95 % oder 99 %, quantifiziert werden. Die korrekte, oft nicht differenzierte Interpretation eines frequentistischen 95 %-Konfidenzintervalls ist, dass bei Konstruktion von unendlich vielen Konfidenzintervallen 95 % dieser den Parameter Θ enthalten würden. Die einfachere Interpretation, Θ läge mit einer Wahrscheinlichkeit von 95 % im Intervall, gilt nur bei Verwendung eines Bayes'schen Ansatzes, welcher Θ als Zufallsvariable ansieht und A-priori-Wissen über die Verteilung von Θ voraussetzt, hier aber nicht weiter behandelt werden soll.

Es gibt auch in der gewöhnlichen frequentistischen Herangehensweise eine Fülle von verschiedenen Möglichkeiten Konfidenzintervalle zu konstruieren. In dieser Arbeit wird nur eine parametrische Methode mit Fisher-z-Transformation für die Anomaliekorrelation (Abschn. 3.6.1) sowie eine nicht-parametrische Permutationsmethode für die 2x2 *scores* verwendet und zum Vergleich das Bootstrapping beschrieben (beides Abschn. 3.6.2).

Formal soll die Unterscheidung zwischen dem aus der Stichprobe berechneten Verifikationsmaß $\widehat{\Theta}$ und dem „wahren“ Wert Θ nur in diesem Abschnitt gemacht werden. Ansonsten werden die berechneten Verifikationsmaße ohne Dach-Akzent geschrieben.

Einschränkungen der Unabhängigkeitsannahme

Eine Annahme, die bei allen beschriebenen Methoden gemacht wird, ist die der Unabhängigkeit der zugrunde liegenden Daten. Diese Grundannahme kann nicht immer vollständig erfüllt werden.

- *Räumliche und zeitliche Korrelation:* Durch die starke räumliche Korrelation der Felder kann nicht jeder Gitterpunktswert als unabhängig angesehen

werden, sondern es muss die Vorhersage für ein Gebiet immer zusammen betrachtet werden. Das heißt, als ein Vorhersage-Beobachtungspaar sind nicht die Werte eines Gitterpunkts anzusehen, sondern die Menge aller Vorhersage- und Beobachtungswerte eines Termins.

Auch zeitlich aufeinander folgende Analysen und Vorhersagen weisen oft eine starke Korrelation auf, was alleine schon an der Güte einer Persistenzprognose (vgl. Abschn. 4.6) erkennbar ist. Die Stärke der Abhängigkeit hängt von der betrachteten Variablen und dem zeitlichen Abstand ab. Für den Niederschlag beispielsweise ist die Forderung der Unabhängigkeit bei einem 24-Stunden-Abstand gut erfüllt (getestet für andere Daten bei Hamill 1999). Bei anderen meteorologischen Größen und kürzeren Zeitintervallen ist sie jedoch aufgrund der zeitlichen Korrelation teilweise verletzt.

- *Korrelation zwischen den Modellen:* Beim Vergleich zweier Vorhersagesysteme Modell 1 und Modell 2 steht meist nicht die Unsicherheit der jeweils erreichten Verifikationsmaße $\widehat{\Theta}_1$ und $\widehat{\Theta}_2$ im Vordergrund, sondern die Frage, ob sich die beiden statistisch signifikant voneinander unterscheiden. Dass diese beiden Aspekte nicht notwendigerweise dasselbe Ergebnis liefern liegt daran, dass die Vorhersagen unterschiedlicher Modelle für denselben Zeitpunkt im Normalfall hochgradig korreliert sind. Überlappen sich also die getrennt berechneten Konfidenzintervalle der *scores* zweier Modelle, so muss das noch nicht heißen, dass ihr Unterschied nicht statistisch signifikant ist (ein Beispiel findet sich bei Jolliffe 2007). Die zu vergleichenden Prognosen sollten also gemeinsam betrachtet werden. Jolliffe (2007) empfiehlt daher die Berechnung von Konfidenzintervallen direkt für die Differenz $\widehat{\Theta}_1 - \widehat{\Theta}_2$, anstatt für die einzelnen *scores*. Beschrieben ist diese Vorgehensweise hier nur für den Permutationstest gemäß Hamill (1999) (Abschn. 3.6.2).

3.6.1 Parametrische Methoden

Parametrische Intervalle basieren auf Annahmen über die Verteilung der Maße für die zugrunde liegende Population. Zumeist wird eine annähernde Gauß-Verteilung von $\widehat{\Theta}$ angenommen, mit dem wahren Wert Θ als Mittelwert und einer von der gewählten Maßzahl abhängigen Varianz. Aus den für die Normalverteilung bekannten Quantilswerten kann das Konfidenzintervall konstruiert werden. Die Herleitung eines solchen für Verhältniszahlen wie POD und POFD ist bei Agresti und Coull (1998) nachlesbar. Nimmt man Unabhängigkeit von POD und POFD an, so kann daraus ferner ein angenäherter Standardfehler für den PSS abgeleitet werden (Stephenson

2000). Für die meisten *skill scores* ist die Verteilung allerdings gar nicht bekannt, für manche ist sie es näherungsweise, entspricht jedoch nicht der Normalverteilung. Für letztere können fallweise Transformationen Abhilfe schaffen.

Fisher-z-Transformation

Bei nahezu normalverteilten Analyse- und Vorhersagedaten ist der Korrelationskoeffizient und folglich auch die AC rechtssteil unimodal verteilt. Mithilfe der Fisher-z-Transformation (3.48) können diese Maße auf annähernde Gauß-verteilung gebracht werden.

$$z_{AC} = \frac{1}{2}\log\left(\frac{1+\mathrm{AC}}{1-\mathrm{AC}}\right) = \operatorname{arctanh}(\mathrm{AC}) \tag{3.48}$$

Die genäherte Verteilung des Schätzwerts ($z_{\widehat{AC}}$) hat den Mittelwert z_{AC} und die Standardabweichung $(N-3)^{-1/2}$. Daher kann ein Konfidenzintervall $[\zeta_l, \zeta_u]$ für die $z_{\widehat{AC}}$ mit Konfidenzniveau von $(1-\alpha)$ wie folgt gewählt werden:

$$\zeta_{u,l} = z_{\widehat{AC}} \pm z_{\alpha/2}(N-3)^{-1/2} \tag{3.49}$$

wobei $z_{\alpha/2}$ hier das Quantil der Standardnormalverteilung (Mittelwert 0, Varianz 1) bezeichnet, das mit der Wahrscheinlichkeit von $\alpha/2$ überschritten wird und nicht mit dem z für die Fisher-z-Transformierte zu verwechseln ist.

Das Konfidenzintervall für die Transformierte z_{AC} kann schlussendlich durch Invertierung von (3.48) auf ein Konfidenzintervall $[\xi_l, \xi_u]$ für die AC umgelegt werden.

$$\xi_{u,l} = \tanh(\zeta_{u,l}) \tag{3.50}$$

Ein fast äquivalentes Vorgehen ist für den 2x2 *skill score* ORSS möglich, da der Logarithmus des OR (logOR) nahezu Gauß-verteilt ist und logOR außerdem zweimal der Fisher-z-Transformation des ORSS entspricht (was aus einfachen Umformungen von (3.34) ersichtlich ist). Die Standardabweichung des logOR ist gegeben durch $1/\sqrt{n_h}$ mit $1/n_h = 1/a + 1/b + 1/c + 1/d$.

3.6.2 Nicht-parametrische Methoden

Ohne Annahme über die Verteilung von $\widehat{\Theta}$ kommen nicht-parametrische *Resampling* Methoden aus. Es sollte jedoch angemerkt werden, dass über die hier beschriebenen nicht-parametrischen Verfahren hinaus, auch parametrisches *Resampling* (Boot-

strapping) existiert.

Bootstrapping

Ein sehr verbreitetes *Resampling* Verfahren ist das Bootstrapping. Aus dem gesamten Verifikationsdatensatz von N Vorhersage-Beobachtungspaaren werden wiederholt Bootstrap-Stichproben von derselben Größe N zufällig „mit Zurücklegen“ gezogen. Für jede dieser Bootstrap Stichproben wird das Verifikationsmaß $\widehat{\Theta}^*$ berechnet. Bei oftmaliger Wiederholung – es wird meist eine Anzahl von Bootstrap Stichproben von $n_{\text{boot}} = 1000$ empfohlen (Ferro 2007) – entsteht eine Verteilung von $\widehat{\Theta}^*$-Werten, aus der auf verschiedene Arten das Konfidenzintervall bestimmt werden kann. Eines der einfachsten ist das Perzentil-Intervall, bei dem für ein Intervall mit dem Konfidenzniveau $(1-\alpha)$, der $(n_{\text{boot}}\alpha/2)$-kleinste und der $(n_{\text{boot}}\alpha/2)$-größte der $\widehat{\Theta}^*$-Werte die untere und obere Grenze angeben.

Permutationstest gemäß Hamill (1999)

Eine oft verwendete Methode zur Bildung von Konfidenzintervallen für den Unterschied in der erreichten Maßzahl zweier Modelle ist ein Permutationstest gemäß Hamill (1999), welcher für 2x2 *scores* geeignet ist. Die Daten liegen in diesem Fall also in Form von Kontingenztabellen jedes Modells für jeden Zeitpunkt vor. Als Nullhypothese wird formuliert, dass sich das berechnete Gesamtmaß über alle Zeitpunkte des Modells 1 nicht statistisch signifikant von dem des Modells 2 unterscheidet, also $H_0 : \widehat{\Theta_1} - \widehat{\Theta_2} = 0$. Die Testgrößen $\widehat{\Theta_1}$ und $\widehat{\Theta_2}$ berechnen sich aus den aufsummierten Kontingenztabellen aller Zeitpunkte. Um H_0 zu überprüfen, werden zufällige Permutationen der Daten gebildet. Konkret wird für jeden Zeitpunkt zufällig ein Modell bzw. dessen Kontingenztabelle ausgewählt, die Selektierten für alle Termine aufaddiert und zur Berechnung von $\widehat{\Theta_1^*}$ herangezogen. Die Kontingenztabellen des zum jeweiligen Zeitpunkt anderen Modells werden ebenfalls summiert und daraus $\widehat{\Theta_2^*}$ errechnet. Bei einer angemessen häufigen Wiederholung des Zufallsprozesses (es wird wieder $n_{\text{perm}} = 1000$ empfohlen) erhält man eine Verteilung von n_{perm} Differenzen $\widehat{\Theta_1^*} - \widehat{\Theta_2^*}$. Wie beim Bootstrapping werden die Perzentilswerte dieser Verteilung zur Konstruktion des Konfidenzintervalls $[t_l, t_u]$ herangezogen. Formal sind die Grenzen also definiert durch

$$\begin{aligned} Pr^*\left[(\widehat{\Theta_1^*} - \widehat{\Theta_2^*}) < t_l\right] &= \frac{\alpha}{2} \qquad \text{und} \\ Pr^*[(\widehat{\Theta_1^*} - \widehat{\Theta_2^*}) < t_u] &= 1 - \frac{\alpha}{2}, \end{aligned} \tag{3.51}$$

wobei die Pr^* die aus der Verteilung berechneten Wahrscheinlichkeiten sind. Unterschiede in $\widehat{\Theta}$, die außerhalb von $[t_l, t_u]$ liegen, können als statistisch signifikant für das gewählte Konfidenzniveau α angesehen werden. Zur graphischen Darstellung kann das Konfidenzintervall der Maßzahl eines der beiden Modelle zugeschrieben und zuaddiert werden: zum Beispiel wird $\widehat{\Theta_1}$ mit einem Intervall $[\widehat{\Theta_1} + t_l, \widehat{\Theta_1} + t_u]$ dargestellt. Solange $\widehat{\Theta_2}$ außerhalb dieses Intervalls liegt, ist der Unterschied zwischen den beiden statistisch signifikant für α.

Permutationstests gehören wie das Bootstrapping zu den *Resampling* Verfahren. Im Unterschied zum Bootstrapping kann man das Vorgehen beim Permutationstest als zufälliges Ziehen „ohne Zurücklegen“ bezeichnen. Zu jedem Termin wird eine zufällige Permutation von Modell 1 und Modell 2 gemacht, für einen Zeitpunkt kann ein Modell nur entweder für $\widehat{\Theta_1^*}$ oder für $\widehat{\Theta_2^*}$ verwendet werden, nicht für beide. In diesem Sinne kann man den beschriebenen Permutationstest prinzipiell auch für den Vergleich von mehr als zwei Modellen verallgemeinern. Dafür werden zu jedem Zeitpunkt die Kontingenztabellen der m verwendeten Modelle zufällig permutiert, die $\widehat{\Theta_j^*}$ (mit $j = 1, 2, \ldots, m$) berechnet und alle möglichen Differenzen $\widehat{\Theta_1^*} - \widehat{\Theta_k^*}$ (mit $k = 2, \ldots, m$) gebildet. Das wie in (3.51) berechenbare Konfidenzintervall kann als Vertrauensbereich für einen statistisch signifikanten Unterschied zwischen $\widehat{\Theta_1}$ und den $\widehat{\Theta_j}$ angesehen werden.

3.7 Skalenunterschiede und Repräsentativität

In vorliegender Arbeit wird die Verifikation auf dem 8 km Gitter der VERA-Analysen durchgeführt und zu diesem Zweck werden alle Modelle auf dieses interpoliert. Die gewählte Auflösung entspricht ungefähr der des COSMO-7, COSMO-EU und ALADIN-FR, für das gröbste Modell, das CMC-GEM-L, verringert sich der Gitterpunktsabstand bei der Interpolation hingegen auf fast die Hälfte und für die am höchsten aufgelösten Modelle AROME, COSMO-2 und COSMO-DE wird er mehr als verdoppelt. Die Interpolation alleine wird also bei letzteren drei eine Glättung der Felder zur Folge haben und auch bei den übrigen entsteht lediglich durch die Abbildung auf ein anderes Gitter ein zusätzlicher, individueller Fehler („*representativeness error*“, Tustison et al. 2001). Dies ist jedoch unvermeidlich, ein einheitliches Verifikationsgitter unerlässlich, da ansonsten wie bei Weygandt et al. (2004) gezeigt, tendenziell am gröberen Gitter die besseren Ergebnisse erzielt werden. Man kann davon ausgehen, dass bei den drei Modellen, deren Auflösung sich nicht sehr stark von den gewählten 8 km unterscheidet, die geringsten Fehler durch die Interpolation gemacht werden.

Ist in der Literatur von Repräsentativitätsfehler oder „*representativeness error*“ die Rede, so wird meistens vom nicht unproblematischen Vergleich zwischen Punktbeobachtungen und Modellprognosen gesprochen (*point-to-area*). Der Vorhersagewert an einem Modellgitterpunkt ist nicht wie eine Punktprognose zu interpretieren, er repräsentiert vielmehr aufgrund der zu erfüllenden physikalischen Erhaltungssätze (Energie, Impuls, etc.) das Mittel über die Fläche bzw. das Volumen der ihn umgebenden Gitterbox und enthält daher nicht die subgrid-skalige Variabilität einer Punktbeobachtung. Göber et al. (2008) zeigen, dass daher nicht einmal ein perfektes Modell bei Verifikation gegen Punktbeobachtungen perfekte Werte erreichen kann. In dieser Arbeit werden zwar bereits analysierte (interpolierte) Beobachtungen mit den Modellen verglichen, dennoch gibt ein Analysepunkt von VERA im Gegensatz zu den Modellen keinen Gitterboxmittelwert wieder, sondern enthält aufgrund der Formulierung von VERA auch Variabilität, deren Skala unter der Gitterpunktsauflösung liegt. Diese kleinskaligen Strukturen sind teilweise in den Modellen mit einer Auflösung von weniger als 3 km – zwar durch die Interpolation abgeschwächt, aber dennoch – enthalten, was einen Vorteil der sehr hochauflösenden Modelle bezüglich dieses Aspekts des Repräsentativitätsfehlers vermuten lässt.

Es lässt sich also zusammenfassen, dass für die Modelle mit einer Auflösung um 8 km (COSMO-7, COSMO-EU, ALADIN-FR) ein geringer Interpolationsfehler gemacht wird, diese aber die kleinen Skalen von VERA nicht enthalten, bei den höherauflösenden Modellen (COSMO-2, COSMO-DE und AROME) hingegen ein größerer Interpolationsfehler gemacht wird, jene aber dafür die kleinräumigen Phänomene erfassen. Die beiden angesprochenen Modellgruppen werden also jeweils von einem Aspekt des Repräsentativitätsfehlers mehr beeinflusst als vom anderen. Beim gröberen CMC-GEM-L gehen beide vergleichsweise stark ein. Die Magnitude der Komponenten des Repräsentativitätsfehlers zu quantifizieren, etwa durch Verifikation auf unterschiedlichen Gittern, ist wünschenswert, geht aber über den Rahmen dieser Arbeit hinaus.

Die gewählte Verifikationssituation „Modelle gegen VERA am Analysegitter“ birgt also an sich schon Fehlerquellen. Im Vergleich zu anderen Möglichkeiten, wie z. B. die Modelle gegen modelleigene Analysen oder Beobachtungen zu verifizieren, sind jene jedoch gering oder aber werden durch andere Vorteile wettgemacht, wie z. B. die Modellunabhängigkeit und die einfache Handhabung.

4 Auswertungen

4.1 Mittlere Übereinstimmung im Tages- und Halbjahresverlauf

Als ersten Schritt sich dem Modellvergleich zu nähern ist es naheliegend, die mittlere Übereinstimmung von Modellen, Analyse und eventuell der Referenzprognose zu untersuchen.

In Abbildung 4.1 sind die gemittelten Tagesverläufe der untersuchten Parameter in VERA, MESOCLIM und den ersten 24 Stunden der Modellprognosen für den Sommer 2007 zu sehen. Durch den Unterschied zwischen Modell- und Analysemittel kann der Bias abgeschätzt werden. Die ersten sechs Stunden der Prognose sind, wie in der Abbildung angedeutet, aufgrund des Modell „*spin-up*“ nicht vertrauenswürdig. Die Muster der Tagesgänge werden von den Modellen für die Temperaturmaße und den reduzierten Bodendruck gut wiedergegeben, wenngleich die meisten Modelle in θ_e und $\mathrm{p_{msl}}$ eine stärkere mittlere Tagesschwankung aufweisen als die Analyse. Der Druck und die potentielle Temperatur werden im Mittel meist unterschätzt, wobei die einheitlichsten und oft stärksten Abweichungen beim Druck eher in der ersten Tageshälfte auftreten, bei θ eher nachmittags.

Das Auffrischen des mittleren Windes im Tagesverlauf bis zum Nachmittag, also der Unterschied $\overline{ff}_{\mathrm{Tag}} - \overline{ff}_{\mathrm{Nacht}}$, wird von den Modellen etwas schwächer (COSMOs) als in der Analyse oder kaum wiedergegeben. Die COSMO Modelle weisen überdies einen signifikanten positiven Bias der Windgeschwindigkeit über alle Tageszeiten auf. Die mittlere Windrichtung, welche bis zum frühen Abend von Südwest gegen West dreht, wird von den Modellen fast durchgehend gegenüber VERA nach rechts verschoben vorhergesagt. Die sechsstündige Niederschlagsmenge wird von den meisten Modellen eher überschätzt. Bemerkenswert gut ist die Übereinstimmung des mittleren Niederschlags in den ersten 24 bzw. 18 Stunden der Prognose der hochauflösenden COSMO Modelle mit dem der Analyse.

Die mittleren Monatsverläufe der Parameter, die in Abb. 4.2 für den 18 UTC Termin

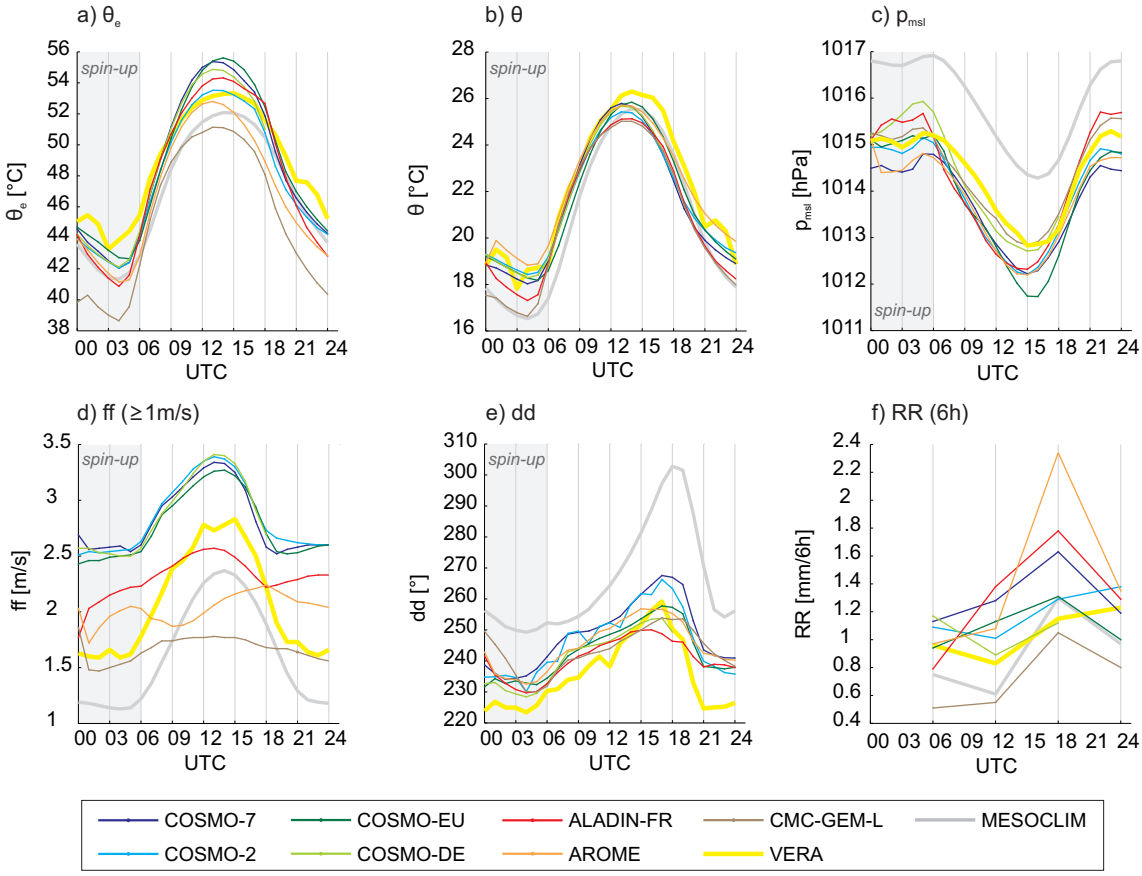

Abbildung 4.1: Mittlere Tagesgänge von a) äquivalentpotentieller, b) potentieller Temperatur, c) reduziertem Luftdruck, d) Windgeschwindigkeit, e) Windrichtung und f) sechsstündigem Niederschlag für Jun.–Aug. 2007 im limitierten D-PHASE Gebiet. Für die Modelle ist die Uhrzeit gleich der Vorhersagezeit in Stunden.

bzw. bei *RR* für 00 UTC dargestellt sind, erlauben gemeinsam mit den Tagesverläufen einen Einblick in die mittleren synoptischen Verhältnisse des Sommers und Herbsts 2007 im limitierten D-PHASE Gebiet. Für den Niederschlag ist anstatt der sechsstündigen, die Summe über 24 Stunden gezeigt, da die Tageswerte für einen jahreszeitlichen Überblick aussagekräftiger sind als nur der Nachmittagsausschnitt. Es muss jedoch dadurch auf die Einbeziehung des COSMO-DE verzichtet werden, da es nur für 18 Vorhersagestunden vorliegt. Bei allen anderen Parametern ist aufgrund der Modellverfügbarkeit der 18 UTC Termin ausgewählt. Der Sommer beginnt mit einem feucht-warmen Juni mit unterdurchschnittlichem Bodendruck und überdurchschnittlichen Windgeschwindigkeiten im Vergleich zum langjährigen Klimamittel. Eine erhöhte zyklonale Aktivität lässt sich erahnen. Ähnliche Bedingungen setzen sich im Wesentlichen im Laufe des Sommers fort, wobei die Temperaturgrößen im Juli und August eher durchschnittlich bis unterdurchschnittlich werden. Das Monatsniederschlagsmaximum wird im August erreicht und kann hauptsächlich auf Starkregenereignisse zwischen dem 6. und 9. August zurückgeführt werden, welche

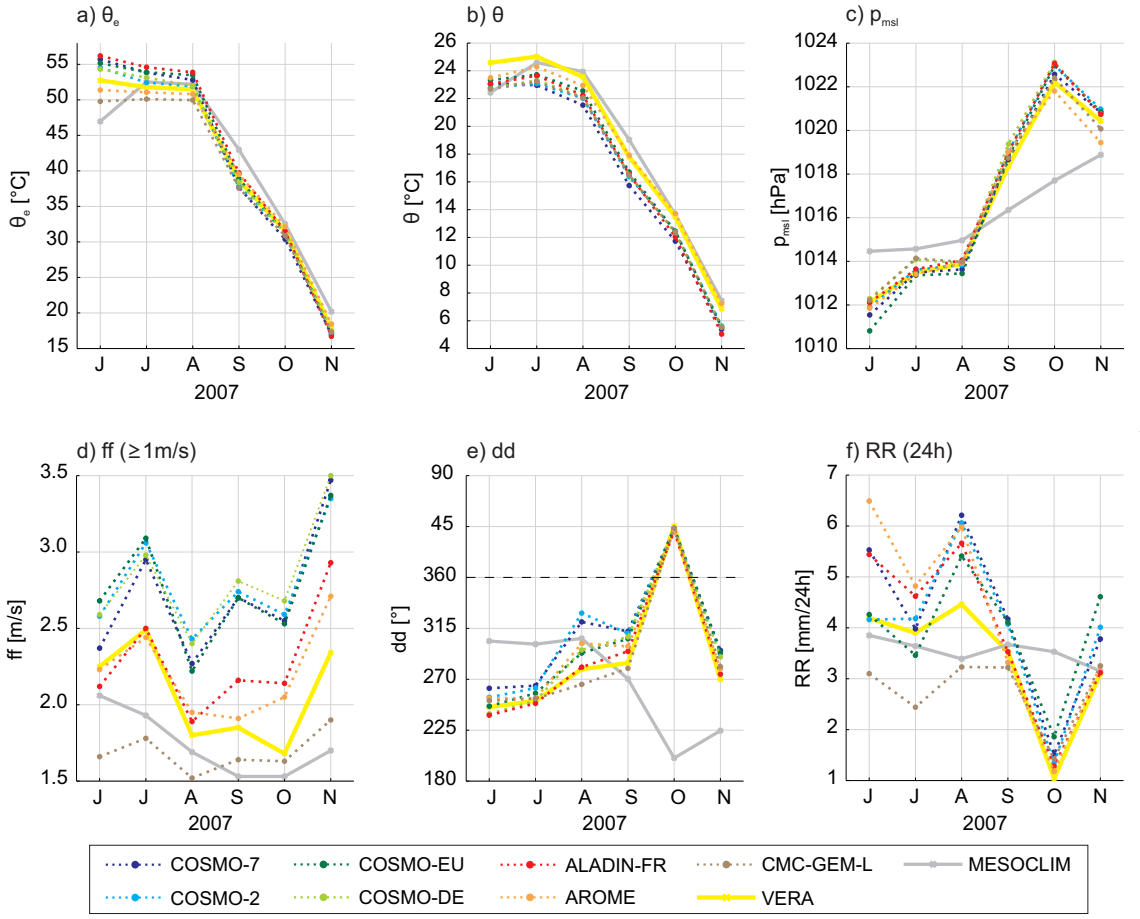

Abbildung 4.2: Mittlere Monatswerte derselben Parameter wie in Abb. 4.1, nur mit 24 h-Niederschlag (f), im limitierten D-PHASE Gebiet. Vorhersagezeit für *RR*: 24 h, für alle anderen Parameter: 18 h (00 UTC Lauf).

zu schweren Überschwemmungen in der Schweiz und in Deutschland führen. Der Herbst 2007 verläuft hingegen trockener, kühler und unter mehr Hochdruckeinfluss als im Klimamittel. Besonders der Oktober sticht als extrem trockener Monat mit einem starken p_{msl}-Maximum heraus. Auch die mittlere Windrichtung schwenkt von den typischen westlichen Richtungen im Oktober aufgrund der Lage der vorherrschenden ausgedehnten und beständigen Hochdruckgebiete auf Nordost.

Da die Modellunterschiede bei θ, θ_e und p_{msl} im Vergleich zur saisonalen Schwankung sehr gering und im Bild nur schwer erkennbar sind, werden jene in Abb. 4.3 anhand des Bias veranschaulicht. Von den verwendeten Modellen wird θ um 18 UTC durchgehend und über weite Teile des Gebiets (nicht gezeigt) unterschätzt, für θ_e treten hingegen im Sommer für die COSMOs und ALADIN-FR positive Biase auf, im Herbst tendenziell negative. Der mittlere reduzierte Bodendruck wird im Sommer noch von den gröberen COSMO Modellen geringer, im Herbst von fast allen Modellen zumeist höher als in der Analyse vorhergesagt. Sowohl in den trockene-

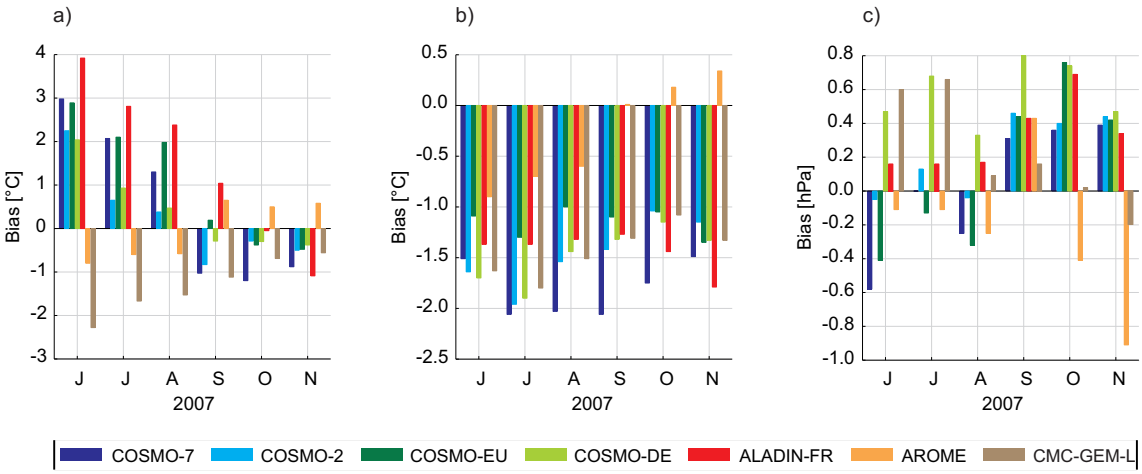

Abbildung 4.3: Monatswerte des Bias für die 18 h Vorhersagen (00 UTC Lauf) von a) θ_e, b) θ und c) p_{msl} im limitierten D-PHASE Gebiet.

ren als auch in den feuchteren Monaten überschätzt der Großteil der Modelle den Niederschlag. Die mittlere Windrichtung wird in den meisten Fällen nach rechts abgelenkt vorhergesagt, im atypischen Monat Oktober wird die mittlere Analyse von den Modellen am besten getroffen. Bezüglich der Windgeschwindigkeit ist wie im Tagesverlauf auch über alle Monate ein positiver Bias zu erkennen, die Überschätzung liegt vor allem im Flachland vor (vgl. Abb. 4.15 b für das COSMO-7).

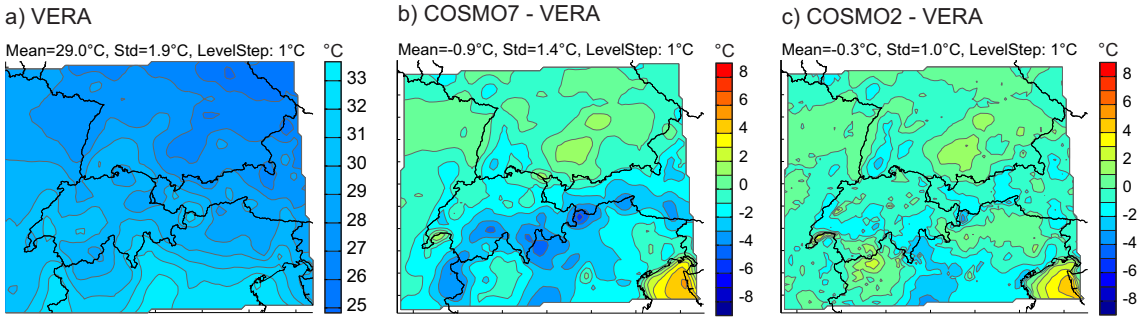

Abbildung 4.4: Mittleres VERA-Feld der äquivalentpotentiellen Temperatur (a) sowie mittlere Abweichungen COSMO 7 - VERA (b) und COSMO 2 - VERA (c) für Sep.–Okt. 2007; 18 h Vorhersagen des 00 UTC Laufs.

Karten der mittleren Felder und der mittleren Abweichungen zwischen Modell und Analyse wie in Abb. 4.4 geben ersten Aufschluss über die räumliche Verteilung der Unterschiede. Wie schon in Abb. 4.3 erkennbar, weisen COSMO-2 und COSMO-7 im Herbst einen negativen über das gesamte Gebiet berechneten Bias der äquivalentpotentiellen Temperatur auf, COSMO-7 stärker als COSMO-2. Das gröbere Modell zeigt die stärksten negativen Abweichungen entlang der Alpen (b), mit geringer werdenden θ_e-Werten in höheren Lagen (nicht explizit gezeigt), was wie auch aus den Radiosondendaten ermittelt, nicht mit den mittleren Beobachtungen im Herbst in

der Region übereinstimmt (vgl. Abb. 2.2). Dieser systematische Fehler an den Alpen ist im feineren Modell nicht bzw. nur andeutungsweise zu sehen (c). Über dem Schwarzwald zeigt VERA im Vergleich zu den Vogesen höhere äquivalentpotentielle Temperaturen (a), die von beiden Modellen unterschätzt werden. Negative mittlere Abweichungen sind des Weiteren in der Poebene zu erkennen, wo VERA stärker als beide Modelle ein θ_e-Maximum aufweist.

Auffällig sind die stark positiven Abweichungen beider Modelle über der Adria. Die Modelle zeigen dort viel deutlicher als VERA eine nach Sonnenuntergang wärmere und feuchtere Luft über dem Meer verglichen mit dem Festland. Da die Analyse in dieser Region auf nur einer (Bogus-) Station über dem Meer und ansonsten Landstationen beruht, kann VERA diese Effekte nicht ausreichend gut erfassen, ein Vergleich in dieser Region ist für die äquivalentpotentielle Temperatur nicht unbedingt sinnvoll. Lässt man für die Berechnung der monatlichen Verifikationsmaße den Adria-Ausschnitt großräumig weg, so ändern sich die Bias Werte am stärksten für COSMO-DE und COSMO-EU: die Überschätzung im Sommer wird geringer, die Unterschätzung im Winter wird stärker, aber das Grundmuster aus Abb. 4.3 bleibt erhalten (Abb. 4.5 b). Um die Fehler der Modelle abseits von dem an der Adria zu beschreiben, wird im Folgenden bei θ_e das „limitierte D-PHASE Gebiet ohne Meer“ verwendet, bei dem mit dem groben Bereich um die Adria etwa 4 % der ursprünglichen Gesamtfläche weggelassen werden.

Die höhere spezifische Wärmekapazität des Wassers gegenüber der Landfläche beeinflusst auch alle weiteren untersuchten meteorologischen Variablen: unter Idealbedingungen herrscht untertags über dem Meer eine geringere potentielle Temperatur und höherer Luftdruck als am Festland, es kommt zu Seewind, der unter Umständen zu konvektivem Niederschlag führt. Die speziellen Verhältnisse über dem Meer sind jedoch bezüglich jener Parameter im Gegensatz zu θ_e in VERA ähnlich wie in den Modellen repräsentiert, der Adria-Ausschnitt muss nicht weggenommen werden.

4.2 Jahreszeitliche Schwankungen der Gütemaße

Mit den mittleren meteorologischen Bedingungen variieren auch die Vorhersageeigenschaften, die über den systematischen Fehler hinausgehen, im Laufe der sechs betrachteten Monate. Bezüglich der zeitlichen Variationen der auf das Klima bezogenen *skill scores* muss immer Vorsicht gewahrt werden, da die variable Güte der Klimareferenzprognose eine große Rolle spielt.

Abbildung 4.5 a zeigt im Sommer einen tendenziell höheren RMSE der 18 h Vor-

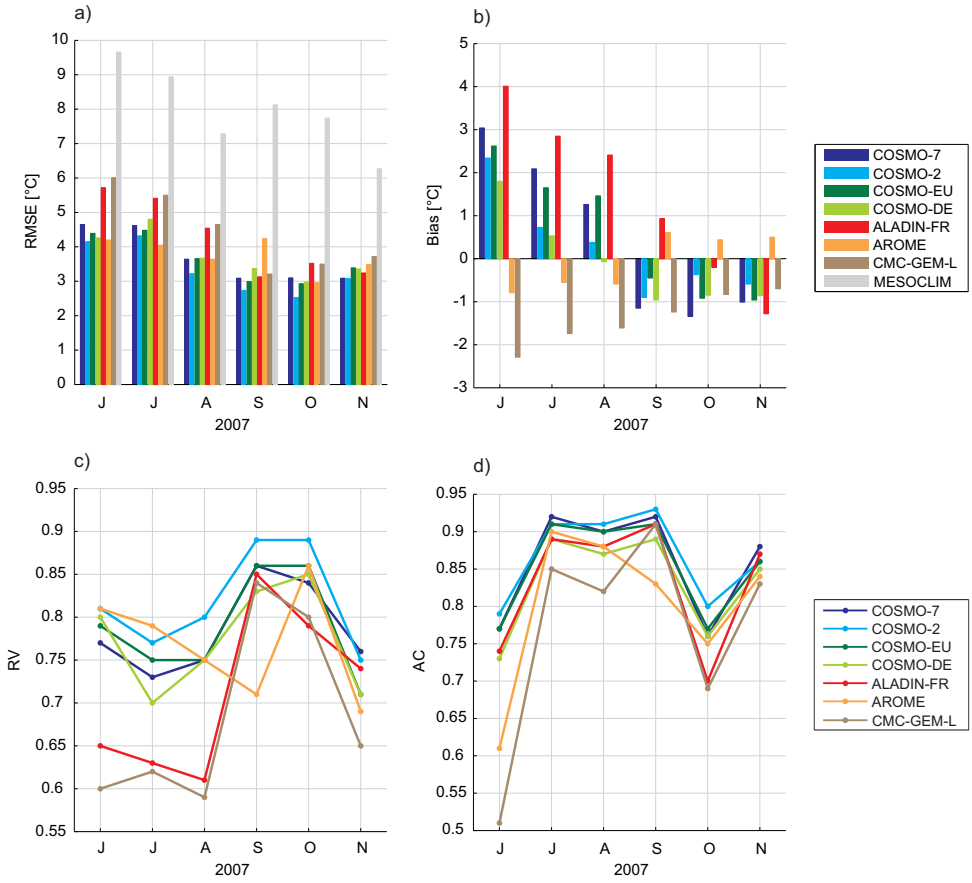

Abbildung 4.5: Monatswerte von a) RMSE, b) Bias, c) RV und d) AC für die 18 h Vorhersagen (00 UTC Lauf) der äquivalentpotentiellen Temperatur im limitierten D-PHASE Gebiet ohne Meer.

hersagen der äquivalentpotentiellen Temperatur als im Herbst und grob gesehen spiegelt sich dieses Verhalten auch in der Reduction of Variance (c) mit besseren Werten im Herbst wider. Die Ausnahme bildet der Monat November, in dem gemäß RV die Modelle schlechter abschneiden als in den anderen Herbstmonaten, was jedoch hauptsächlich auf das Sinken des Klima-RMSE (grau) zurückzuführen ist. Die klimaähnlichen Verhältnisse machen das Erreichen guter *scores* schwieriger. Die Anomaliekorrelationen (d) folgen hingegen keinem jahreszeitlichen Signal. Im Juli ist die Verbesserung des RMSE gegenüber der Klimareferenzprognose (RV) für die meisten Modelle etwas schlechter als im Juni, gleichzeitig zeigen alle Modelle eine starke Verbesserung bezüglich der AC zwischen diesen beiden Monaten, die „Muster“ in den Anomalien werden also im Juli besser erkannt, obwohl der mittlere quadratische Fehler eher steigt.

In allen Modellketten zeigen die gröberen Modelle höhere absolute Bias-Fehler (b) als die feineren, Ausnahmen sind der September für COSMO-EU und -DE und der Oktober für ALADIN/AROME. Die RV zeigt ein durchgehend besseres Abschneiden

des COSMO-2 gegenüber dem COSMO-7, während dies bezüglich der AC im Juli und November nicht der Fall ist. In der COSMO-Kette des DWD dominiert hingegen bezüglich der *skill scores* eher das gröbere Modell, wenn auch nur knapp und nicht in allen Monaten.

Die unterschiedlichen Verläufe von RV und AC machen deutlich, dass wie besprochen in realen Situationen, in denen ein nicht vernachlässigbarer Bias Fehler auftritt, oft nicht die genäherte lineare Beziehung (3.22) zwischen RV und AC gilt. Für das Beispiel des COSMO-2 im Monat August lässt sich der Zusammenhang jedoch gut verifizieren, da die Bedingungen (Bias und Mittel der Analyseanomalien nahe Null, ähnliche Varianz bei Modell und Analyse) hinreichend gut erfüllt sind: $2 \cdot \mathrm{AC} - 1 = 2 \cdot 0.87 - 1 = 0.74 \approx \mathrm{RV}$.

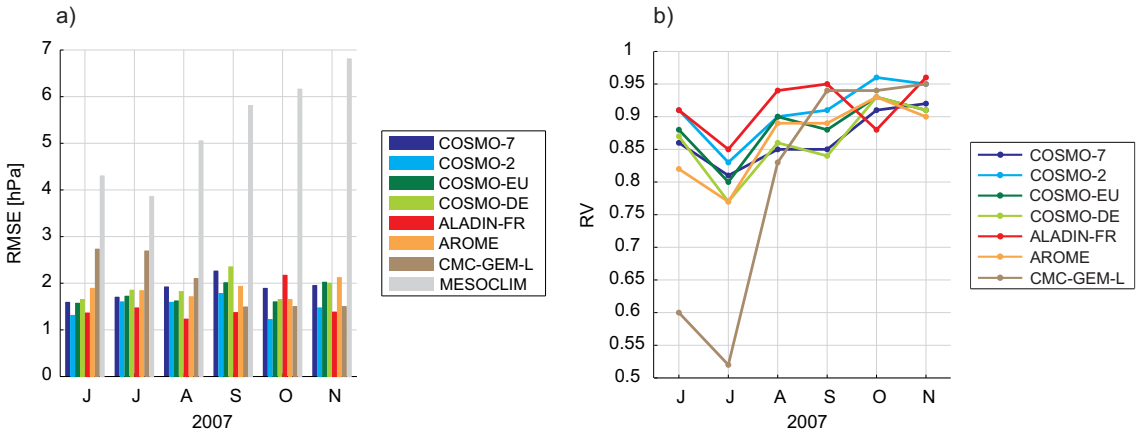

Abbildung 4.6: Monatswerte von a) RMSE und b) RV für die 18 h Vorhersagen (00 UTC Lauf) des reduzierten Luftdrucks im limitierten D-PHASE Gebiet.

Ähnlich wie bei der äquivalentpotentiellen Temperatur erreichen die Vorhersagen des auf Meeresniveau reduzierten Luftdrucks im Herbst tendenziell bessere RV-Werte als im Sommer (Abb. 4.6 b). Beim Druck kann dieses Verhalten jedoch vor allem dem steigenden RMSE der Klimareferenzprognose zugeschrieben werden (a). Analog zur äquivalentpotentiellen Temperatur gestalten sich die Modellunterschiede: COSMO-2 schneidet besser ab als COSMO-7, COSMO-EU hingegen meist besser als COSMO-DE.

Umgekehrt wie beim Luftdruck und der äquivalentpotentiellen Temperatur verhält sich der jahreszeitliche Verlauf der RV für die potentielle Temperatur mit tendenziell besseren Sommer- als Herbstwerten (Abb. 4.7 a). Zur besseren Übersichtlichkeit sind in Abb. 4.7 nur je zwei Modelle dargestellt. Zu sehen sind Ergebnisse für die 12- sowie die 24-stündigen Vorhersagen gültig jeweils um 12 UTC bzw. 00 UTC, was zusätzlich Einblick in die Tagesvariationen gibt. Um 12 UTC sind die thermischen Druckgebilde (Hitzetiefs) vor allem im Sommer stark ausgeprägt und werden

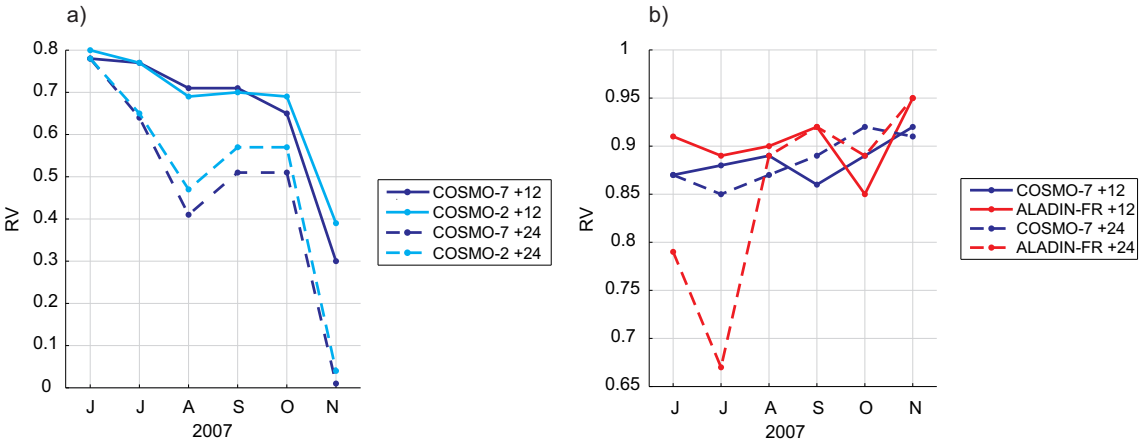

Abbildung 4.7: Monatswerte der RV für a) die potentielle Temperatur und b) den reduzierten Luftdruck, für die Vorhersagezeiten 12 h (durchgezogene Linien) und 24 h (gestrichelt) vom 00 UTC Lauf im limitierten D-PHASE Gebiet.

von den Modellen gut wiedergegeben. Um 00 UTC sind hingegen die sich bildenden (Kälte-)Hochs nicht sehr stark in θ zu sehen, da das Signal überlagert wird von der höheren potentiellen Temperatur an den Höhenlagen gegenüber der Umgebung (trockenstabile Schichtung). Die Muster in θ gestalten sich komplexer und werden somit von den Modellen ungenauer wiedergegeben. Der berechtigte Einwand, die Verschlechterung zwischen 12 und 00 UTC könnte auch nur auf die größer werdende Vorhersagezeit zurückzuführen sein, kann durch einen Vergleich widerlegt werden, der zeigt, dass auch die RV-Werte der Prognosen für 12 UTC des nächsten Tages (+36 h) noch über denen für 00 UTC bleiben. Der jahreszeitliche Verlauf ist für beide Uhrzeiten ähnlich. Die sehr geringen RV Werte im November, besonders in der Nacht, sind vor allem durch die Nähe zum klimatologischen Mittel bedingt.

Beim Luftdruck (b) ist in den gezeigten Modellen die RV im Sommer tendenziell für die 12 UTC Vorhersage besser, im Herbst für die 00 UTC Vorhersage. Das Drucksignal der nächtlichen Kältehochs ist im Herbst stärker ausgeprägt als das der mittäglichen Hitzetiefs und kann deswegen besser reproduziert werden.

Einen Vergleich der 18 h Vorhersagen für den gesamten Sommer mit jenen für den Herbst bietet das Taylordiagramm in Abb. 4.8 für mehrere meteorologische Parameter, am Beispiel der COSMO-DE Vorhersagen. Da auch die äquivalentpotentielle Temperatur enthalten ist, wird das limitierte D-PHASE Gebiet ohne Meer betrachtet. Es sind prinzipiell jene Variablen ausgewählt, die annähernde Normalverteilung aufweisen und für die daher der Pearson Korrelationskoeffizient problemlos angewendet werden kann. Zusätzlich wird die Windrichtung einbezogen, die aufgrund des zyklischen Wertebereichs im Allgemeinen keiner eigentlichen Normalverteilung folgt. Unter Einhaltung der in Abschn. 3.5 besprochenen Bedingungen an die Mittelwert-

und Differenzenbildung, durch die der Wertebereich anschaulich auf den Mittelwert zentriert wird, ist eine Korrelationsberechnung dennoch möglich aber unüblich. Es sollte die Bevorzugung anderer Verifikationsmaße in Betracht gezogen werden.

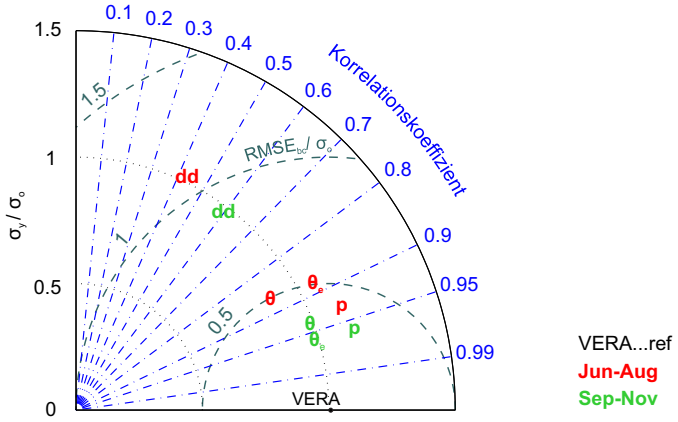

Abbildung 4.8: Taylor Diagramm für die 18 h COSMO-DE Vorhersagen (00 UTC Lauf) von θ, θ_e, p_{msl} (im Bild: p) und dd ($ff \geq 1$m/s) im Sommer und im Herbst 2007 im limitierten D-PHASE Gebiet ohne Meer.

Bezüglich der linearen Korrelation und des mit σ_o normierten RMSE_{bc} schneiden die Vorhersagen im Herbst für alle Parameter besser ab als im Sommer. Dieser Unterschied tritt genauso, und sogar verstärkt, auch in den nicht-normierten RMSE_{bc} auf (nicht gezeigt), da die Standardabweichungen der Analyse für die gewählten Variablen im Herbst immer größer sind als im Sommer.

Die eher sprunghaften Felder der Windrichtung weisen relativ schwache Korrelationen und relativ hohe mittlere Fehler auf. Im Sommer erreicht der Druck die besten Ergebnisse, potentielle und äquivalentpotentielle Temperatur erhalten kaum unterscheidbare RMSE_{bc}- und Korrelationsergebnisse. Die Feuchte, die bei θ_e gegenüber θ als ein, wie man meinen möchte, die Vorhersage möglicherweise erschwerender Faktor hinzukommt, beeinflusst die Verifikationsergebnisse nicht. Im Herbst stehen die θ_e-Vorhersagen sogar auf demselben Niveau der linearen Korrelation wie die Druckprognosen, θ liegt knapp dahinter.

4.3 Stratifizierung nach Niederschlagsintensität

Zur Evaluierung der Niederschlags- (oder Windgeschwindigkeits-) Vorhersagen über die gesamte Spannweite der Intensitäten ist es sinnvoll, nicht nur Absolutwerte als variable Schwellwerte zu verwenden, sondern auch die Quantile der Verteilungen der aufgetretenen und prognostizierten Niederschlagsmengen (vgl. Abschn. 3.4.3).

Abbildung 4.9 zeigt die Quantilswerte für Analyse und die 18-stündigen Vorhersagen

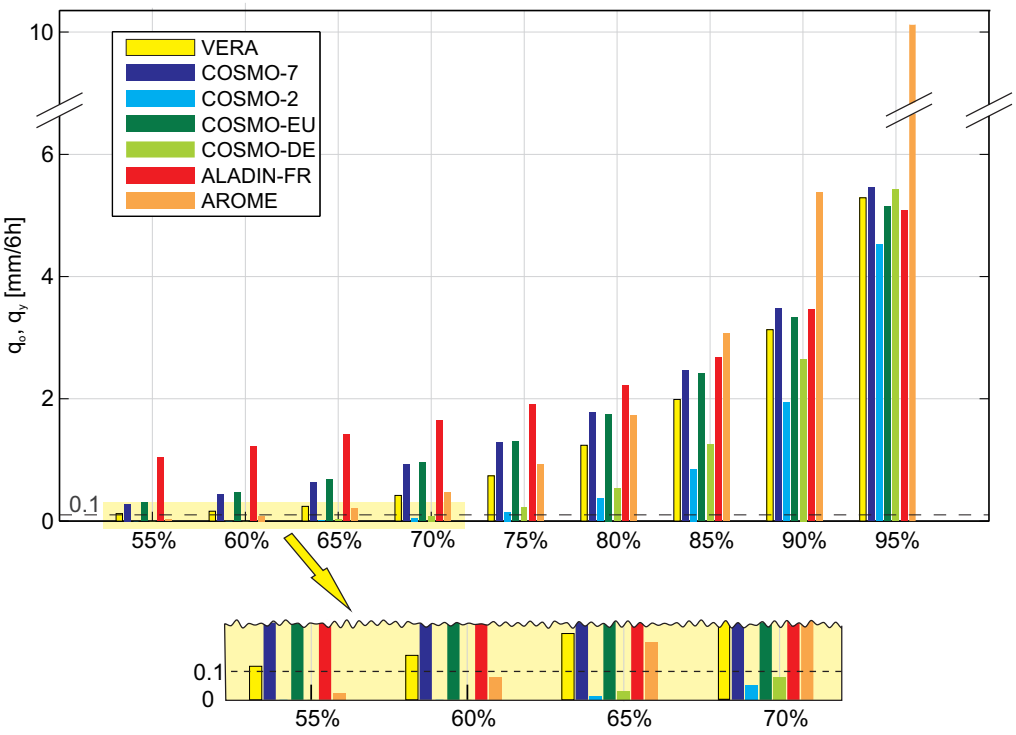

Abbildung 4.9: Quantilswerte der Analyse und der 18 h Vorhersagen (00 UTC Lauf) des Niederschlags [mm/6h] im COPS-Gebiet, Jun.–Aug. 2007. Hervorgehoben sind die Perzentile, deren zugehörige Regenmengen für ein oder mehrere Modelle noch unter 0.1 mm/6 h liegen.

aus dem 00 UTC Lauf des sechsstündigen Niederschlags im COPS-Gebiet im Sommer 2007. Erst ab dem 65 % Quantil liegt der Großteil der Modelle (vier von sechs) über dem 0.1 mm Limit für messbaren Niederschlag. Deshalb werden in den Auswertungen für dieses Beispiel nur die Quantile $\geq$ 65% gezeigt. COSMO-2 und COSMO-DE erreichen die Grenze erst bei einer Überschreitungshäufigkeit von 75 %, Ergebnisse für die darunter liegenden können nicht als sinnvoll angesehen werden. Eine Niederschlagsmenge >2 mm/6h ist an 15 % aller Zeit- und Ortspunkte der Analyse aufgetreten ($q_o(85\%) = 2$ mm). Die höchste betrachtete Quantilswahrscheinlichkeit von 95 % entspricht einem Analyse-Schwellwert von 5.3 mm/6h.

Die von den Modellen erreichten quantilsabhängigen Werte für den PSS_{qb} und die *QD'* sind in Abb. 4.10 a und c dargestellt. Wie anzunehmen, sinkt der „potentielle *skill*“ mit steigenden Quantilen, die (relativ) höheren Intensitäten bzw. ihre Überschreitung wird mit geringerem *skill* vorhergesagt als die niedrigeren.

Die *QD'* nähert sich hingegen für höhere Quantile bei allen Modellen außer AROME dem Bestwert Null. Dies bedeutet zwar nicht, dass bei höheren Niederschlagsmengen nicht größere absolute Fehler gemacht werden (die absolute QD wird teilweise für höhere Quantile maximal), aber der Bias-Fehler normiert mit dem Mittelwert aus

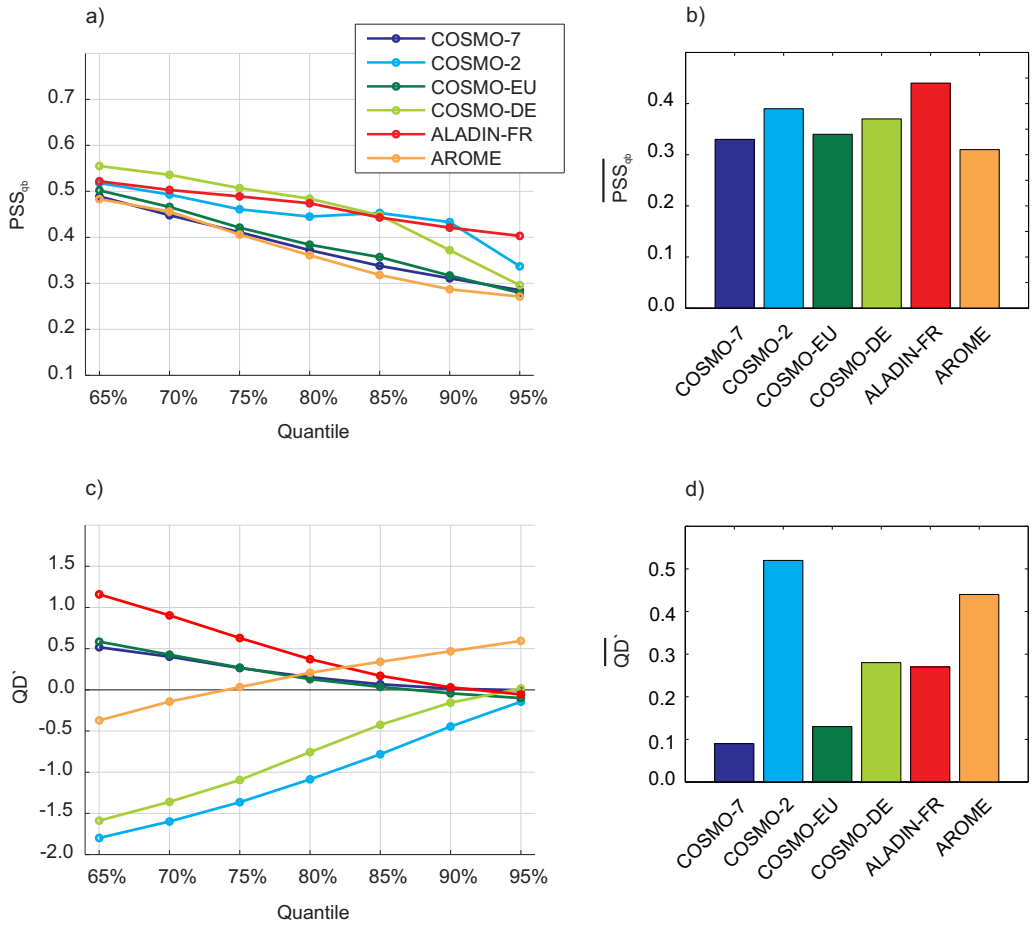

Abbildung 4.10: PSS_{qb} und normierte Quantilsdifferenz (*QD'*) für dasselbe Beispiel wie in Abb. 4.9. a) und c) zeigen PSS_{qb} und *QD'* über alle höheren Quantile, b) und d) die jeweils über alle Quantile integrierten Größen.

Modell- und Analyse-Intensitätsschwellwert sinkt mit größer werdenden Schwellen. Die feineren Modelle (außer AROME) unterschätzen die Niederschlagsmengen, die etwas gröberen Modelle überschätzen sie fast durchwegs. Über alle Quantile integriert (Teilabbildungen b und d) lässt sich erkennen, dass die höher auflösenden COSMO Modelle einen höheren „potentiellen *skill*“ als ihre gröberen Gegenstücke erreichen, jedoch gleichzeitig auch einen stärkeren Bias ($\overline{QD'}$) aufweisen. AROME schneidet bezüglich beider Maße schlechter ab als das ihm übergeordnete ALADIN-FR, schon Abb. 4.9 zeigt, dass das in Testphase befindliche AROME fast doppelt so hohe maximale (also 95% Quantils-) Niederschlagsmengen enthält wie die Analyse.

Durch diese Evaluierungen anhand der quantilsbasierten Maßzahlen kann trotz eines recht starken Bias der hochauflösenden COSMO Modelle ein Vorsprung im „potentiellen *skill*“ ihrer Niederschlagsprognosen im Vergleich zu ihren gröberen Gegenstücken erkannt werden.

Wird, wie in Abb. 4.11 b, der klassische PSS bei absoluten Schwellwerten herangezo-

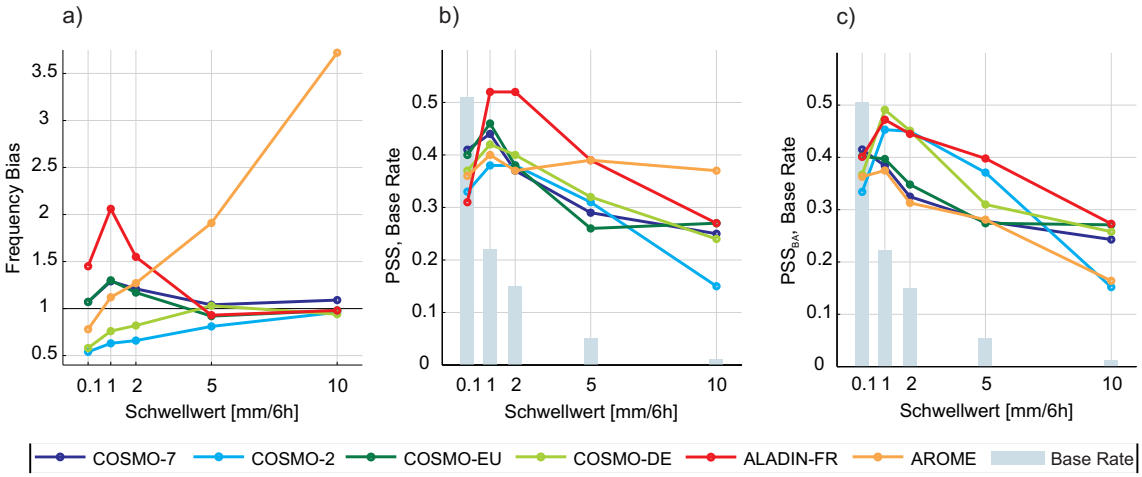

Abbildung 4.11: Frequency Bias (a), PSS (b) und Bias-adjustierter PSS (PSS_{BA}) (c) für dasselbe Beispiel wie in Abb. 4.9.

gen, so sind diese Muster nicht erkennbar. Die Reihung der Modelle ändert sich für jeden Schwellwert. Das ALADIN-FR sticht beispielsweise mit hohen PSS Werten bei den mittleren Schwellen heraus. Der *skill* steigt vom 0.1 mm zum 1 mm Schwellwert und sinkt erst dann mit steigender Intensität. Die Entscheidung, ob Niederschlag fällt oder nicht, ist also schwerer als die Entscheidung ob dieser 1–2 mm/6h übersteigt. Für den höchsten Schwellwert von 10 mm/6h ist die Base Rate so gering, dass die PSS Werte nicht mehr vertrauenswürdig sind. Der Frequency Bias (a) ist durch das Verhältnis der Auftrtittshäufigkeiten gegeben und daher nicht direkt mit der relativen Quantilsdifferenz, welche (relative) Unterschiede in der Magnitude angibt, vergleichbar. Dennoch liefern die beiden Größen ein qualitativ ähnliches Bild von tendenzieller Überschätzung (der Häufigkeit bzw. der Magnitude) durch die gröberen Modelle und Unterschätzung durch die hochauflösenden COSMOs. Auch das Annähern der QD' an den Bestwert mit steigenden Schwellen ist im Frequency Bias wiederzufinden.

Wird eine Bias-Adjustierung bei fixen Schwellen durchgeführt (c), so erhält man erwartungsgemäß ein ähnliches Bild wie beim quantilsbasierten PSS. Zumindest für die mittleren Schwellwerte 1–5 mm/6h wird wieder die Überlegenheit der höher auflösenden COSMOs über die gröberen deutlich. Für die kleinste Schwelle von 0.1 mm/h ist festzuhalten, dass bei Bias-Adjustierung der feineren Modelle ihr Vorhersageschwellwert auf unter 0.1 mm/6h reduziert wird (vgl. Abb. 4.9), was praktisch keinen Sinn macht. Für die feinen Modelle ist daher zusätzlich zur größten auch die kleinste Schwelle wenig aussagekräftig. Der an fixen Schwellwerten Bias-adjustierte PSS und der quantilsbasierte PSS bringen also ein qualitativ übereinstimmendes Ergebnis, was hier als eine gegenseitige Kontrolle der Methoden angeführt werden

kann. Der Vorteil der quantilsbasierten *scores* ist die Möglichkeit der Berechnung von integralen Gesamtmaßen und macht jene Vorgehensweise zur attraktiveren.

4.4 Windrichtung

Die Häufigkeitsverteilungen der Windrichtung im Sommer 2007 für die Modelle, VERA und die Klimamittel sind in Abb. 4.12 anhand von Windrosen gezeigt. Die dominierenden Windrichtungen in der Analyse sind, wie es typisch für die mittleren Breiten ist, klar die westlichen. Genauer betrachtet ist die häufigste Windrichtung in VERA eher Westsüdwest, während der klimatologische Mittelwind im betrachteten Gebiet großteils aus Nordwesten kommt. Die Modelle geben die beobachtete

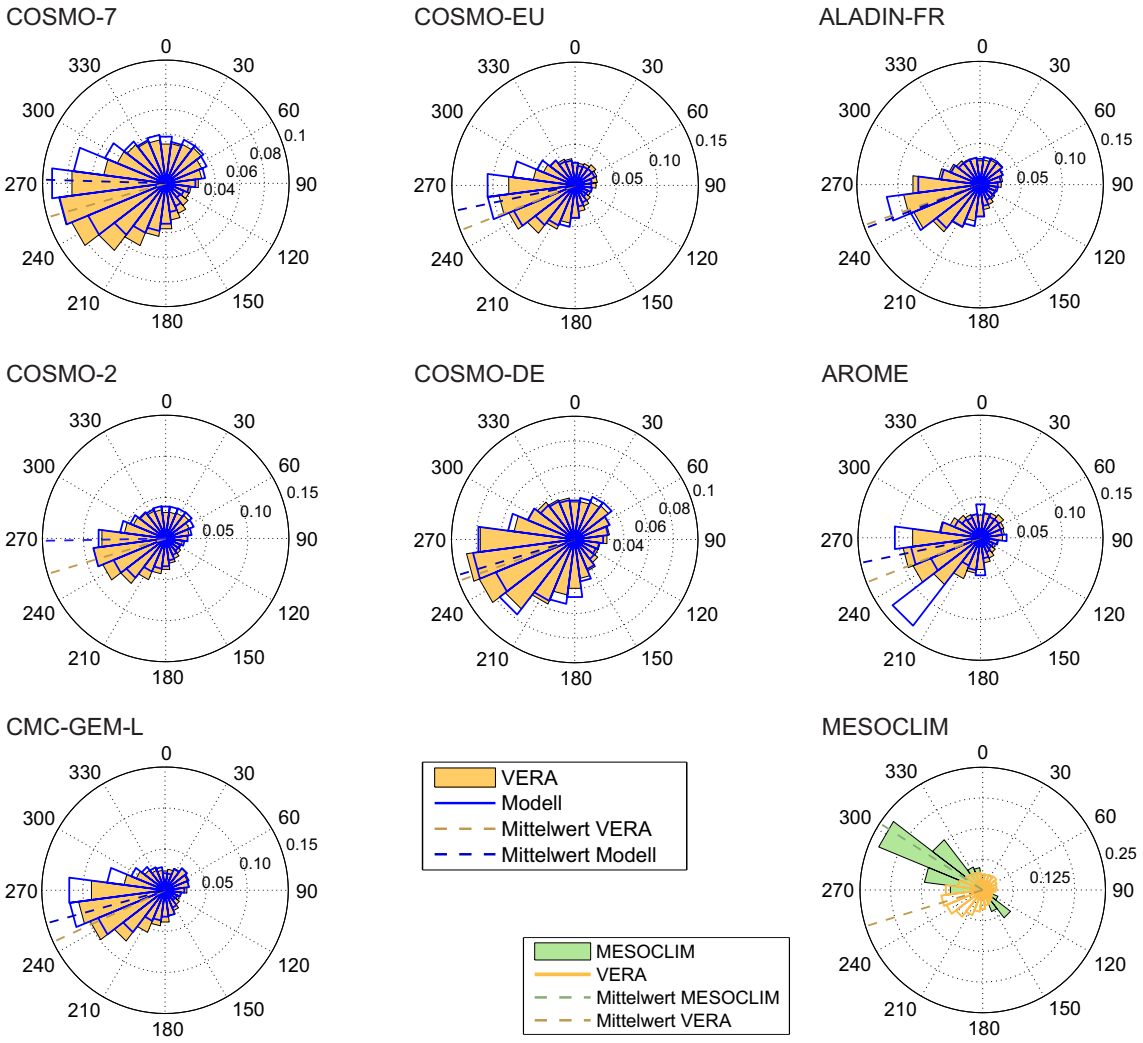

Abbildung 4.12: Häufigkeitsverteilungen der vorhergesagten, analysierten und mittleren klimatologischen Windrichtungen (bei $ff \geq 1\text{m/s}$) im Sommer 2007 im limitierten D-PHASE Gebiet.

Verteilung recht gut wieder, COSMO-7 ist um das Maximum herum ziemlich genau um eine Kategorie gegen die Analyse nach rechts verschoben, auch COSMO-EU und CMC-GEM-L enthalten mehr nordwestliche Richtungen als beobachtet. Beim COSMO-2 wird die Häufigkeit der südwestlichen Richtungen zugunsten der nordöstlichen unterschätzt, auch COSMO-DE überschätzt die nordöstlichen eher. ALADIN-FR überschätzt lediglich das Maximum, die Verteilung in AROME hingegen ergibt ein unstetiges, weniger realistisches Muster mit einem sehr starken Maximum bei Südwest und Überschätzungen genau in Nord, Ost, Süd, und West.

Zur Evaluierung der *dd*-Vorhersagen mit den 2x2-Maßen wie in Abb. 4.13 wird die etwas gröbere Unterteilung in die acht Hauptwindrichtungen getroffen. Auch hier ist die Häufigkeitsverteilung der Analysen (Base Rate) eingezeichnet und man kann, wie schon in Abb. 4.2, erkennen, dass im Herbst (b) im Vergleich zum Sommer (a) zunehmend Winde aus Nordosten auftreten und sich die Häufigkeiten der Westan die der Nordostwinde angleichen. Der SEDS für alle Modelle zeigt sowohl im Sommer als auch im Herbst ein Maximum in der NO-Richtung, im Herbst erzielen zudem einige Modelle für West und Südwest hohe Werte (> 0.3), im Sommer für Ost. Beim Vergleich der Modelle fällt auf, dass ALADIN-FR über viele Richtungen am besten abschneidet, AROME, wie schon anhand der Häufigkeitsverteilung erahnt, im Sommer oft am schlechtesten. Die SEDS-Werte der verschiedenen COSMO Modelle weichen nur sehr wenig voneinander ab, im Sommer schneiden die gröberen COSMOs für die nördlichen Richtungen (NW bis O) etwas besser ab als die höher aufgelösten, für die südlichen (SO bis W) ist es umgekehrt. Diese Unterschiede in der Modellperformance zeigen sich sehr ähnlich bei Verwendung anderer 2x2 *skill scores*, wie z. B. HSS.

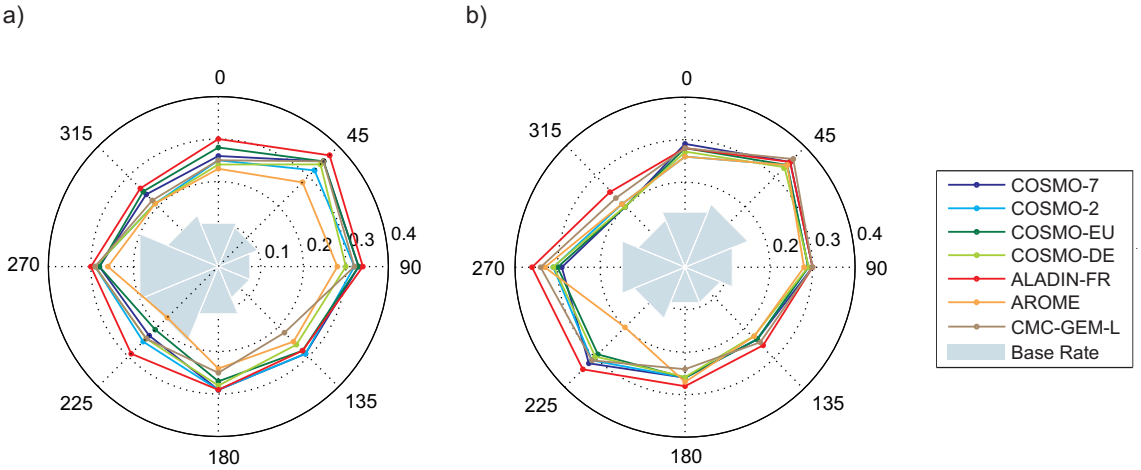

Abbildung 4.13: SEDS für die 18 h Modellprognosen (00 UTC Lauf) der acht Hauptwindrichtungen im limitierten D-PHASE Gebiet, a) Sommer, b) Herbst 2007. Blaue Balken: Base Rate.

Wie in Abschn. 3.5 beschrieben, können unter Beachtung gewisser Vorgaben zur Richtung der Abweichungen, etc. auch Mittelwerte und darauf basierende kontinuierliche Maße für die Windrichtung berechnet werden. Die mittleren Fehler von *dd* können für den gesamten Sommer schon aus Abb. 4.12 aus den Abständen der Mittelwerte abgeschätzt werden, die monatlichen Werte sind in Abb. 4.14 a aufgetragen. Die höchsten Biase treten im August, September sowie im November auf, wenn der analysierte Mittelwind ziemlich genau aus Westen kommt (vgl. Ab. 4.2). Der Oktober ist der einzige Monat, in dem alle Modelle die mittlere Windrichtung nach links verschoben vorhersagen, für den Rest des Halbjahres dominiert eine Rechtsablenkung. Ein Grund für die sehr starke Rechtsablenkung bei den (Schweizer) COSMO Modellen, die wie in Abb. 4.15 für das COSMO-7 ersichtlich, systematisch im Flachland ohne störenden Gebirgseinfluss zu finden ist, könnte ein eventuell zu schwacher Reibungsterm in der Grenzschichtparametrisierung der Modelle sein. Eine zu schwach angesetzte Reibung würde nicht nur die weniger ausgeprägte Winddrehung gegenüber den Isobaren zum tieferen Druck hin, sondern auch die im Vergleich zu VERA zu hohen Windgeschwindigkeiten (vgl. Abb. 4.1 und 4.2) in den COSMO Modellen erklären. Außerdem hat die Abweichung der Höhe der untersten Modellschicht von der Höhe, in der der Wind gemessen wird (10 m über Grund), Einfluss auf die Verschiebung des Windes.

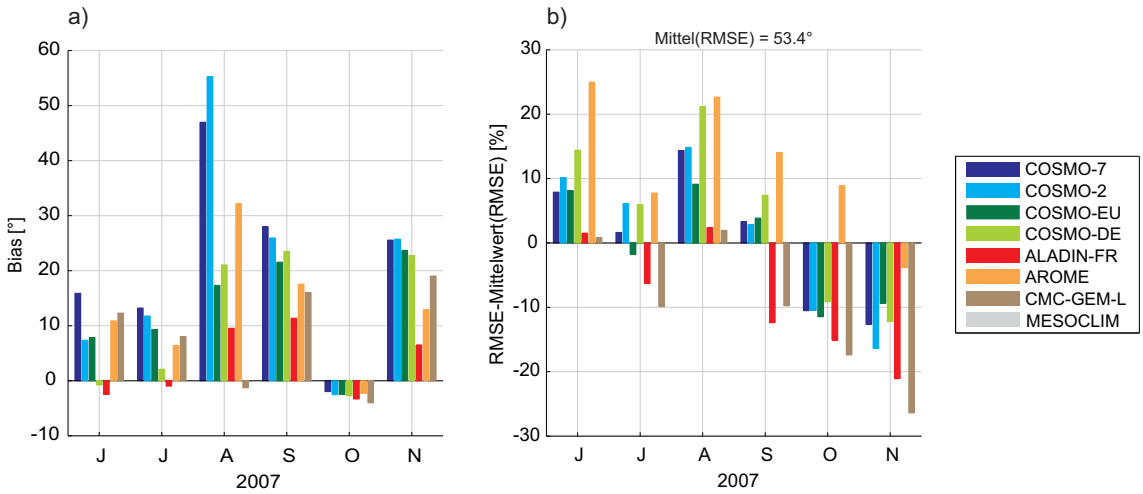

Abbildung 4.14: Monatswerte des Bias (a) und der Abweichungen RMSE - Mittelwert(RMSE) (b) für die 18 h Vorhersagen (00 UTC Lauf) der Windrichtung ($ff \geq 1$m/s) im limitierten D-PHASE Gebiet.

Wie besprochen herrscht im Oktober im Unterschied zu den anderen Monaten und zum Klima nordöstlicher Mittelwind vor, welcher dennoch von allen Modellen gut wiedergegeben wird. Sehr hohe Biase weisen COSMO-2 und COSMO-7 im August auf, wo ihr mittlerer Fehler die Größenordnung der „Standardabweichung des Fehlers“, also des RMSE, erreicht.

Die RMSEs sind in Teilabbildung 4.14 b nicht in ihren Absolutwerten, sondern in ihren Abweichungen vom Mittel aller im Bild enthaltenen, zu vergleichenden RMSEs angegeben, stark negative Werte kennzeichnen also die besten Ergebnisse. Diese Darstellungsweise wird im Folgenden auch für andere Maßzahlen gewählt, wenn Modell- oder Terminunterschiede besser sichtbar gemacht werden sollen. Der RMSE der Windrichtung erreicht die besten Werte des Halbjahres im Oktober und November, zwei Monate, die bezüglich mittlerer Windrichtung und mittlerer Windgeschwindigkeit sehr unterschiedlich sind (eher schwach aus NO bzw. eher stark aus W). Im Modellvergleich schneiden ALADIN-FR und CMC-GEM-L bezüglich des RMSE am besten ab, auch im Bias erreicht vor allem das ALADIN-FR meist die geringsten Werte. Die gröbere Auflösung ist für jene kein Nachteil bei den Windrichtungsvorhersagen. In den COSMO-Ketten ist jedoch keine eindeutige Tendenz im Vergleich zwischen den unterschiedlich feinen Modellen zu erkennen.

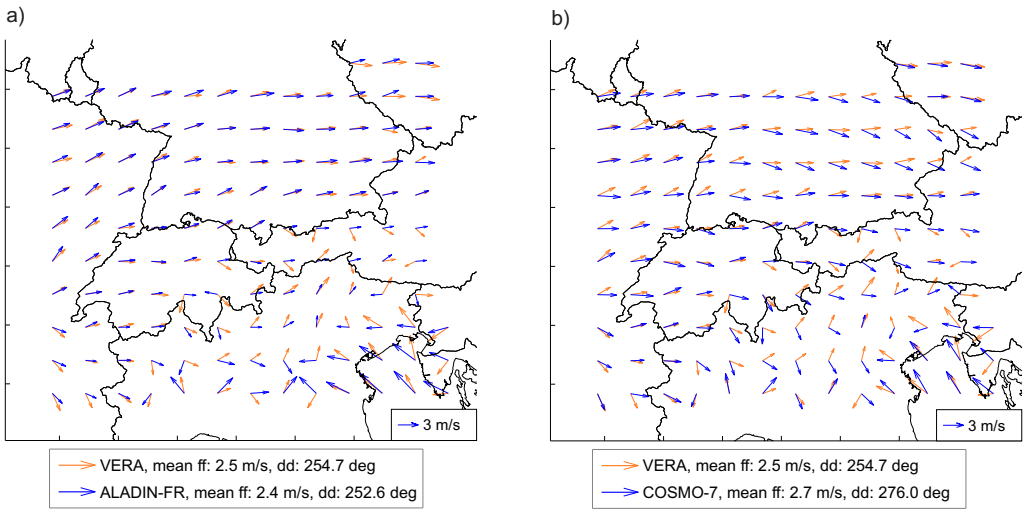

Abbildung 4.15: Mittlere Windvektoren im Sommer 2007, VERA vs. Modell a) ALADIN-FR b), COSMO-7; 18 h Vorhersagen des 00 UTC Laufs. Gezeichnet sind Mittel über je 7x7 Gitterpunkte.

Abbildung 4.15 zeigt die Felder der mittleren analysierten und vorhergesagten Windvektoren zweier Modelle, von denen das ALADIN-FR bezüglich Windrichtung und -geschwindigkeit bessere Resultate erreicht als das COSMO-7 (vgl. vorhergehende Abbildungen). Die Richtungsverteilungen aus Abb. 4.12 werden hier also nicht mehr (bzw. nur wenig) räumlich zusammengefasst und mit den mittleren Windstärken versehen. Man kann erkennen, dass in beiden Modellen die Windrichtungsvorhersagen in der nördlichen, weniger gebirgigen Hälfte des Gebiets besser mit VERA übereinstimmen als in den Alpen. Dennoch liegt im Norden ein systematischer Modellunterschied mit einer leichten Rechtsablenkung im COSMO-7 über ein großes Gebiet hinweg vor. Die größten Abweichungen treten jedoch südlich des Alpenhauptkamms

und in der Poebene auf, wo VERA eine stärkere südwestliche Komponente enthält, der Wind also nach der Umströmung wieder näher an die Alpen herangeführt wird als in den Modellen, wo verbreitet auch mittlerer Nordostwind herrscht. Von der um 18 UTC im Vergleich zur Landfläche noch kühleren Adria strömt Seewind Richtung Festland. Die mittleren Geschwindigkeiten, welche in dieser Darstellung nur diejenigen $\geq$1 m/s enthalten, stimmen insgesamt in beiden Modellen gut mit VERA überein.

4.5 Windgeschwindigkeit

Ähnlich wie der Niederschlag kann die Windgeschwindigkeit über die Spanne der Intensitäten mithilfe der quantilsbasierten Maße und der klassischen 2x2 *skill scores* untersucht werden. Die Auswertung der *ff*-Vorhersagen für den Herbst 2007 im limitierten D-PHASE Gebiet anhand der quantilsbasierten *scores* (Abb. 4.16) zeigt sehr geringe Modellunterschiede von nur wenigen Prozent im integralen *skill* (b). Ein hauchdünner Vorsprung der gröberen Modelle ist erkennbar, das ALADIN-FR erreicht den höchsten Wert.

Bemerkenswert ist das Ansteigen des PSS bis zu relativ hohen Quantilen von ca. 75%–80%, was beobachteten Geschwindigkeiten im Bereich von 2–3 m/s entspricht. Die korrekte Vorhersage der häufigen, sehr niedrigen Windgeschwindigkeiten gestaltet sich also schwierig für die Modelle. Stärkere Unterschiede als im *skill* sind in der Quantilsdifferenz zu finden: ALADIN-FR und AROME schneiden im integralen Maß (d) signifikant besser ab als die anderen Modelle. AROME weist zwar in den niedrigen Quantilen eine sehr starke Unterschätzung auf, durch die gute Performance in den höheren Quantilen und deren stärkere Gewichtung erhält es dennoch den besten Wert für $\overline{QD'}$. Alle übrigen Modelle überschätzen die Windgeschwindigkeit durchwegs.

Für die Überschreitung des Schwellwerts von 3 m/s, also die Vorhersage des Auftretens von mäßigen bis stärkeren Windgeschwindigkeiten, für welche in etwa der höchste „potentielle *skill*“ erreicht wurde, soll weiters der Monatsverlauf des Frequency Bias (a) sowie des HSS (b) in Abb. 4.17 genauer betrachtet werden. Vor allem in den Monaten niedriger Base Rate (b, Balken) überschätzen die COSMO Modelle, für die aufgrund der sehr starken Ähnlichkeit untereinander stellvertretend nur das COSMO-EU gezeigt ist, die Auftrittshäufigkeit jener Windgeschwindigkeiten stark. Diese Modelle enthalten also viel weniger jahreszeitliche Variation in der vorhergesagten Häufigkeit als die Analyse. Besser und über die Monate hin konstan-

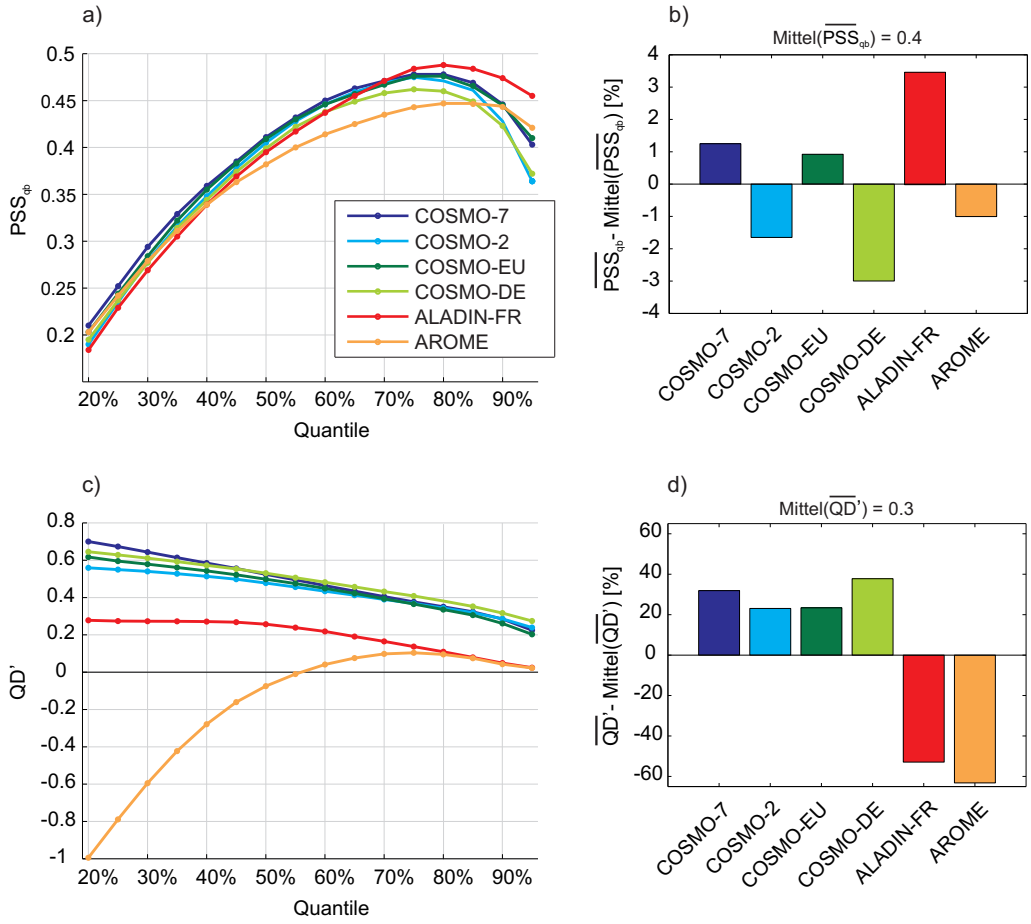

Abbildung 4.16: PSS_{qb} und QD' für die 18 h Vorhersagen (00 UTC Lauf) der Windgeschwindigkeit im Herbst im limitierten D-PHASE Gebiet. a) und c) zeigen PSS_{qb} und QD' über alle höheren Quantile, b) und d) die integralen Größen in ihren Abweichungen vom Mittel.

ter sind die Frequency Bias Werte für ALADIN-FR und AROME. CMC-GEM-L ist das einzige Modell, das die Häufigkeiten durchgehend unterschätzt. Sowohl bezüglich des Skills (HSS) als auch im Frequency Bias erreicht wieder das ALADIN-FR über fast alle Monate die besten Werte.

In den Monaten August und Oktober ist die Häufigkeit von Windgeschwindigkeiten $\geq$ 3m/s ähnlich gering, der Mittelwind ähnlich schwach (vgl. Abb. 4.2). Die Vorhersagen bei gegebenem Schwellwert weisen aber für den Oktober einen signifikant höheren *skill* auf als für den August, wo für alle Modelle das Halbjahresminimum des HSS erreicht wird. In den vorwiegenden Nordostlagen des Oktobers werden also die wenigen über schwachen Wind hinaus gehenden Geschwindigkeiten mit höherem *skill* vorhergesagt als in den eher auf West konzentrierten Regimes ähnlich windschwacher Monate.

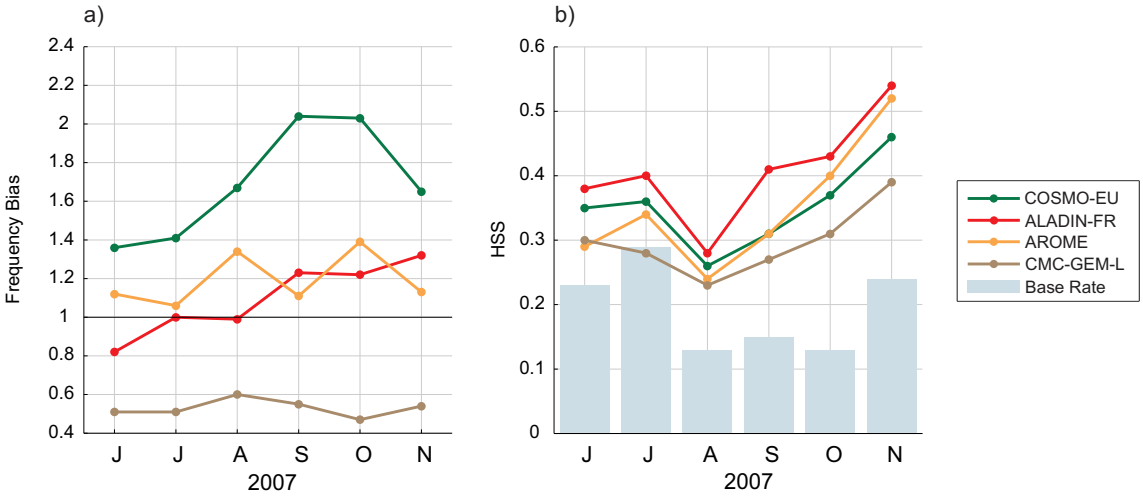

Abbildung 4.17: Halbjahresverlauf des a) Frequency Bias und b) HSS für die 18 h Vorhersagen (00 UTC Lauf) von $ff \geq 3\,\mathrm{m/s}$ im limitierten D-PHASE Gebiet.

4.6 Einfluss der Referenzprognose auf die skill scores

Die Wahl der Referenzprognose beeinflusst maßgebend die Eigenschaften und das Verhalten der *skill scores*. Für die 2x2-*skill scores* ist im klassischen Sinne nur eine Zufallsprognose als Referenz vorgesehen, es wird hier jedoch kurz auch auf die Möglichkeit einer Klima-Referenzprognose aus Abschn. 3.4.4 eingegangen. Bei den kontinuierlichen Gütemaßen wird das Verhalten von Persistenz- und Klimareferenzprognose gegenüber gestellt.

4.6.1 Zufall vs. Klima als Referenz für 2x2 skill scores

Abbildung 4.18 gibt Aufschluss über das Zustandekommen des klassischen HSS (b) aus dem PC und dem PC einer Zufallsprognose (a) für Niederschlagsmengen $\geq 0.1\,\mathrm{mm}$. Als erster Gesichtspunkt wird durch die Streuung der $\mathrm{PC_{rand}}$ ersichtlich, dass die verwendete Zufallsprognose, wie in 3.2 diskutiert, nicht beliebig, sondern von der Modellprognose abhängig ist. Der Percent Correct für die Modelle erreicht sein Maximum im sehr trockenen Monat Oktober, der hohe Wert ist also vor allem auf die große Anzahl der *correct rejections* zurückzuführen. Auch durch Zufall kann unter diesen Bedingungen ein hoher PC erreicht werden, $\mathrm{PC_{rand}}$ kommt im selben Monat zu seinem Höchstwert, was im resultierenden HSS zu einem Ausglätten der Spitze im Oktober führt. Über die Sommermonate hingegen ist die Wahrscheinlichkeit, durch eine Zufallsprognose einen guten PC zu erreichen, für viele Modelle fast konstant, was durch vergleichbare aufgetretene Niederschlagshäufigkeiten bedingt ist. Der Verlauf des PC der Modelle mit einem Maximum im Juli (außer beim

ALADIN) spiegelt sich daher im HSS wider.

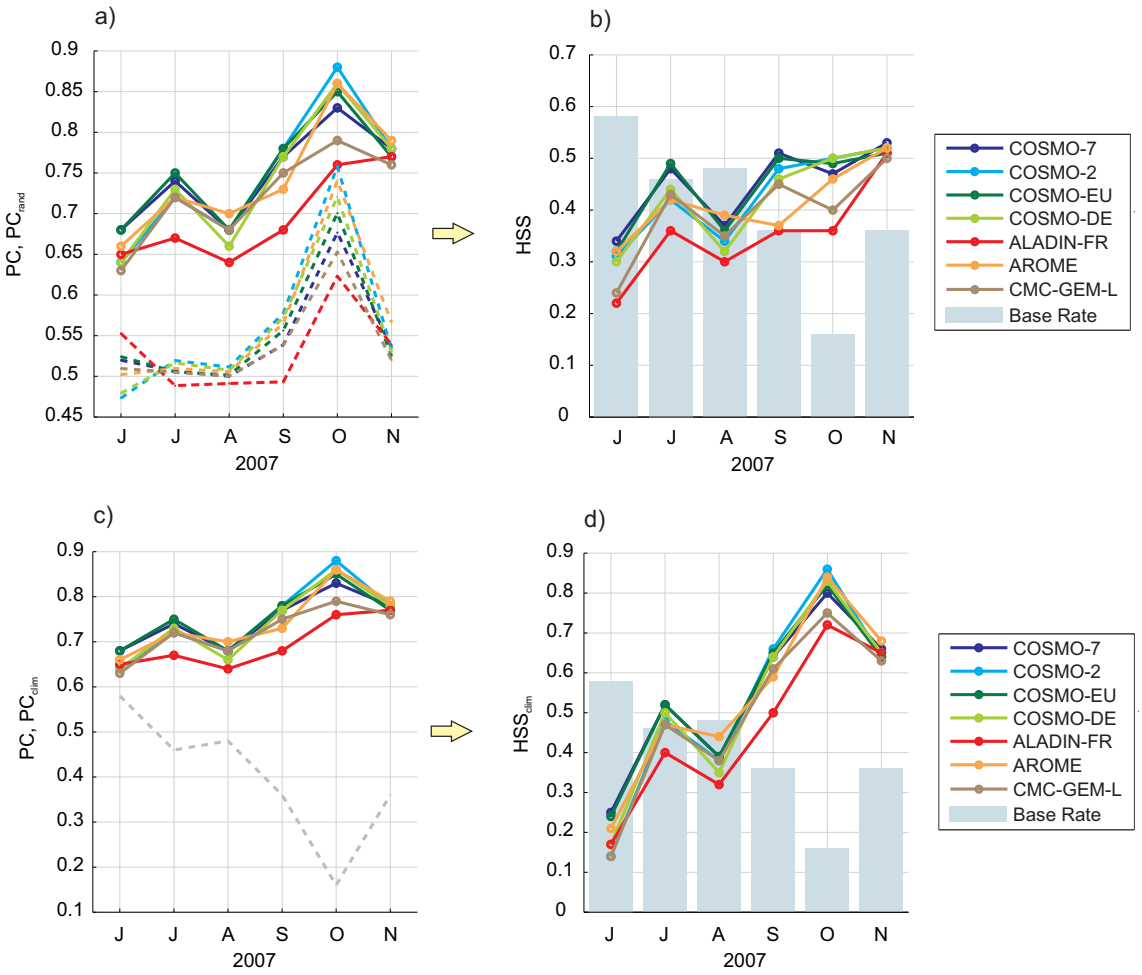

Abbildung 4.18: PC der Modelle vs. a) PC der Zufallsprognose (gestrichelt) und c) PC der Klimareferenzprognose (gestrichelt) sowie die resultierenden HSS (b) und $\mathrm{HSS_{clim}}$ (d) für die 18 h Vorhersagen (00 UTC Lauf) von $RR(6\,\mathrm{h}) \geq 0.1\,\mathrm{mm}$ im limitierten D-PHASE Gebiet. Zu beachten ist die ungleiche Achsenskalierung.

Der $\mathrm{HSS_{clim}}$ (d) zeigt eine zeitliche Entwicklung, die sich signifikant von der des HSS unterscheidet. Im Oktober tritt mit einem starken Maximum im $\mathrm{HSS_{clim}}$ die größte Abweichung auf. Die Erklärung dafür ist in der Klimareferenzprognose zu finden, welche wegen der ungewöhnlichen Trockenheit im Oktober schlecht abschneidet. Das Oktober-Maximum im $\mathrm{HSS_{clim}}$ kommt dadurch zustande, dass die Verhältnisse stark vom Klimamittel abweichen und die Modelle sehr oft korrekterweise keinen Niederschlag vorhersagen. Der $\mathrm{HSS_{clim}}$ und der $\mathrm{ETS_{clim}}$ sind für klimatologisch „extreme“ Situationen also keine sehr angemessenen Indikatoren der Prognoseleistung, da alleine die Schwäche der Klimareferenzprognose auch bei relativ schwacher Modellvorhersage hohe *scores* verursacht. Beim $\mathrm{HSS_{clim}}$ kommt die starke Gewichtung der *correct rejections* hinzu, welche bei seltenen Ereignissen problematisch ist.

4.6.2 Klima vs. Persistenz als Referenz für kontinuierliche Prädiktanden

Aufgrund der typischen mittleren Zeitskala meteorologischer Systeme erreichen Persistenzprognosen für Druck und Temperaturmaße im Kurzfristbereich oft sehr gute Ergebnisse. Ein Indikator dafür ist der Autokorrelationskoeffizient der Analyse über mehrere Zeitverzögerungen im Vergleich mit dem (Pearson-) Korrelationskoeffizienten zwischen VERA und Klima bzw. VERA und Modell.

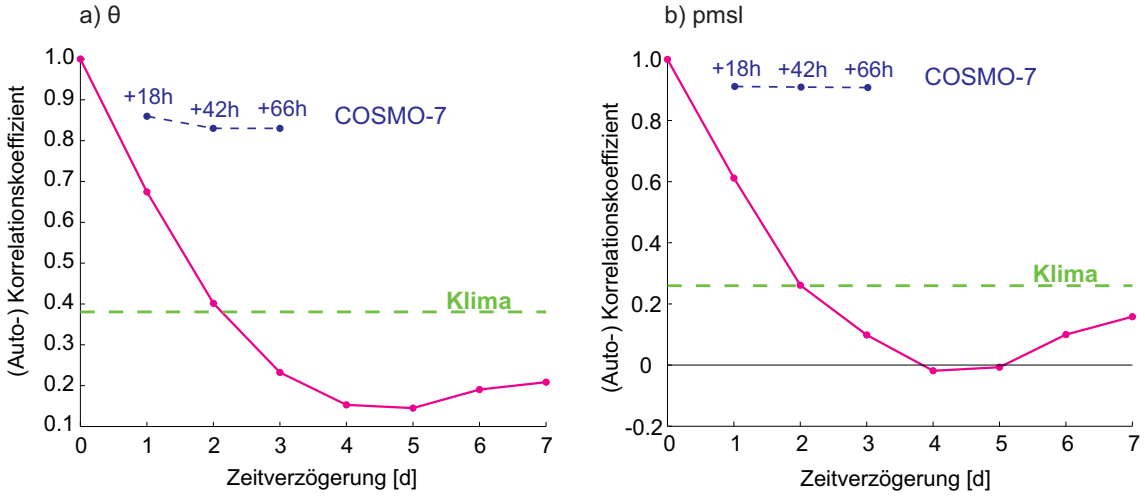

Abbildung 4.19: Autokorrelationskoeffizient der VERA Analysen mit größer werdender Zeitverzögerung (magenta) für a) θ und b) p_{msl}, Jun.–Aug. 2007, 18 UTC, limitiertes D-PHASE Gebiet. Zusätzlich eingezeichnet sind die Korrelationskoeffizienten der Analysen mit den 18, 42 bzw. 66 h COSMO-7 Vorhersagen (blau) und mit den MESOCLIM Klimamitteln (grün). Für letztere existiert keine Zeitverzögerung.

Abbildung 4.19 a zeigt, dass für die potentielle Temperatur im Sommer 2007 um 18 UTC eine Persistenzprognose in den ersten beiden Vorhersagetagen besser mit der Analyse korreliert als die Klimareferenzprognose. Ab einer Vorhersagedauer von drei Tagen ist die Korrelation mit der Klimareferenzprognose höher. Das Minimum erreicht die Autokorrelation bei einer Zeitverzögerung von fünf Tagen, danach steigt sie wieder leicht an, jedoch auf einem zu niedrigen Niveau, als dass man auf eine Wiederkehr von typischen Strukturen in θ schließen könnte. Für den Druck (b) zeigt sich ein sehr ähnliches Verhalten, im Beispiel wird die Korrelation mit Klima und Persistenzprognose am zweiten Vorhersagetag genau gleich. Die Korrelation mit der Modellprognose ist bei beiden Parametern sehr hoch und sinkt über die ersten drei Vorhersagetage für den Dreimonatszeitraum nur sehr schwach. Erwartungsgemäß schneidet das Modell besser als jede Persistenz- oder Klimareferenzprognose ab, außer für den künstlichen Fall ohne Zeitverzögerung, der definitionsgemäß eine perfekte Autokorrelation ergibt.

In der Praxis wird für eine Persistenzprognose üblicherweise die kürzestmögliche Zeitverzögerung gewählt. Wenn man annimmt, dass die Herausgabe der Persistenzprognose ebenso wie die Initialisierung des Modells um 00 UTC geschieht, so beträgt für die ersten 24 Stunden der Prognose die Verzögerung der Persistenzprognose also einen Tag, für die 25–48 h Prognosen zwei Tage, usw. In Abb. 4.20 ist der Einfluss der unterschiedlichen Referenzprognosen auf die Reduction of Variance über die gesamte dreitägige Prognosedauer des COSMO-7 erkennbar. Der erste Teil der Ab-

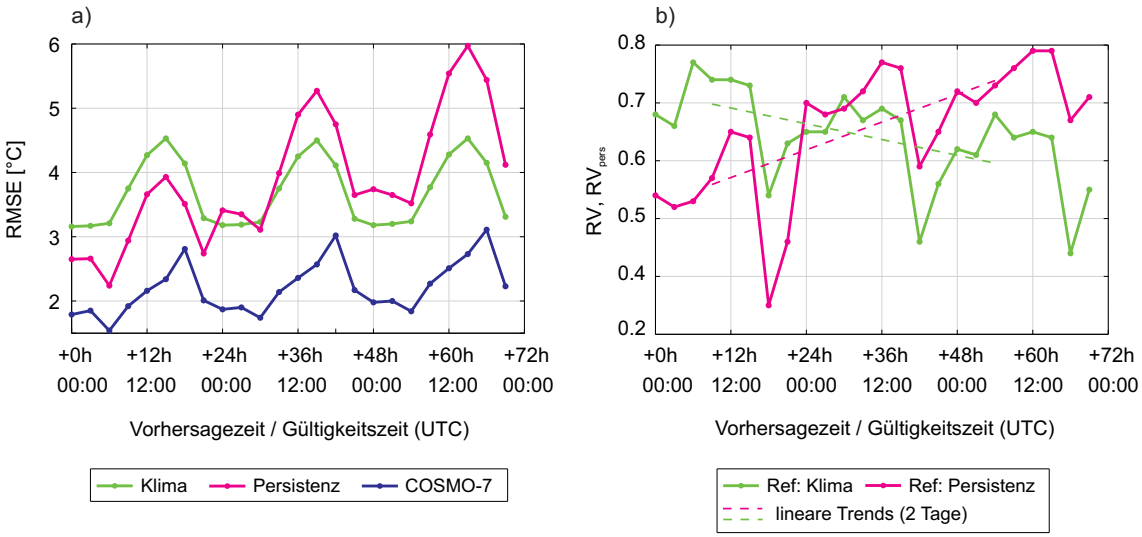

Abbildung 4.20: a) RMSE der COSMO-7-, Klima- und Persistenzprognose, b) RV der COSMO-7 Prognosen mit Klima und Persistenz als Referenz für θ-Vorhersagen im Jun.–Aug, limitiertes D-PHASE Gebiet. 2007.

bildung (4.20 a) zeigt die RMSEs der Klima-, Modell- und Persistenzprognose, deren Quadrate die Basis der RV bilden. Hier werden bereits am zweiten Tag (bis auf den 06 UTC Termin) von der Persistenzvorhersage schlechtere Werte erreicht als vom Klima, nicht wie bei der Korrelation (Abb. 4.19) erst am dritten. Alle Prognosen weisen nachmittags (12–18 UTC) höhere mittlere quadratische Fehler auf als nachts (21–06 UTC), jedoch sind die Tagesverläufe im Detail nicht gleich. So zeigen die Referenzprognosen den schlechtesten (maximalen) RMSE um 15 UTC, also in etwa dann, wenn die höchsten potentiellen Temperaturen des Tages erreicht werden. Der Modell-RMSE wird erst drei Stunden später, wenn θ schon wieder sinkt, am größten. Die Bestwerte des RMSE erhalten Persistenz und Modell um 06 UTC, wenn das Tagesminimum von θ typischerweise schon knapp überschritten ist, während die Klimareferenzprognose die Temperaturen zwischen 21 und 06 UTC mit relativ konstantem Fehler bewertet.

Lässt man diese tageszeitlichen Schwankungen außer Acht, so steigt der RMSE der Persistenzprognose mit der Vorhersagezeit schneller, als der der Modellprognose und

somit wächst auch die resultierende RV mit der Prognosedauer an. Da jedoch keine tatsächliche Steigerung der Prognoseleistung, sondern nur die immer leichter zu übertreffende Persistenzvorhersage dafür verantwortlich ist, kann argumentiert werden, dass die Persistenz für Vergleiche zwischen Vorhersagezeiten über mehrere Tage keine geeignete Referenzprognose darstellt. Der RMSE der Klimareferenzprognose verläuft hingegen, da diese nicht vom Herausgabezeitpunkt abhängt, für jeden Vorhersagetag gleich. Die auf das Klima referenzierte RV gibt das Sinken der Modellleistung mit der Vorhersagezeit akkurat wieder.

Die gezeigten Beispiele beschreiben das typische Verhalten von Persistenz und Klimareferenzprognose, wie es auch für andere Parameter und Zeiträume auftritt. Eine Persistenzprognose als Referenz für *skill scores* zu verwenden hat den Nachteil, dass sie mit steigender Zeitverzögerung schnell schlechter wird und daher Vergleiche zwischen unterschiedlichen Prognosedauern unmöglich macht. Für Evaluierungen mit fixer Prognosedauer ist sie aber durchaus eine mögliche Alternative zur Klimareferenzprognose.

Für den gezeigten dreimonatigen Zeitraum sind beide Referenzprognosen bezüglich RMSE oder Korrelation immer schlechter als die Modellprognose. Werden jedoch einzelne Termine oder kurze Zeitspannen betrachtet, so können Klima- oder Persistenzprognose sehr wohl fallweise besser abschneiden als ein Modell, negative *skill scores* sind die Folge. Besonders häufig tritt dieser unerwünschte Effekt einer „zu guten“ Referenzprognose bei der Persistenz und geringen Zeitverzögerungen von nur einem Tag auf.

Bei *skill score*-Vergleichen über verschiedene Zeitspannen, wie den hier gezeigten monatlichen Werten im Halbjahresverlauf muss immer der Einfluss der Referenzprognose auf das Ergebnis mitberücksichtigt werden. Wenn sich zwei Monate sehr stark in Bezug auf ihre Nähe zum Klimamittel bzw. in Bezug auf die Häufigkeit von persistenten Situationen voneinander unterscheiden, so sind die Unterschiede in den resultierenden *skill scores* unter Umständen unverhältnismäßig groß. Der Verlauf der unreferenzierten Modellleistung wird vom Signal der Referenzprognose überlagert und verdeckt.

4.7 Unsicherheiten der skill scores

Der Permutationstest nach Hamill (1999) ist speziell darauf ausgelegt, die statistische Signifikanz von Unterschieden in den 2x2-*scores* zweier Modelle zu zeigen. In Abb. 4.21 zeigt sich, dass der Unterschied zwischen den jeweils ineinander ge-

nesteten COSMO Modellen bezüglich des PSS für die niedrigeren Schwellwerte im gezeigten Beispiel zu 95 % signifikant ist. Bei den höheren Schwellwerten (5 bzw. 10 mm) trifft dies nicht mehr zu, jedoch sind die Auftritshäufigkeiten (Base Rates) dieser so gering, dass die PSS-Werte selbst nicht mehr unbedingt vertrauenswürdig sind, ein breites Konfidenzintervall ist die logische Folge.

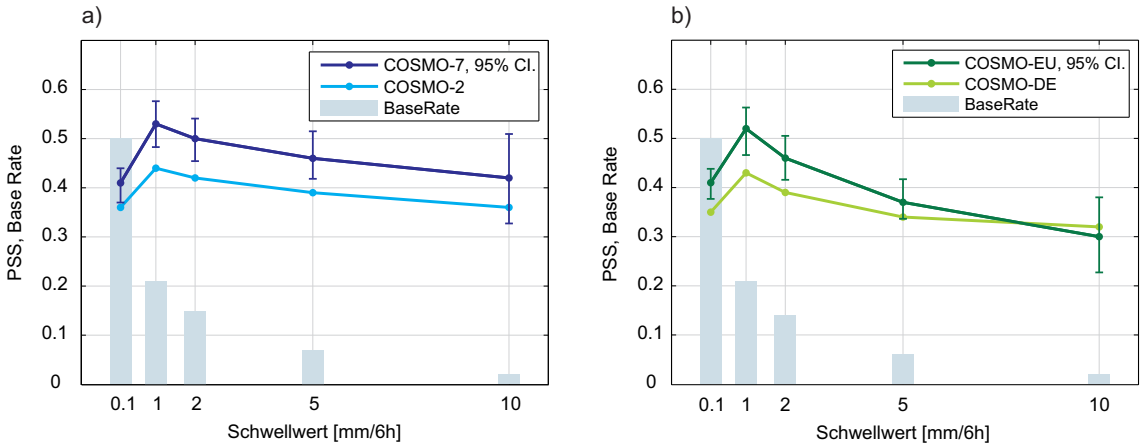

Abbildung 4.21: 95 % Konfidenzintervalle für den PSS, berechnet mit dem Permutationstest aus Abschn. 3.6.2 für den Unterschied zwischen je zwei genesteten Modellen, 18 h Vorhersagen (00 UTC Lauf) für $RR(6\,\mathrm{h})$ im Sommer 2007 im limitierten D-PHASE Gebiet.

Die gröberen Modelle schneiden hier in den Niederschlagsvorhersagen über das limitierte D-PHASE Gebiet besser ab als ihre feineren Gegenstücke. Im Gegensatz zum kleineren COPS Gebiet, wo nicht die intensivsten Niederschläge auftreten, kommt stärker das „*double-penalty*“ Problem zum Tragen: Werden Niederschläge zwar möglicherweise mit der korrekten Intensität aber mit einer raumzeitlichen Verschiebung vorhergesagt, so wird dies von den klassischen punktweisen Maßzahlen „doppelt bestraft“: einmal an der Stelle, wo der Niederschlag vorhergesagt wurde, aber nicht aufgetreten ist und einmal dort, wo er aufgetreten ist, aber nicht vorhergesagt wurde. Besonders trifft dieses Problem die hochauflösenden Modelle bei starken Niederschlagsereignissen, da sie oft kleinskalige, leicht verschobene Strukturen enthalten, die in den gröberen Modellen gar nicht oder abgeschwächt vorhergesagt werden und so dort geringere „Fehler“ ergeben. Des Weiteren ist zu betonen, dass die PSS Werte in Abb. 4.21 nicht Bias-adjustiert sind. Die Verwendung des $\mathrm{PSS_{qb}}$ würde nicht signifikant unterscheidbare Ergebnisse für alle COSMO Modelle bringen.

Der in Abschn. 4.3 besprochene Unterschied im „potentiellen *skill*“ zwischen den höher und niedriger auflösenden COSMO Modellen bei der Niederschlagsvorhersage im COPS Gebiet wird ebenfalls mit dem Hamill'schen Permutationstest auf Signifikanz getestet. Abbildung 4.22 zeigt, dass nicht für alle Quantile die Differenz die gewählten Signifikanzkriterien erfüllt. Besonders für den höchsten Quantilswert ist

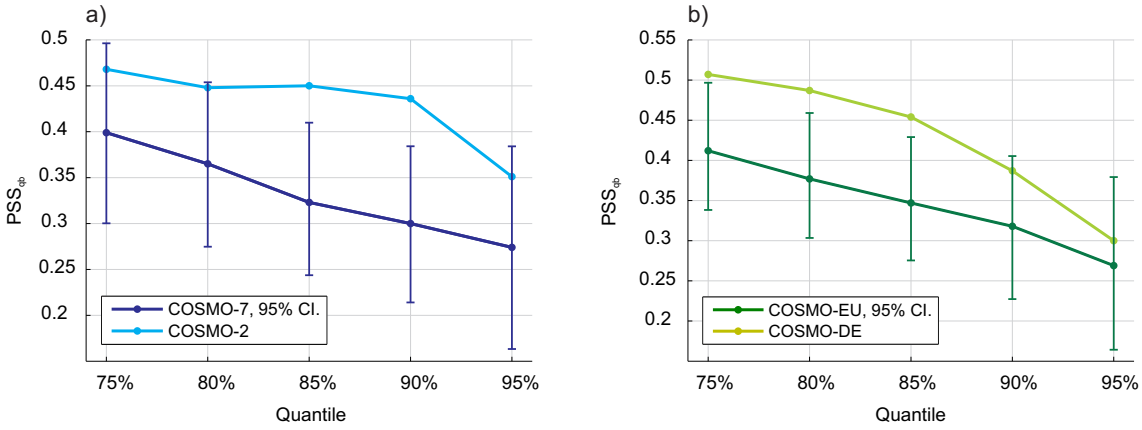

Abbildung 4.22: 95 % Konfidenzintervalle des PSS_{qb} für das Beispiel aus Abb. 4.10 a, berechnet mit dem Permutationstest aus Abschn. 3.6.2 für den Unterschied zwischen je zwei genesteten Modellen. Gezeigt sind nur die Quantile ab 75 %, da erst dann für alle Modelle $RR \geq 0.1$ mm ist (Abb. 4.9).

ein signifikanter Unterschied zwischen den Modellen nur schwer erreichbar. Die Aussage, dass die feineren COSMO Modelle besser als die gröberen abschneiden, kann also für einige Intensitätsbereiche durch statistische Signifikanz untermauert werden. Für manche kann sie es knapp nicht, soll aber nicht aus diesem Grund gänzlich verworfen werden.

Abbildung 4.23 zeigt ein Beispiel für den mit der parametrischen Methode aus Abschn. 3.6.1 berechneten Konfidenzbereich für die Anomaliekorrelation der COSMO-7 Vorhersagen der potentiellen Temperatur. Da die oberen und unteren Grenzen definitionsgemäß nur von der Stichprobengröße und dem Wert des Maßes selbst abhängen, erhalten übereinstimmende AC-Werte für unterschiedliche Vorhersagezeiten dieselben Konfidenzintervalle. Die Ergebnisse für das COSMO-EU liegen für alle Vorhersagezeiten innerhalb des Konfidenzbereichs des COSMO-7. Hier wurde nicht explizit die Signifikanz des Unterschieds zwischen den beiden Modellen untersucht, laut Jolliffe (2007) ist es also wie besprochen theoretisch dennoch möglich, dass eine signifikante Differenz existiert. Aufgrund ihres gleichartigen Aufbaus und der extrem kleinen Unterschiede im Beispiel ist aber ein sehr ähnliches Abschneiden der beiden Modelle durchaus plausibel.

Zum Verlauf der AC über den dreitägigen Vorhersagezeitraum kann man wie schon bei der RV (Abb. 4.20) festhalten, dass die tageszeitlichen Schwankungen deutlich stärker sind als das Sinken des Verifikationsmaßes über die Vorhersagetage für eine fixe Tageszeit. Der Tagesgang der AC folgt in etwa dem mittleren Tagesgang der potentiellen Temperatur selbst, die Maxima werden um 15 UTC erreicht. Um diesen Zeitpunkt herum sind zwar auch die die Abweichungen der Analyse bzw. des Modells vom Klima (gezeigt nur durch den Klima-RMSE in Abb. 4.20) am größten, was

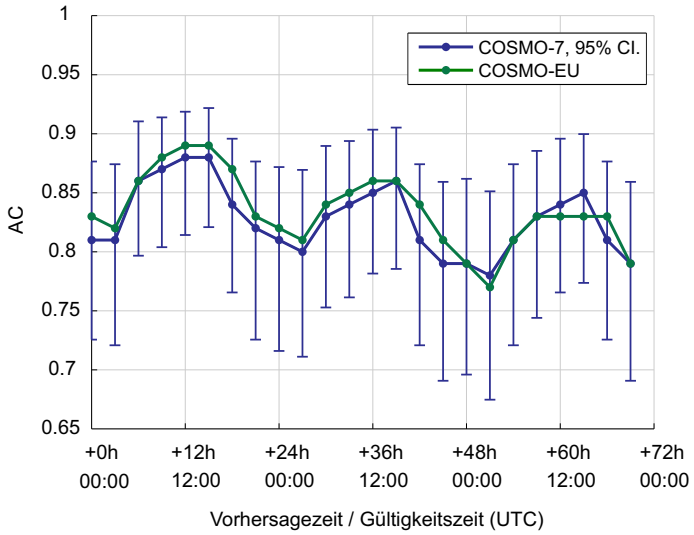

Abbildung 4.23: 95 % Konfidenzintervalle berechnet mit der parametrischen Methode aus Abschn. 3.6.1 für die AC von COSMO-7 Vorhersagen der potentiellen Temperatur im Sommer 2007 im limitierten D-PHASE Gebiet.

aber kein Widerspruch dazu ist, dass gleichzeitig die lineare Korrelation ebendieser (AC) am stärksten ist. Auch gemäß der AC lässt sich also, wie bei Abb. 4.7 für die RV angedeutet, feststellen, dass die nächtliche potentielle Temperatur von den Modellen schlechter mit der Analyse übereinstimmend vorhergesagt wird als die nachmittägliche.

4.8 Modellleistung während der Frontdurchgänge

Fronten sind ein bestimmendes Element im Wettergeschehen der mittleren Breiten und bewirken im Idealfall typische und relativ schnelle Änderungen in allen hier untersuchten meteorologischen Variablen. Es macht daher Sinn, die Modellvorhersagen für diese Fälle gesondert zu betrachten. Die in Abschn. 2.8 in Tabelle 2.2 angeführten Ereignisse bezeichnen Kaltfronten oder kaltaktive Okklusionen, die das „limitierte D-PHASE Gebiet“ oder große Teile dessen überqueren. Diese werden im Folgenden, so auch in der Zeitachse von Abb. 4.24, mit ihrem Anfangsdatum gekennzeichnet.

In Abb. 4.24 sind die RV Werte für die θ_e-Vorhersagen jedes Kaltfront-Ereignisses (+9 bis +18 h, 00 und 12 UTC Lauf) im Vergleich (gestrichelt) zu den Monatswerten der 18 h Vorhersagen zu sehen. Die verschiedenen Fälle erreichen sehr unterschiedliche Ergebnisse, von teilweise negativem *skill* am 02. Juli bis zu sehr hohen Werten um 0.9 für alle Modelle am 05. Oktober oder am 24. November. Fronten und ihre konkreten Auswirkungen können einerseits aufgrund der oft kräftigen Änderungen im Wettergeschehen große Herausforderungen für die Prognosemodelle darstel-

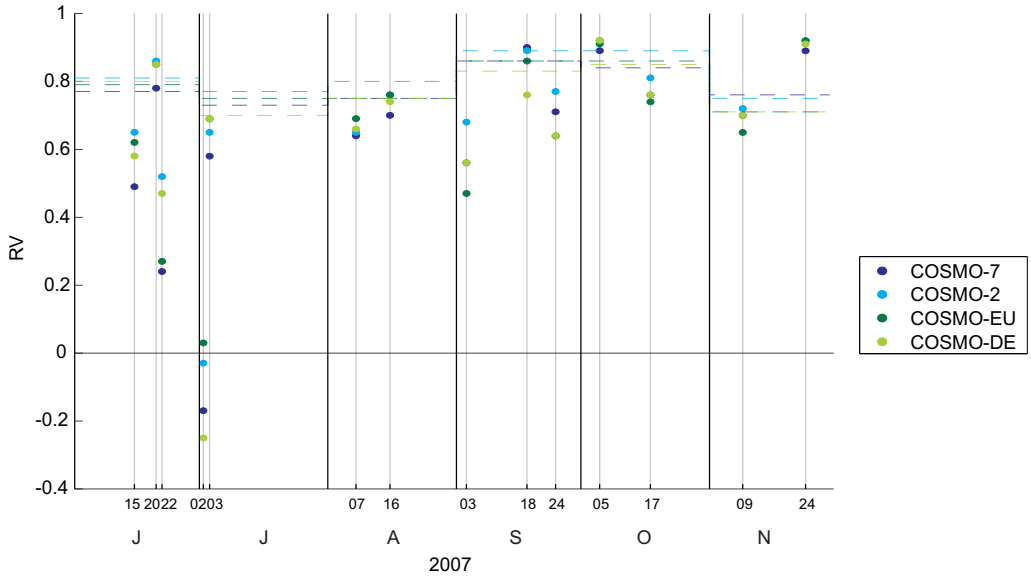

Abbildung 4.24: RV für die θ_e-Vorhersagen während der 14 Kaltfront-Durchgänge (je 9–12 Termine), Vorhersagezeiten: 09–18 h, Modellläufe: 00 und 12 UTC, limitiertes D-PHASE Gebiet ohne Meer. Zusätzlich eingezeichnet (gestrichelt) sind die Monatswerte der RV für die 18 h Vorhersagen (00 UTC Lauf).

len, andererseits sind ihre Verläufe oft schon lange in Vorhinein beobachtbar und die Durchgänge in Mitteleuropa gut vorhersagbar. Es lässt sich eine Überlegenheit des COSMO-2 gegenüber dem COSMO-7 in 13 von 14 Fällen erkennen, wenn sie auch teilweise sehr schwach ist. Unentschieden fällt hingegen der Vergleich zwischen COSMO-DE und COSMO-EU aus. Die Breite der Modellunterschiede variiert stark von Ereignis zu Ereignis, tendenziell ist die Streuung für die einzelnen Kaltfrontdurchgänge größer als im monatlichen Vergleich. Die Ausnahmen bilden besonders jene Fronten mit sehr hohen RV Ergebnissen, wie z. B. für den Fall des 24. November mit einem parallel zur Höhenströmung bereits verwellenden, schwachen, langsamen System, das für alle betrachteten Modelle ähnlich gut kurzfristig vorhersehbar ist.

Aufgrund der starken Unterschiede zwischen den verschiedenen Frontdurchgängen lohnt es sich, den Verlauf von einzelnen Kaltfront-Ereignissen genauer zu betrachten. Abbildung 4.25 zeigt RV (a) und AC (b) für θ_e über die Dauer des Frontdurchgangs vom 18. September, bei dem eine Kaltfront das Gebiet von Nordwest nach Südost überquert und ein relativ starkes Signal im Bodenfeld vorhanden ist (Abb. 4.26). Die AC erreicht ein Maximum am Höhepunkt des Einflusses der Kaltfront auf das Gebiet, in der RV treten zugleich eher die geringsten Werte auf. Die Modell- und Analyseanomalien im Vergleich zum Klima (nicht dargestellt) zeigen positive Abweichungen vor der Front und negative dahinter, liefern aber qualitativ dieselben Muster wie die absoluten Felder in Abb. 4.25. Am stärksten korrelieren Modell- und Analyseanomalien also, wenn die Anzahl der Gitterpunkte mit positiven Werten

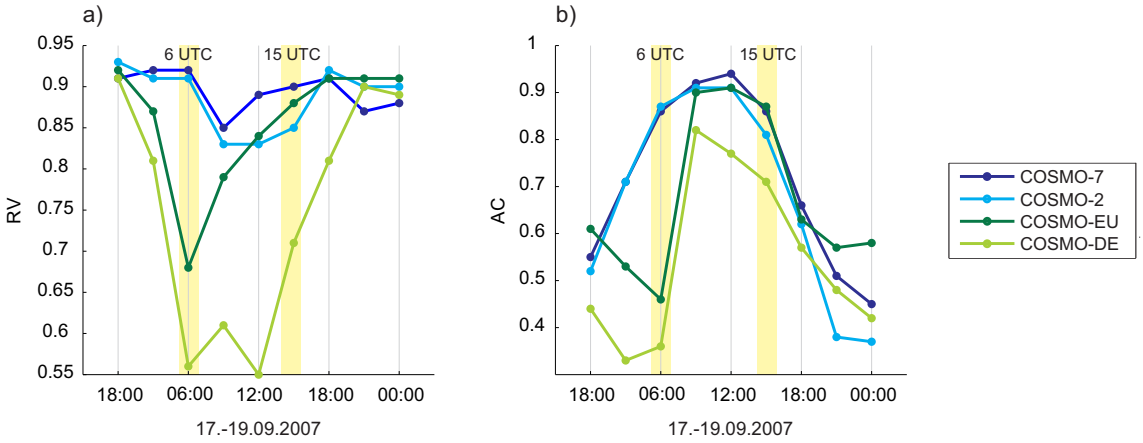

Abbildung 4.25: RV (a) und AC (b), der +9 bis +18 h Vorhersagen (Modelläufe 00 und 12 UTC) für θ_e während eines Kaltfrontdurchgangs im limitierten D-PHASE Gebiet. Gelb markiert sind die Termine für die Karten in Abb. 4.26.

in etwa gleich der mit negativen Werten ist, weil die Front quer über der Domäne liegt. Dies ist zwischen 9 und 12 UTC, also genau zwischen den beiden gezeigten Terminen, der Fall. Der RMSE kann gleichzeitig trotz nur relativ geringer raumzeitlicher Verschiebung der Front, welche jedoch ein großes Gebiet betrifft, sehr hoch werden. Zu dem markierten Termin um 06 UTC ist der Unterschied zwischen den beiden COSMO Modellfamilien besonders groß, um 15 UTC nähern sich die Modelle bereits wieder an, obwohl COSMO-DE immer noch schlechter als COSMO-2 abschneidet. Die Gründe für die Unterschiede der beiden hochauflösenden Modelle sind in den Analyse- und Modellfeldern für diese beiden Termine in Abb. 4.26 auszumachen. Um 6 UTC hat die Kaltfront in der Analyse bereits den Schwarzwald und den Nordwesten der Schweiz überquert, im COSMO-2 liegt sie nur knapp dahinter, während sie im COSMO-DE noch vor Erreichen des Elsass steht. Um 15 UTC hat die Front die Alpen passiert und die Modellunterschiede sind geringer. Dennoch liegt die Luftmassengrenze im COSMO-DE immer noch zurück und das θ_e-Signal verwellt zusätzlich im nördlichen sichtbaren Abschnitt. Es enthält also nur wenig bzw. keine Abkühlung östlich von Vorarlberg.

Für eine weitergehende Analyse dieses Ereignisses wird im nächsten Abschnitt (4.9.1) die Entwicklung der Kaltfront über die COPS-Untergebiete mit Einbeziehung der Windvorhersagen verfolgt.

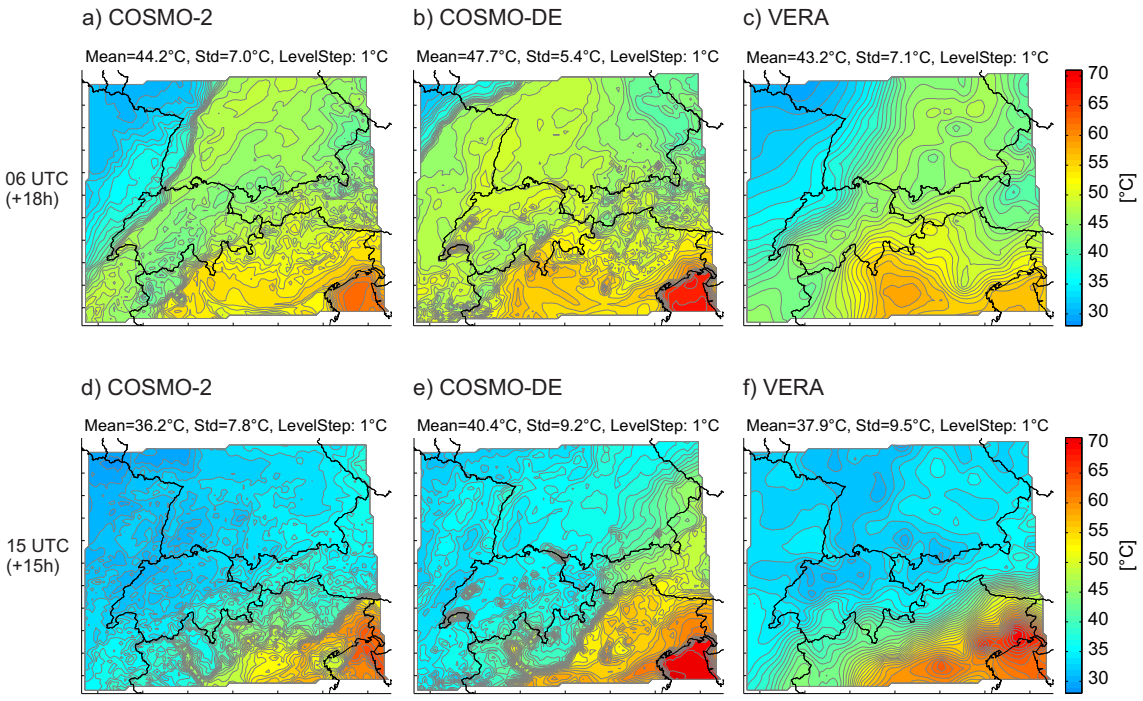

Abbildung 4.26: Äquivalentpotentielle Temperatur zu zwei ausgewählten Terminen des Kaltfrontdurchgangs am 18. 09. 2007 für COSMO-2 (a, d), COSMO-DE (b, e) und VERA (c, f).

4.9 Räumliche Stratifizierung

4.9.1 Das Kaltfrontereignis vom 18.09.

Eine geographische Unterteilung im komplexen Gelände des COPS Gebiets wird gemäß Abb. 2.4 durchgeführt. Die im vorigen Abschnitt 4.8 bereits besprochene Kaltfront vom 18. 09. 2007 überquert die Vogesen und den Schwarzwald von West nach Ost und ist nahezu parallel zu den Gebirgszügen, nur etwas mehr gegen NO und SW, geneigt. Abbildung 4.27 zeigt θ_e-Abweichungen Modell - VERA und die Windvektoren beider für COSMO-2 und COSMO-7. Wie für das COSMO-2 und den 06 UTC Termin schon in Abb. 4.26 erkennbar, ist die Front im Modell gegenüber der Analyse nach hinten (Westen) verschoben. In den Differenzenfeldern (Abb. 4.27) zeigt sich dies durch ein Band hoher positiver θ_e-Anomalien, dessen vorderer (östlicher) Rand in etwa die Analyse-Front kennzeichnet, während der hintere (westliche) Gradient die Modell-Front wiedergibt, gestört nur durch kleinskaligere Signale im jeweils anderen Feld. Exakter markieren die zyklonalen Windsprünge, von Südwest vor der Front auf West bis Nordwest danach, ihre Lage. Um 03 UTC hat die Kaltfront in der Analyse die Vogesen knapp überschritten, in COSMO-2 und COSMO-7 erreicht sie gerade die Gebietsgrenze. Drei Stunden später reicht die Analyse-Front

nur mehr in den südöstlichsten Teil des COPS-Gebiets, in COSMO-2 spannt sie sich über den Schwarzwald während sie in COSMO-7 weitgehend knapp östlich des Rheins, im Süden noch westlich des Rheins liegt. Durch das Auftreffen auf die im Vergleich zu den Vogesen relativ scharfe Kante des Schwarzwalds, an der die kalte Luft am Boden zunächst aufgehalten wird, verstärken sich die θ_e-Gradienten der Fronten. Zu sehen ist dies vor allem in den Modellen, da der Zeitpunkt an dem die Analyse Front am Schwarzwald liegt, nicht aufgelöst ist. Die Modelle, vor allem COSMO-2, zeigen generell einen stärkeren Gradienten als die Analyse.

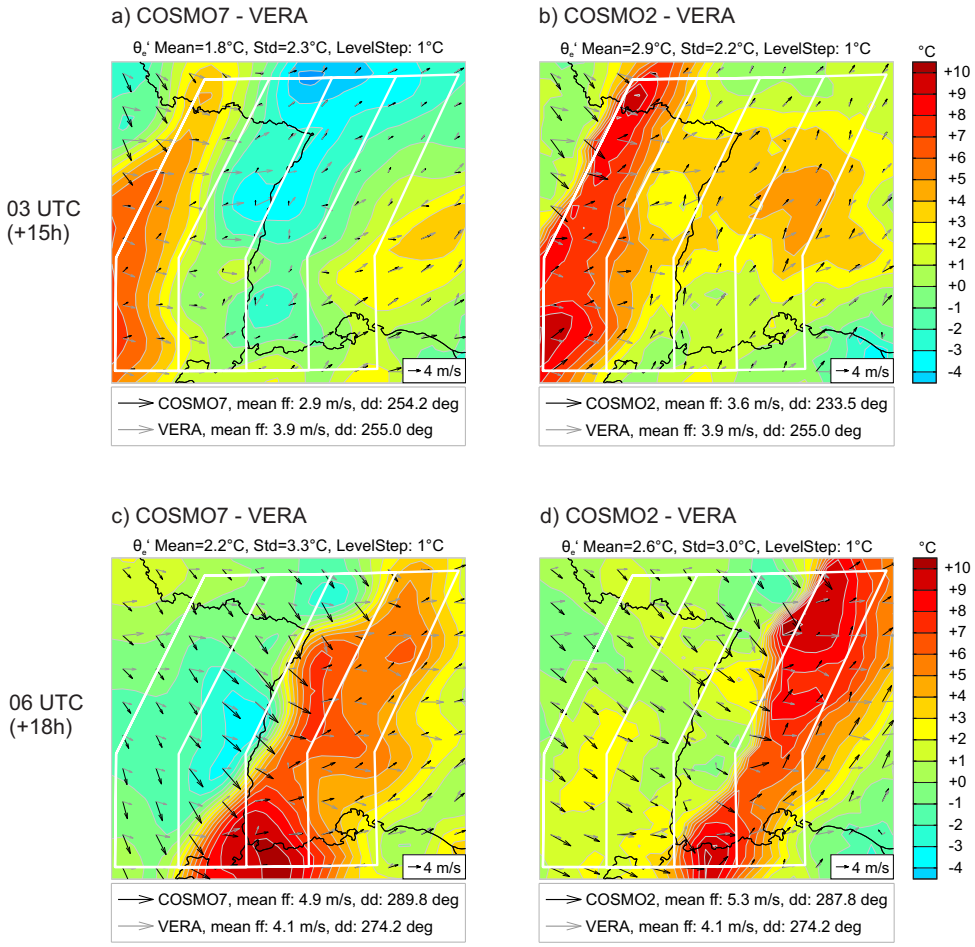

Abbildung 4.27: θ_e-Differenzen Modell - VERA und Windvektoren beider (gezeichnet an 1 von 9 Gitterpunkten) für zwei Termine am 18. 09. 2007 im COPS Gebiet für COSMO-7 (a, c) und COSMO-2 (b, d).

Beide Modelle und beide Termine, in denen die Kaltfront das Gebiet beeinflusst, werden in jeder Teilabbildung von 4.28 anhand des RMSE bzw. der AC für θ_e und Wind verglichen, wobei wieder die Abweichungen vom Mittel aufgetragen sind. Die RMSEs der äquivalentpotentiellen Temperatur (a) lassen sich direkt aus den Differenzen in Abb. 4.27, mit den größten Werten im Einflussbereich des Zwischenfrontgebiets, ableiten. Für das COSMO-2 ist dies um 03 UTC das Vogesen-Gebiet

und um 06 UTC der Schwarzwald. COSMO-7 liegt um 06 UTC weiter zurück und weist somit große Fehler sowohl im Rhein- als auch im Schwarzwald-Gebiet auf. Um 03 UTC in den Vogesen ist der Fehler des COSMO-7 beträchtlich geringer als der des COSMO-2, obwohl beide Modellfronten ähnlich weit zurück liegen, da im gröberen Modell sowie in VERA der θ_e-Gradient nicht so scharf ausgeprägt ist wie im hochauflösenden Modell.

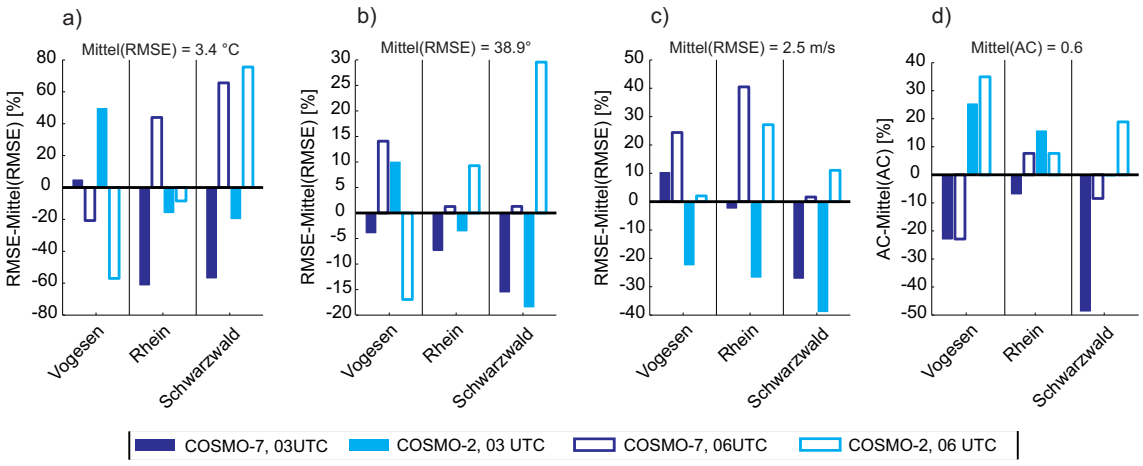

Abbildung 4.28: Abweichungen RMSE - Mittelwert(RMSE) in den COPS Untergebieten für a) θ_e, b) *dd* (*ff* $\geq$ 1m/s) und c) *ff* ($\geq$ 1m/s), sowie AC - Mittelwert(AC) für θ_e (d) für das Bsp. aus Abb. 4.27.

In der Windrichtung ist das Band der durch die unterschiedliche Frontenlage bedingten Abweichungen gegenüber dem in θ_e etwas nach Osten verschoben, da wie besprochen die genaue Position der Front, wo auch der Windsprung auftritt, am warmen (für die Kaltfront vorderen) Rand des θ_e-Gradienten zu finden ist. Im gezeigten Beispiel treten aber auch außerhalb des Fronteinflusses, vor allem für das COSMO-7, relativ starke systematische Abweichungen zwischen Modell und VERA auf. Besonders ab einem gewissen Abstand (etwa 50 km) hinter der Front ist das Windfeld im Modell gegenüber dem in VERA signifikant nach rechts verschoben. Diese systematische Rechtsablenkung ist konsistent mit der in den mittleren sommerlichen Windfeldern in einem großen Teil der Nordhälfte des limitierten D-PHASE Gebiets (Abb. 4.15). Die Bodendruckmuster sind im Mittel und im betrachteten Einzelfall gegeneinander verschoben. Für *dd* ist der Verlauf der RMSEs über die Gebiete (Abb. 4.28 b) beim COSMO-2 ähnlich wie für θ_e, mit den höchsten RMSEs in den jeweiligen Einflussgebieten der Front, wobei hier zum zweiten Termin (Schwarzwald) signifikant größere Fehler gemacht werden als zum ersten (Vogesen). COSMO-7 zeigt hingegen nach Osten hin sinkende RMSEs, alle 03 UTC Werte sind besser als alle 06 UTC Werte. Das heißt, beim COSMO-7 machen die systematischen Fehler hinter der Front mehr oder gleich viel aus wie die Fehler im Fronteinflussbereich.

Die Windgeschwindigkeit frischt mit Durchzug der Front auf und diese Verstärkung kann danach einige Zeit anhalten. Das Band der frontal bedingten Abweichungen ist also gegenüber dem der *dd* nach Westen verschoben und ist breiter. Zum 03 UTC Termin nehmen die RMSEs daher nach Osten hin ab (d), denn das Auffrischen nach der Analysefront liegt vollständig im Vogesen Gebiet. Zum 06 UTC Termin liegt das RMSE-Maximum für beide Modelle im Rhein Gebiet, da die Modellfronten noch nicht über den Schwarzwald hinaus sind.

Die Frage, welches der beiden COSMO Modelle die wesentlichen Parameter während des Frontdurchgangs besser voraussagt, ist nicht eindeutig beantwortet. Ein grober Blick auf Abb. 4.27 zeigt, dass die Lage der Front um 03 UTC von beiden in etwa gleich, um 06 UTC von COSMO-2 besser vorhergesagt wird, die Magnitude der θ_e-Abweichungen aber in COSMO-7 durch die schwächeren Gradienten vor allem um 03 UTC geringer ausfällt. Aus diesem Grund ist der RMSE für θ_e um 03 UTC für das gröbere Modell besser, um 06 UTC nähern sich die beiden an. Der geringere Ortsfehler führt jedoch in den Anomaliekorrelationen zu einem besseren Abschneiden des COSMO-2 gegenüber dem COSMO-7 (Abb. 4.28 d). Bezüglich der Windgeschwindigkeitsvorhersagen ist überwiegend das COSMO-2 überlegen. Für die Windrichtung erreicht in den Gebieten des Frontaleinflusses das COSMO-7 die besseren RMSEs.

In allen Parametern schneiden, wenn man jeweils das Haupteinflussgebiet der Front betrachtet, die Modelle um 03 UTC (Vogesen) besser ab als um 06 UTC (Schwarzwald, bzw. Rhein für *ff*). Beim COSMO-7 ist das vor allem auf das stärkere Zurückfallen der Front zum zweiten Termin zurückzuführen, beim COSMO-2 für θ_e auf das Verschärfen des Gradienten. Für die Windgeschwindigkeit liegt der Grund jedoch darin, dass die Unterschiede nach der Modellfront (um 06 UTC im Gebiet) stärker sind als die nach der Analysefront (03 UTC).

Die Verbesserung in der Prognose der Lage der Front mit sinkender Vorhersagezeit wird in Abb. 4.29 deutlich. Wie in Abb. 4.27 c sind die Differenzenfelder zwischen COSMO-7 und VERA zum 06 UTC Termin gezeigt, hier jedoch für steigende Vorhersagezeiten vom Initialisierungstermin 00 UTC. Bei den längeren Laufzeiten von 30 und 54 Stunden (b und c) liegt die vorhergesagte Front noch weiter zurück als bei der 18 h Prognose in Abb. 4.27 c. In der 54 h Vorhersage steht die Modellfront noch am westlichen Rand des COPS Gebiets. Die sechsstündige Vorhersage (a) gibt hingegen eine im nördlichen Teil sehr gut mit der Analyse übereinstimmende Frontenlage, an der Dreiländergrenze im Süden liegt die Modellfront jedoch auch hier noch zurück. Der Bias, im Bild angegeben als Kartenmittel der Anomalien („θ'_e *Mean*“), wird für diese Prognose Null.

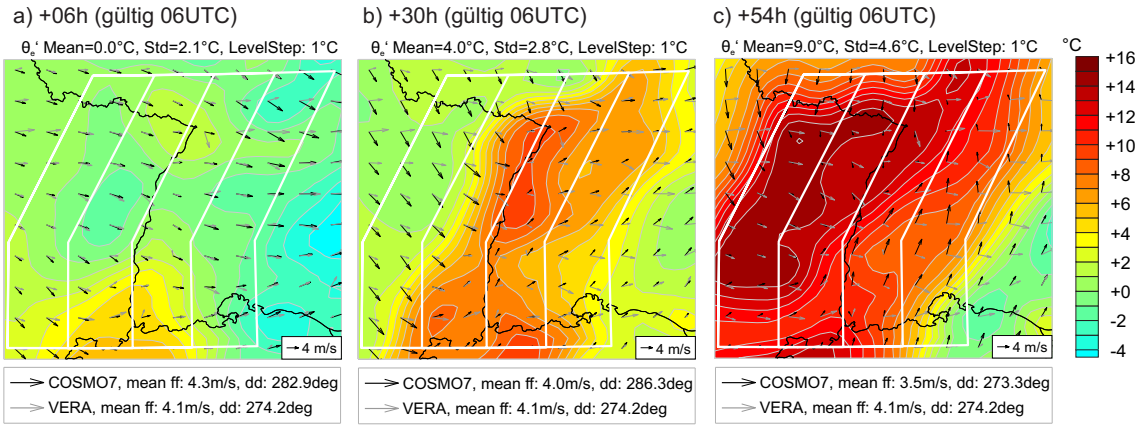

Abbildung 4.29: θ_e-Differenzen COSMO7 - VERA und Windvektoren beider (gezeichnet an 1 von 9 Gitterpunkten) am 18. 09. 2007 06 UTC im COPS Gebiet für steigende Vorhersagezeit: a) +06 h, b) +30 h, c) +54 h (00 UTC Lauf)

4.9.2 Systematische Unterschiede in den COPS-Gebieten

Auch bei der Untersuchung der Modellleistung über längere Zeiträume können Unterschiede in den Vorhersagen für die COPS-Teilregionen auftreten. Als Beispiel ist ein gebietsabhängiges Gefälle in der Wiedergabe der Bodendruckmuster in Abb. 4.30 sichtbar gemacht. Die 18 h Vorhersagen im Juni liegen für alle Modelle bezüglich der Korrelation und des Bias-korrigierten RMSE in den Vogesen näher an der Analyse als im Schwarzwald. Die Ergebnisse für das Rhein Gebiet, das sich aus Teilflächen von beiden zusammensetzt, liegen, wie aufgrund des relativ glatten Verlaufs der Druckfelder zu erwarten, dazwischen (nicht eingezeichnet).

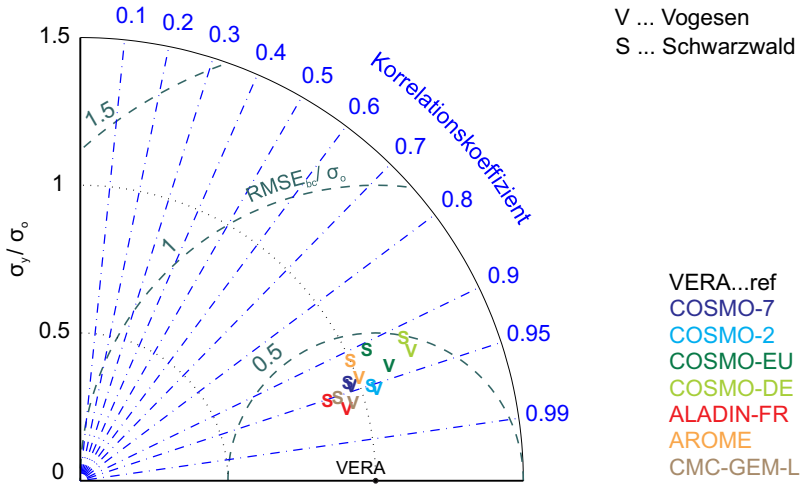

Abbildung 4.30: Taylor Diagramm für die 18 h Vorhersagen (00 UTC Lauf) des reduzierten Drucks im Juni 2007 in Vogesen und Schwarzwald.

Trotz dieser Regelmäßigkeit sind die Modellunterschiede weiterhin größer als die Unterschiede zwischen den Regionen. Soll dennoch ein Grund für letztere ausgemacht

werden, so liegen diese mit hoher Wahrscheinlichkeit in der speziellen Topographie der beiden Gebirgszüge, welche in den Vogesen nach Westen hin viel flacher verläuft als am Schwarzwald und welche auch mit im Mittel flacheren Druckgebilden in dieser Region in Zusammenhang steht. Am weitesten im Taylordiagramm von der Analyse entfernt und somit im Modellvergleich am schlechtesten gemäß der Fehlermaße sind COSMO-DE und COSMO-EU. Generell wird der reduzierte Bodendruck jedoch erwartungsgemäß für alle Modelle mit sehr hohen Korrelationen vorhergesagt.

5 Conclusio

Neben den eigentlichen Auswertungen der Modellleistung enthält vorliegende Arbeit eine umfangreiche Erarbeitung der „klassischen“, also nicht-räumlichen, nicht-objektbasierten Verifikationsmethoden für deterministische Modellvorhersagen meteorologischer Bodenfelder. Besondere Aufmerksamkeit gilt der Gruppe der *skill scores*, denn die Vorhersagegüte im Vergleich zu einer Referenzprognose ist ein besonders reizvolles Kennzeichen. Es zeigt sich, dass für binäre Prädiktanden eine große Anzahl an unterschiedlichen Formulierungen von *skill scores* existiert, die verschiedene, in Tabelle 3.2 zusammengefasste Attribute aufweisen. Dennoch ist es schwierig eine eindeutige Empfehlung hinsichtlich der Wahl des adäquaten Maßes abzugeben. Zum einen bringen einige in typischen Fällen kaum unterschiedliche Ergebnisse (HSS, PSS), zum anderen können die jeweiligen Eigenschaften situationsbedingt erwünscht oder unerwünscht sein, wie z.B. die Möglichkeit eines perfekten Wertes für Vorhersagen mit Bias beim ORSS. In jedem Fall muss bei Verwendung von *skill scores* dieser Art gleichzeitig der Bias berücksichtigt werden, sehr nützlich ist oft eine Bias-Adjustierung. Für die kontinuierlichen Prädiktanden geben Reduction of Variance und Anomaliekorrelation wichtige, im Allgemeinen nicht redundante Aspekte der Vorhersageleistung wieder.

Nicht zu unterschätzen ist der Einfluss der Referenzprognose, bzw. der aktuellen meteorologischen Situation im Vergleich zur Referenzsituation, auf die *skill scores*. Für die kontinuierlichen Prädiktanden wird in vorliegenden Auswertungen eher das Klima als Referenz gegenüber der Persistenz bevorzugt, da es allgemeiner, also auch für Vergleiche über steigende Vorhersagezeiten, anwendbar ist und da für die sehr kurzfristigen Prognosen von weniger als 24 Stunden besonders bei Bewertungen von kurzen Zeitspannen oder Einzelfällen die Persistenzprognose nur sehr schwer von den Modellen zu übertreffen ist. Grundsätzlich ist es meist sinnvoll, vor der *skill score* Auswertung auch grundlegendere Maße ohne Referenzbezug zu berechnen.

Eine Quantifizierung der Unsicherheit eines Verifikationsmaßes ist wünschenswert, besonders der auf die Signifikanz von Unterschieden ausgerichtete Permutationstest gemäß Hamill (1999) ist für den Modellvergleich nützlich. Bei der Interpretation von Konfidenzintervallen muss jedoch immer die zur Berechnung verwendete Methode

bekannt sein. So wird beispielsweise die Breite des beschriebenen einfachen parametrischen Konfidenzintervalls für die AC zum überwiegenden Teil von der Stichprobengröße bestimmt.

Aus den Ergebnissen des Modellvergleichs für den Sommer und Herbst 2007 lassen sich einige zentrale Punkte zusammenfassen:

- Für die Vorhersagen der Temperaturgrößen und des Luftdrucks schneidet bezüglich RMSE, RV und teils des Bias (θ, θ_e) in den monatlichen Ergebnissen, sowie beim Großteil der Frontdurchgänge das COSMO-2 besser ab als das übergeordnete COSMO-7. In der COSMO Kette des DWD fällt der Vergleich zwischen feinerem und gröberem Modell COSMO-DE und COSMO-EU hingegen teils umgekehrt, teils in dieselbe Richtung aus und oftmals sind die Verifikationsergebnisse der beiden kaum unterscheidbar.

  In vielen Fällen liegen die Maße für COSMO-7 und COSMO-EU auf ähnlichem Niveau, von dem sich COSMO-2 bemerkbar absetzt, COSMO-DE jedoch nicht. Der Hauptunterschied der beiden COSMO Familien liegt im antreibenden Globalmodell, dennoch machen sich in den Verifikationsergebnissen für die relativ glatten meteorologischen Parameter in den untersuchten Beispielen die stärksten Unterschiede in der kleinsten „*nesting*“-Stufe bemerkbar. Dies ist ein erstaunliches Ergebnis, da die COSMO Modelle der jeweils (annähernd) gleichen Auflösung dieselbe Modellphysik und dieselben Datenassimilationsmethoden verwenden. Man kann daher darauf schließen, dass auch die genauen Gitterspezifikationen und Berechnungsdomänen einen nicht unerheblichen Einfluss auf das Vorhersageergebnis haben.

- Für den sechsstündigen sommerlichen Nachmittagsniederschlag in der COPS Region zeigt sich ein höherer „potentieller“ *skill* der feiner aufgelösten COSMO Modelle im Vergleich zu ihren gröberen Gegenstücken, allerdings bei einem gleichzeitig höheren Bias. Für das größere limitierte D-PHASE Gebiet verschwindet der Vorsprung im *skill* jedoch. Aufgrund des bekannten „*double penalty*“ Problems ist es vor allem für den Niederschlag ratsam, auf räumliche oder objektbasierte Verifikationsmethoden zurückzugreifen, besonders im Vergleich unterschiedlicher Modellauflösungen. Dennoch bietet die Verwendung der quantilsbasierten Maßzahlen bereits deutliche Vorteile gegenüber den klassischen *skill scores*.

- Sowohl für die Vorhersagen der Windrichtung als auch der Windgeschwindigkeit erreicht ALADIN-FR, das einzige hydrostatische Modell, in den durchgeführten Auswertungen zumeist die besten Ergebnisse im Vergleich zu den ande-

ren. Auch das CMC-GEM-L liegt anders als für die übrigen Parameter bei der Windrichtungsprognose im Vorderfeld des Modellvergleichs. Der verhältnismäßig höhere Gitterpunktsabstand scheint eine VERA-ähnliche Windvorhersage nicht zu erschweren. Die COSMO Modelle weisen hingegen hohe positive Biase der Windgeschwindigkeit und, besonders COSMO-2 und COSMO-7, im Flachland eine starke Rechtsablenkung des Windes gegenüber VERA, das heißt eine weniger starke Drehung zum tieferen Druck hin, auf. Dies legt die Vermutung nahe, dass bei diesen Modellen die Reibung in der Grenzschichtparametrisierung möglicherweise zu schwach angesetzt ist.

- Das kanadische CMC-GEM-L liegt mit seiner Auflösung noch eine Stufe unter den gröberen Modellen der anderen Modellketten und der Analyse, das heißt, es weist den höchsten Gitterpunktsabstand auf. Es ist daher nicht verwunderlich, dass das Modell in vielen Auswertungen, mit Ausnahme derer für die Windrichtung, vergleichsweise schwächere Ergebnisse erzielt. Für das AROME lassen sich in manchen Teilresultaten noch Schwierigkeiten in der Testphase erahnen.

Im Laufe der Erstellung dieser Arbeit wurde eine Fülle von weiterführenden Gesichtspunkten und Ideen aufgeworfen, welche jedoch aus verschiedenen Gründen nicht miteinbezogen werden konnten. Einige sollen hier kurz erwähnt werden, um eventuell Anstöße für zukünftige Auswertungen zu geben.

- Zusätzlich zu den hier betrachteten, hochauflösenden LAMs, sollten idealerweise auch die jeweiligen zugehörigen Globalmodelle in den Vergleich eingeschlossen werden, um drei Stufen einer Modellkette und eine mögliche Fehlerfortpflanzung untersuchen zu können. Erste Auswertungen in diese Richtung wurden im Rahmen des Projekts VERITA, seitdem diese Daten zur Verfügung stehen, bereits durchgeführt.
- Es wäre außerdem wünschenswert, die saisonalen Vergleiche auf den Winter und den Frühling auszuweiten, also eine Zeitspanne von zumindest einem Jahr zu betrachten. Dadurch könnten überdies weitere spezifische Wetterlagen zusätzlich zu den Kaltfrontdurchgängen herausgefiltert werden. Auch ein möglichst großes räumliches Gebiet ist für detaillierte Evaluierungen anzustreben und würde eine Unterteilung in klimatologisch und topographisch verschiedenartige Regionen erlauben.
- Wie bereits besprochen, sollte jedes berechnete Gütemaß mit einer Angabe von dessen Unsicherheit versehen werden. Aufgrund des hohen Rechenaufwandes wird das in dieser sowie in vielen anderen Arbeiten noch nicht durchgängig

gemacht.

- Im Bezug auf die Wahl der Referenzprognose für Skill Scores kann man in Erwägung ziehen, eine Mischung aus den hier behandelten Klima- und Persistenzprognosen zu formen. Es könnte beispielsweise zu Beginn der Vorhersageperiode die Persistenz herangezogen werden und erst ab einem gewissen Zeitpunkt das Klima, sodass immer die „bessere“ der beiden Referenzprognosen übertroffen werden muss.
- Um die Unterschiede zwischen den ineinander genesteten LAMs genauer herauszuarbeiten, wäre es aufschlussreich, nur die inneren Ausschnitte der Modellgebiete zu vergleichen. Es ist zu vermuten, dass in der Nähe der Ränder der Einfluss des übergeordneten Modells dominiert, und die Andersartigkeit des jeweils feineren Modells erst im mittleren Bereich der Domäne am stärksten zum Ausdruck kommt.

Abschließend ist zu erwähnen, dass Teile dieser Arbeit im Mai 2011 im Rahmen der ICAM (31st International Conference on Alpine Meteorology) als Vortrag und im November 2011 am 4. Österreichischen Meteorologentag als Poster präsentiert wurden.

Literatur

Agresti, A., und B.A. Coull, 1998: Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126.

Ament, F., und M. Arpagaus, 2009a: dphase_cosmoch2: COSMO model forecasts (2.2 km) run by MeteoSwiss for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_cosmoch2. `http://dx.doi.org/10.1594/WDCC/dphase_cosmoch2` (abgerufen am 08.03.2012).

Ament, F., und M. Arpagaus, 2009b: dphase_cosmoch7: COSMO model forecasts (7 km) run by MeteoSwiss for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_cosmoch7. `http://dx.doi.org/10.1594/WDCC/dphase_cosmoch7` (abgerufen am 08.03.2012).

Baldauf, M., J. Förstner, S. Klink, T. Reinhardt, C. Schraff, A. Seifert, und K. Stephan, 2011: Kurze Beschreibung des Lokal-Modells Kürzestfrist COSMO-DE (LMK) und seiner Datenbanken auf dem Datenserver des DWD. DWD.

Baldauf, M., K. Stephan, S. Klink, C. Schraff, A. Seifert, J. Foerstner, T. Reinhardt, C.J. Lenz, und M. Denhard, 2009: dphase_lmk: LMK (COSMO-DE) high resolution model forecasts run by DWD for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_lmk. `http://dx.doi.org/10.1594/WDCC/dphase_lmk` (abgerufen am 08.03.2012).

Bauer, H.S., T. Weusthoff, M. Dorninger, V. Wulfmeyer, T. Schwitalla, T. Gorgas, M. Arpagaus, und K. Warrach-Sagi, 2011: Predictive skill of a subset of models participating in D-PHASE in the COPS region. *Quarterly Journal of the Royal Meteorological Society*, **137**, 287–305.

Bazile, E., 2009: dphase_aladfr: ALADIN-France operational model forecasts run by Meteo-France for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_aladfr. `http://dx.doi.org/10.1594/WDCC/dphase_aladfr` (abgerufen am 08.03.2012).

Bougeault, P., P. Binder, A. Buzzi, R. Dirks, J. Kuettner, R. Houze, R.B. Smith, R. Steinacker, und H. Volkert, 2001: The MAP special observing period. *Bulletin of the American Meteorological Society*, **82**, 433–462.

Bouttier, F., 2010: The Météo-France NWP system: description, recent changes

and plans. Météo-France.

Bubnova, R., A. Horanyi, und S. Malardel, 1993: International project ARPEGE-ALADIN. *EWGLAM Newsletter*, **22**, 117–130.

Casati, B., G. Ross, und D. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, **11**, 141–154.

Coles, S., J. Heffernan, und J. Tawn, 1999: Dependence measures for extreme value analyses. *Extremes*, **2**, 339–365.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, und A. Staniforth, 1998: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Monthly Weather Review*, **126**, 1373–1395.

Doms, D., J. Förstner, E. Heise, H.J. Herzog, D. Mironov, M. Raschendorfer, T. Reinhardt, B. Ritter, R. Schrodin, J.P. Schulz, und G. Vogel, 2011: A description of the nonhydrostatic regional COSMO-model, Part II: Physical parameterization. Deutscher Wetterdienst, `http://www.cosmo-model.org` (abgerufen am 08.03.2012).

Doms, G., 2011: A description of the nonhydrostatic regional COSMO-model, Part I: Dynamics and numerics. Deutscher Wetterdienst, `http://www.cosmo-model.org` (abgerufen am 08.03.2012).

Doolittle, M., 1885: The verification of predictions. *Bulletin of the Philosophical Society of Washington*, **7**, 122–127.

Dorninger, M., T. Gorgas, T. Schwitalla, M. Arpagaus, M. Rotach, und V. Wulfmeyer, 2009: Joint D-PHASE–COPS dataset (JDC dataset). Universität Wien.

Eigenmann, R., N. Kalthoff, T. Foken, M. Dorninger, M. Kohler, D. Legain, G. Pigeon, B. Piguet, D. Schüttemeyer, und O. Traulle, 2011: Surface energy balance and turbulence network during the Convective and Orographically-induced Precipitation Study (COPS). *Quarterly Journal of the Royal Meteorological Society*, **137**, 57–69.

Ferro, C.A.T., 2007: A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting*, **22**, 1089–1100.

Finley, J., 1884: Tornado predictions. *American Meteorological Journal*, **1**, 85–88.

Fischer, C., T. Montmerle, L. Berre, L. Auger, und S.E. Ştefănescu, 2005: An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. *Quarterly Journal of the Royal Meteorological Society*, **131**, 3477–3492.

Gandin, L., und A.H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Monthly Weather Review*, **120**, 361–370.

Göber, M., E. Zsótér, und D.S. Richardson, 2008: Could a perfect model ever satisfy a naïve forecaster? On grid box mean versus point verification. *Meteorological Applications*, **15**, 359–365.

Gilbert, G.F., 1884: Finley's tornado predictions. *American Meteorological Journal*, **1**, 166–172.

Gorgas, T., 2006: Vergleich von numerischen Wettervorhersage-Modellen mittels VERA. Diplomarbeit, Universität Wien.

Gorgas, T., und M. Dorninger, 2012a: Concepts for a pattern-oriented analysis ensemble based on observational uncertainties. *Quarterly Journal of the Royal Meteorological Society*, **138**, 769–784.

Gorgas, T., und M. Dorninger, 2012b: Quantifying verification uncertainty by reference data variation. Angenommen bei: Meteorologische Zeitschrift.

Gruber, C., 2005: Evaluierung von modellierten Windfeldern im Alpenraum. Diplomarbeit, Universität Wien.

Hamill, T.M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14**, 155–167.

Hamill, T.M., und J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, **132**, 2905–2923.

Hogan, R.J., C.A.T. Ferro, I.T. Jolliffe, und D.B. Stephenson, 2010: Equitability revisited: Why the "equitable threat score" is not equitable. *Weather and Forecasting*, **25**, 710–726.

Hogan, R.J., E.J. O'Connor, und A.J. Illingworth, 2009: Verification of cloud fraction forecasts. *Quarterly Journal of the Royal Meteorological Society*, **135**, 1494–1511.

Jenkner, J., C. Frei, und C. Schwierz, 2008: Quantile-based short-range QPF evaluation over Switzerland. *Meteorologische Zeitschrift*, **17**, 827–848.

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

Jolliffe, I.T., und D.B. Stephenson (Hrsg.), 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, R. Jenne, und D. Joseph, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, **77**, 437–471.

Kiesenhofer, S., 2011: Vorhersageverifikation hochauflösender Modelle mit MET und VERA. Diplomarbeit, Universität Wien.

Knabl, T., 2004: Klimatologie des Bodendruckfeldes im Alpenraum. Diplomarbeit, Universität Wien.

Mailhot, J., S. Bélair, L. Lefaivre, B. Bilodeau, M. Desgagné, C. Girard, A. Glazer, A.M. Leduc, A. Méthot, A. Patoine, A. Plante, A. Rahill, T. Robinson, D. Talbot, A. Tremblay, P. Vaillancourt, A. Zadra, und A. Qaddouri, 2006: The 15-km version of the Canadian regional forecast system. *Atmosphere-Ocean*, **44**, 133–149.

Mason, I., ., 1989: Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Magazine*, **37**, 75–81.

McTaggart-Cowan, R., 2009: dphase_cmcgeml: regional GEM model driving (low resolution) forecast run by CMC for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_cmcgeml. `http://dx.doi.org/10.1594/WDCC/dphase_cmcgeml` (abgerufen am 08.03.2012).

Mesinger, F., 2008: Bias adjusted precipitation threat scores. *Advances in Geosciences*, **16**, 137–142.

Milbrandt, J.A., M.K. Yau, J. Mailhot, und S. Bélair, 2008: Simulation of an orographic precipitation event during IMPROVE-2. Part I: Evaluation of the control run using a triple-moment bulk microphysics scheme. *Monthly Weather Review*, **136**, 3873–3893.

Murphy, A., und Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Murphy, A.H., 1991: Forecast verification: Its complexity and dimensionality. *Monthly Weather Review*, **119**, 1590–1601.

Murphy, A.H., 1996: The Finley affair: A signal event in the history of forecast verification. *Weather and Forecasting*, **11**, 3–20.

Peirce, C., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Reuter, H., M. Hantel, und R. Steinacker, 2001: *Bergmann-Schaefer: Lehrbuch für Experimentalphysik*, Band 7: Erde und Planeten, Meteorologie, S. 132–310. De Gruyter.

Rotach, M.W., P. Ambrosetti, C. Appenzeller, M. Arpagaus, L. Fontannaz, F. Fundel, U. Germann, A. Hering, M.A. Liniger, M. Stoll, A. Walser, F. Ament, H.S. Bauer, A. Behrendt, V. Wulfmeyer, F. Bouttier, Y. Seity, A. Buzzi, S. Davolio, M. Corazza, M. Denhard, M. Dorninger, T. Gorgas, J. Frick, C. Hegg, M. Zappa, C. Keil, H. Volkert, C. Marsigli, A. Montaini, R. McTaggart-Cowan, K. Mylne, R. Ranzi, E. Richard, A. Rossa, D. Santos-Muñoz, C. Schär, M. Staudinger, Y. Wang, und J. Werhahn, 2009: MAP D-PHASE: Real-time demonstration of weather forecast quality in the Alpine

region. *Bulletin of the American Meteorological Society*, **90**, 1321–1336.

Schraff, C., und R. Hess, 2011: A description of the nonhydrostatic regional COSMO-model, Part III: Data assimilation. Deutscher Wetterdienst, `http://www.cosmo-model.org` (abgerufen am 08.03.2012).

Schulz, J.P., und M. Denhard, 2009: dphase_lme: LME (COSMO-EU) model forecasts run by DWD for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_lme. `http://dx.doi.org/10.1594/WDCC/dphase_lme` (abgerufen am 08.03.2012).

Schulz, J.P., und U. Schättler, 2009: Kurze Beschreibung des Lokal-Modells Europa COSMO-EU (LME) und seiner Datenbanken auf dem Datenserver des DWD. DWD.

Seity, Y., 2009: dphase_arome: AROME +30hours forecasts run by Meteo-France CNRM/GAME for the MAP D-PHASE project. World Data Center for Climate. DOI:10.1594/WDCC/dphase_arome. `http://dx.doi.org/10.1594/WDCC/dphase_arome` (abgerufen am 08.03.2012).

Steinacker, R., C. Häberli, und W. Pöttschacher, 2000: A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data. *Monthly Weather Review*, **128**, 2303–2316.

Steinacker, R., D. Mayer, und A. Steiner, 2011: Data quality control based on self-consistency. *Monthly Weather Review*, **139**, 3974–3991.

Steinacker, R., M. Ratheiser, B. Bica, B. Chimani, M. Dorninger, W. Gepp, C. Lotteraner, S. Schneider, und S. Tschannett, 2006: A mesoscale data analysis and downscaling method over complex terrain. *Monthly Weather Review*, **134**, 2758–2771.

Stephenson, D., B. Casati, C.A.T. Ferro, und C.A. Wilson, 2008: The Extreme Dependency Score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications*, **15**, 41–50.

Stephenson, D.B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, **15**, 221–232.

Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, **106**, 7183–7192.

Tustison, B., D. Harris, und E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research*, **106**, 11775–11784.

Weusthoff, T., F. Ament, M. Arpagaus, und M.W. Rotach, 2010: Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Monthly Weather Review*, **138**, 3418–3433.

Weygandt, S.S., A.F. Loughe, S.G. Benjamin, und J.L. Mahoney, 2004: Scale

sensitivities in model precipitation skill scores during IHOP. In: *22nd Conf. on Severe Local Storms*, Hyannis, MA. American Meteorological Society.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences, International Geophysics Series*, Band 59. Academic Press.

Wulfmeyer, V., A. Behrendt, C. Kottmeier, U. Corsmeier, C. Barthlott, G.C. Craig, M. Hagen, D. Althausen, F. Aoshima, M. Arpagaus, H.S. Bauer, L. Bennett, A. Blyth, C. Brandau, C. Champollion, S. Crewell, G. Dick, P. Di Girolamo, M. Dorninger, Y. Dufournet, R. Eigenmann, R. Engelmann, C. Flamant, T. Foken, T. Gorgas, M. Grzeschik, J. Handwerker, C. Hauck, H. Höller, W. Junkermann, N. Kalthoff, C. Kiemle, S. Klink, M. König, L. Krauss, C.N. Long, F. Madonna, S. Mobbs, B. Neininger, S. Pal, G. Peters, G. Pigeon, E. Richard, M.W. Rotach, H. Russchenberg, T. Schwitalla, V. Smith, R. Steinacker, J. Trentmann, D.D. Turner, J. van Baelen, S. Vogt, H. Volkert, T. Weckwerth, H. Wernli, A. Wieser, und M. Wirth, 2011: The Convective and Orographically-induced Precipitation Study (COPS): the scientific strategy, the field phase, and research highlights. *Quarterly Journal of the Royal Meteorological Society*, **137**, 3–30.

Yessad, K., 2011: Basics about ARPEGE/IFS, ALADIN and AROME in the cycle 38 of ARPEGE/IFS. Météo-France, CNRM, GMAP.

Zappa, M., M.W. Rotach, M. Arpagaus, M. Dorninger, C. Hegg, A. Montani, R. Ranzi, F. Ament, U. Germann, G. Grossi, S. Jaun, A. Rossa, S. Vogt, A. Walser, J. Wehrhan, und C. Wunram, 2008: MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmospheric Science Letters*, **9**, 80–87.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AC	Anomaly Correlation
B	Bias
BR	Base Rate
COPS	Convective and Orographically induced Precipitation Study
D-PHASE	Demonstration of Probabilistic Hydrological and Atmospheric Simulation of flood Events in the Alpine region
DWD	Deutschen Wetterdienst
EDS	Extreme Dependency Score
ETS	Equitabe Threat Score
GTS	Global Telecommunication System
HSS	Heidke Skill Score
IMGW	Institut für Meteorologie und Geophysik der Universität Wien
JDC	Joint D-PHASE/COPS
LAM	Limited-Area Model
MAE	Mean Absolute Error
MESOCLIM	MESOscale Alpine CLIMatology
MSE	Mean Square Error
NWP	Numerical Weather Prediction
OR	Odds Ratio
ORSS	Odds Ratio Skill Score
PC	Percent Correct

POD	Probability Of Detection
POFD	Probability Of False Detection
PSS	Peirce Skill Score
QD	Quantile Difference
RMSE	Root Mean Square Error
ROC	Relative Operating Characteristics
RV	Reduction of Variance
SEDS	Symmetric Extreme Dependency Score
TS	Threat Score
VERA	Vienna Enhanced Resolution Analysis
WDCC	World Data Center for Climate
WWRP	World Weather Research Programme

Danksagung

Herrn Univ.-Prof. Dr. Reinhold Steinacker möchte ich für die Betreuung dieser Diplomarbeit danken.

Außerordentlicher Dank gilt Ass.-Prof. Mag. Dr. Manfred Dorninger, der mir die Mitarbeit im Projekt VERITA ermöglicht hat und in zahlreichen Besprechungen zusammen mit Mag. Theresa Gorgas und Mag. Stefan Kiesenhofer den Schlüsselimpuls sowie wertvolle Anregungen für die Erstellung dieser Arbeit gegeben hat. Auch meinen weiteren hilfsbereiten Kolleginnen und Kollegen am IMGW danke ich für den Austausch in technischen, fachlichen und bürokratischen Fragen.

Meine Korrekturleserinnen haben dankenswerterweise nicht nur die sprachliche Qualität dieser Arbeit verbessert, sondern mich auch stets mental unterstützt.

Nicht zuletzt möchte ich besonders meiner Familie danken, die mich immer zum Studium ermutigt und mir so den notwendigen Rückhalt gegeben hat.

Lebenslauf

Persönliche Daten

Name	Sarah Umdasch
Geburtsdatum	17.12.1985
Geburtsort	Linz
Nationalität	Österreich

Ausbildung

2005 – heute	**Universität Wien**, *Diplomstudium der Meteorologie und Geophysik*, erster Studienabschnitt abgeschlossen: Juli 2006, gewählter Studienzweig: Meteorologie.
2008	**Universidad Complutense de Madrid**, *ERASMUS-Auslandsaufenthalt*.
1996 – 2004	**BG/BRG Khevenhüller**, *Linz*, Ablegung der Reifeprüfung: Juni 2004.

Berufstätigkeit

03/2010 – heute	**Universität Wien**, *Wissenschaftliche Mitarbeiterin im FWF-Projekt: 'VERITA' (P20925, PI.: M. Dorninger).*
03/2010 – 02/2012	**Universität Wien**, *Tutorin für die Lehrveranstaltung Synoptisch-dynamische Meteorologie 1 und 2.*
07/2009	**Zentralanstalt für Meteorologie und Geodynamik**, *Praktikum - Abteilung Umweltmeteorologie, Wien.*
08/2008 – 09/2008	**MeteoServe Wetterdienst GmbH**, *Praktikum.*
09/2004 – 06/2005	**Au-Pair Aufenthalt**, *Paris.*

Konferenzbeiträge

Umdasch, S., S. Kiesenhofer, T. Gorgas und M. Dorninger, 2011, *An intercomparison of high resolution NWP models during MAP D-PHASE - Precipitation forecasts*, 31st International Conference on Alpine Meteorology (ICAM), Aviemore, Schottland, (Vortrag).

Umdasch, S., S. Kiesenhofer und M. Dorninger, 2011, *Vorhersagequalität hochauflösender NWP Modelle während MAP D-PHASE*, 4. Österreichischer MeteorologInnentag 2011, Klagenfurt, (Poster: ausgezeichnet mit dem ersten Posterpreis der Österreichischen Gesellschaft für Meteorologie).