7. Multiple lineare Regression

Die Kreuzspektrumanalyse hat gezeigt, daß zu prognostischen Zwecken verwendbare Phasenbeziehungen zwischen dem Druckverlauf an einem festen Ort und der Temperatur in Berlin über einen längeren Zeitraum wohl nicht existieren. Dennoch deuten die Ergebnisse z.T. auf das Vorhandensein zyklischer Zusammenhänge hin, die sich vom typischen „roten“ Verlauf unterscheiden. Daher werden im folgenden wieder Möglichkeiten untersucht, die weitere Entwicklung aus dem Zustand der atmosphärischen Zirkulation zu fixen Zeitpunkten vorherzusagen. Es wird also der „klassische“ Ansatz, daß ursprünglich nahe beieinander liegende Zustände zumindest eine Zeitlang einen vergleichbaren Verlauf nehmen, wieder aufgegriffen. Da die Autokorrelations-Spektralanalyse keine sinnvollen Alternativen aufzeigte, wurden diese Untersuchungen wieder in zeitlich starrem Rahmen durchgeführt, d.h. die Prognosen basieren auf monatlichen Mittelwerten der Einflußgrößen (Bodendruck und 500-hPa-Geopotential) und beziehen sich ebenfalls auf einzelne Kalendermonate. Dadurch ist die Vergleichbarkeit zu den früheren Arbeiten zur langfristigen Temperaturprognose für Berlin gewahrt.
In einem ersten Schritt soll das Vorhersagepotential der Gitterpunkte selbst – gemeint ist natürlich der monatliche Mittelwert des Luftdrucks am Ort des Gitterpunktes – bestimmt werden. Dies entspricht der Klärung der Frage, inwieweit bestimmte nordhemisphärische Regionen einen Einfluß auf die Folgetemperaturen von Berlin haben. Eine derartige Untersuchung führte *Dettmann* (2000) zwar bereits durch, jedoch weichen sowohl Methodik als auch die verwendeten Daten in wesentlichen Punkten voneinander ab. Die methodischen Unterschiede betreffen die Auswahl der Prediktoren und die Erstellung der finalen Prognosen: Existieren mehrere signifikante Gitterpunkte, so verwendet *Dettmann* unterschiedliche Mittelungsverfahren zur Kombination der einzelnen, durch einfache lineare Regression bestimmten Vorhersagen. In dieser Arbeit erfolgt die Berechnung dagegen mittels multipler linearer Regression. Um trotz der großen Zahl potentieller Prediktoren ein möglichst stabiles Modell entwickeln zu können, wurde zur Selektion der Prediktoren ein Screening-Verfahren benutzt.

7.1 Mathematisch-statistische Grundlagen

Das Modell der multiplen (auch mehrfachen) linearen Regression setzt einen oder mehrere Prediktoren $x_1,x_2,...,x_k$ in Beziehung zu einer einzigen Zielgröße y:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k \; . \tag{7.1}$$

Bei den Koeffizienten $a_1,a_2,...,a_k$ handelt es sich um die partiellen Regressionskoeffizienten zwischen y und x_i, die sich durch die jeweilige Ausschaltung der Einflüsse aller übrigen Prediktoren ergeben.
Für eine Stichprobe des Umfanges n liegt ein Gleichungssystem vor, welches n Wiederholungen der Gleichung 7.1 umfaßt – eine Gleichung für jeden Beobachtungstermin. Zur Berechnung der Regressionskoeffizienten $a_0,a_1,...,a_k$ existieren mehrere unterschiedliche Möglichkeiten. Das hier verwendete Verfahren ist *Taubenheim* (1969) entnommen:

Kennzeichnet man die bisher y genannte Zielgröße mit dem Index 0, so läßt sich folgende symmetrische Determinante B aufstellen:

$$\mathrm{B} = \begin{vmatrix} s_0^2 & s_{01} & s_{02} & \cdots & s_{0k} \\ s_{10} & s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{20} & s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{k0} & s_{k1} & s_{k2} & \cdots & s_k^2 \end{vmatrix} . \tag{7.2}$$

Bei den Werten entlang der Diagonalen handelt es sich um die Varianzen der K+1 Größen (K unabhängige und eine abhängige Variable), bei den übrigen um die jeweiligen Kovarianzen.
Die Koeffizienten a_j ($j \neq 0$) ergeben sich dann zu

$$a_j = -(-1)^j \frac{\mathrm{B}_{0j}}{\mathrm{B}_{00}}, \tag{7.3}$$

worin B_{0j} diejenige Unterdeterminante von B ist, die sich durch Streichung der 0-ten Zeile und der j-ten Spalte ergibt.
Die Konstante a_0, im Falle der einfachen linearen Regression der Achsenabschnitt, läßt sich wie folgt berechnen:

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - \ldots - a_k\bar{x}_k . \tag{7.4}$$

Wie im Falle der einfachen linearen Korrelation ist die Güte der Anpassung durch den Anteil der Varianz der Regressionswerte $\hat{y}$ an der Gesamtvarianz s_y^2 definiert:

$$r^2 = \frac{\frac{1}{n-1}\sum (\hat{y} - \bar{y})^2}{s_y{}^2} . \tag{7.5}$$

Die durch Gleichung 7.5 definierte Größe wird als mehrfache Bestimmtheit bezeichnet. Zieht man ihre Quadratwurzel, so erhält man den bekannten Korrelationskoeffizienten r. Er ist im Falle der multiplen Regression auf positive Werte beschränkt, da negative Werte keine sinnvolle Aussage enthalten.

Soll eine Prüfung der Signifikanz des Mehrfach-Korrelationskoeffizienten r mittels klassischen Testverfahrens durchgeführt werden, so ist der „F-Test“ geeignet: Überschreitet die Testgröße

$$F = \frac{n-1-K}{K} \frac{r^2}{1-r^2} \tag{7.6}$$

die in vielen Standardlehrbüchern der Statistik zu findenden Signifikanzschwellen, so ist r mit der entsprechenden Irrtumswahrscheinlichkeit signifikant von Null verschieden. Als Freiheitsgrade sind $f_1 = K$ sowie $f_2 = n-1-K$ zu verwenden.

7.2 Screening-Verfahren

Gilt es, ein auf multipler linearer Regression beruhendes Vorhersagemodell zu entwickeln, für das eine große Zahl potentieller Prediktoren zur Verfügung steht, so muß eine Auswahl getroffen werden, welche Prediktoren tatsächlich verwendet werden. Selbst im Idealfall, daß ein Teil der potentiellen Prediktoren von physikalischer Bedeutung ist, während ein derartiger Zusammenhang bei anderen nicht gegeben ist, ist es normalerweise nicht sinnvoll, alle physikalisch relevanten Prediktoren mit einzubeziehen. Grund hierfür ist der zumeist vorhandene Zusammenhang der Einflußgrößen untereinander, der zu Redundanzen führen würde. Ist für keinen der potentiellen Prediktoren ein physikalischer Zusammenhang auszumachen, so wird die Wahl der richtigen Prediktoren weiter erschwert.
Zur Lösung dieses Problems werden sogenannte schrittweise Verfahren verwendet, die in der englischsprachigen Literatur als „screening regression“ bezeichnet werden. Es handelt sich im Prinzip um Filteralgorithmen, die durch Hinzunahme bzw. Weglassen der einzelnen unabhängigen Variablen x_i das bestmögliche Prediktoren-Kollektiv bestimmen. Im Vergleich zu anderen Verfahren sind sie „leicht verständlich, benötigen wenig Rechenoperationen und bieten dem Benutzer im Laufe der Rechnung jederzeit die Möglichkeit, interaktiv einzugreifen“ (*Fahrmeir und Hamerle*, 1984). Das am häufigsten verwendete Screening-Verfahren ist als „Vorwärtsselektion“ („forward selection“ oder „stepwise regression“) bekannt:
Es existiere ein Pool L potentieller Prediktoren. Man geht zunächst von einem Modell aus, das nur die Konstante a_0 enthält: $y = a_0$. Im ersten Schritt wird für jeden der L potentiellen Prediktoren der einfache lineare Zusammenhang mit dem Prediktanden y bestimmt. Derjenige potentielle Prediktor, welcher die bestmögliche lineare Regression liefert, wird als x_1 ausgewählt, so daß das Vorhersagemodell nun $y = a_0 + a_1x_1$ lautet. In der zweiten Stufe werden neue Regressionsmodelle für die verbliebenen L-1 potentiellen Prediktoren berechnet, die jedoch alle das in Folge des ersten Schrittes ausgewählte x_1 enthalten ($y = a_0 + a_1x_1 + a_ix_i$). Ausgewählt wird wiederum derjenige potentielle Prediktor, der in Verbindung mit x_1 die besten Resultate erzielt. Diese Prozedur wird bei jeder nachfolgenden Stufe wiederholt, so daß sich die Zahl der Prediktoren mit jedem Schritt um einen erhöht. Die Regressionskoeffizienten verändern sich dabei in der Regel ebenfalls, da die Prediktoren fast immer untereinander korreliert sind (vergl. Abschnitt 7.1).

Eine Alternative bietet u.a. die „Rückwärtselimination“ („backward elimination“). Sie geht vom vollständigen Modell mit L Prediktoren aus und eliminiert den jeweils unbedeutendsten. Dabei ist keineswegs garantiert, daß letztlich dieselben Prediktoren ausgewählt werden wie bei der Vorwärtsselektion. Beide Verfahren sind lediglich als Werkzeug zur Entwicklung eines zufriedenstellenden Vorhersagemodells im Sinne einer „black box“ geeignet, keineswegs aber zur Ermittlung derjenigen Variablen, die in einem physikalischen Zusammenhang mit dem Prediktanden stehen.

Zur Beantwortung der Frage, welcher der beste Prediktor ist bzw. wann die Prozedur abzubrechen ist, wird im Rahmen der klassischen Statistik zumeist die Testgröße F (Gl. 7.6) herangezogen. In jeder Stufe des Verfahrens wird stets derjenige Prediktor ausgewählt, dessen Miteinbeziehung das maximale F, d.h. den höchsten Mehrfach-Korrelationskoeffizienten, zur Folge hat. Das Abbruchkriterium ist dann erfüllt, wenn eine vorher festgelegte Signifikanz der Regression durch das Hinzufügen eines weiteren Prediktors mit keinem der potentiellen Prediktoren mehr erreicht wird.
Es hat sich jedoch gezeigt, daß die Verwendung der Testgröße F die Gefahr des Overfitting in sich birgt, wenn eine Vielzahl potentieller Prediktoren zur Verfügung steht (vergl. Abschnitt 4.1). Der Grund hierfür ist, daß die Prediktoren in den einzelnen Stufen der Regression nicht zufällig selektiert werden, sondern daß der jeweils beste genommen wird (vergl. z.B. *Wilks*, 1995). Dadurch ist selbst eine vollständig auf Zufällen beruhende Anpassung nicht auszuschließen. Um dieser Gefahr entgegenzuwirken, empfiehlt es sich, mit Cross-Validation zu arbeiten, wobei zwei gängige Varianten existieren: Erstens die Abschätzung der Prognosengüte des vorab festgelegten Modells, wobei das Kollektiv der Prediktoren während der gesamten Prozedur unverändert bleibt (vergl. Abschnitt 4.1). Alternativ kann die Cross-Validation auch direkt zur Variablenselektion verwendet werden. Dann wird die Testgröße F durch ein Fehlermaß wie z.B. den MSE ersetzt, welches jeweils am künstlichen Verifikationskollektiv bestimmt wird (*Michaelsen*, 1987). Gleichzeitig fungiert das Fehlermaß auch als Abbruchkriterium: Wird durch das Hinzufügen eines weiteren Prediktors keine Verringerung des Vorhersagefehlers mehr erreicht, so ist der Screening-Vorgang abzubrechen. Um die Schärfe des Abbruchkriteriums zu steigern, kann zusätzlich ein Mindestbetrag eingeführt werden, um den sich die Prognosen verbessern müssen.
Ein Nachteil der Variablenselektion mittels Cross-Validation ist, daß diese Methode keine unabhängige Schätzung der Vorhersageleistung mehr liefert. Soll die Prognosengüte dennoch vorab geschätzt werden, so ist eine „double Cross-Validation“ durchzuführen (vergl. z.B. *Mosteller und Tukey*, 1977). Hierbei handelt es sich um die Verschachtelung zweier Cross-Validation Prozeduren. Da das Screening-Verfahren für jedes künstliche Verifikationskollektiv (an Hand dessen die Prognosengüte abgeschätzt wird) neu durchlaufen wird, können verschiedene Prediktoren selektiert werden. Im Extremfall ist es möglich, daß das Kollektiv in allen Fällen unterschiedlich zusammengesetzt ist. Nach dem Kriterium der Unverzerrtheit würde dies jedoch darauf hindeuten, daß das Modell instabil ist.

7.3 Vorgehensweise

Um das Risiko instabiler Modelle weitestgehend zu minimieren, sollte im Rahmen der multiplen linearen Regression unbedingt ein Examinationskollektiv verwendet werden. Wie bereits in Abschnitt 4.1 erwähnt, ist dies die sicherste Methode, Scheingüte

auszuschließen. Zu diesem Zweck wurde ein ausreichend langer Datensatz benötigt, wodurch auf die Verwendung des 500-hPa-Geopotentials verzichtet werden mußte. Die monatlichen Mittelwerte des Bodendrucks des Datensatzes „ds010.1“ reichen hingegen über den Beginn der Dahlemer Klimareihe hinaus zurück. Somit stand der insgesamt 90 Jahre umfassende Zeitraum 1909 bis 1998 zur Verfügung, wobei jedoch nur bei einem Teil aller Gitterpunkte der Nordhemisphäre im gesamten Zeitraum für jeden Monat Werte vorhanden sind. Daher konnten lediglich 389 Punkte verwendet werden, d.h. das Gitternetz weist z.T. beträchtliche räumliche Lücken auf (siehe dazu auch Kapitel 2).
Die Aufteilung des Untersuchungszeitraumes in ein Entwicklungs- und ein Examinationskollektiv geschah im Verhältnis 2 zu 1. An Hand der Jahre 1909 bis 1968 wurden die Modelle entwickelt, zur Beurteilung der Prognosenleistung dienten die Jahre 1969 bis 1998. Die Berechnung der Anomalien sowie die Trendbereinigung (siehe dazu auch Abschnitt 4.2) wurden vorab der Teilung, d.h. nur einmal für den vollständigen Datensatz, durchgeführt. Da real existierende Klimatrends praktisch nie exakt linear verlaufen, ist der Mittelwert der einzelnen Kollektive in den meisten Kalendermonaten somit von Null verschieden. Dies hat einen Bias der Referenzprognose zur Folge, da für die Klimavorhersage Null, d.h. keine Abweichung, verwendet wurde. Zusätzliche Konsequenz ist eine Abnormität des Erwartungswertes des mittleren Prognosefehlers (siehe dazu auch Abschnitt 7.5).
Neben der Temperatur wurde auch der Niederschlag als Prediktand getestet. Aufgrund der im 3. Kapitel vorgestellten Ergebnisse wurde statt der gesamten monatlichen Niederschlagsmenge die Zahl der Tage mit meßbarem Niederschlag verwendet.
Die Berechnung von Anomalien sowie die Trendbereinigung führten dazu, daß die Zahl der Tage mit meßbarem Niederschlag nicht mehr in Form von ganzen Zahlen vorlag. Um die Regressionen möglichst exakt durchführen zu können, wurde eine Dezimalstelle beibehalten, d.h. es wurde die bei der Temperatur übliche Genauigkeit gewählt. Auch die Berechnungen der Vorhersageleistung (RV-Werte) basieren auf derartig gerundeten Zahlen. Auf diese Weise sollte die Vergleichbarkeit zur Klimaprognose – hier ist die Angabe einer Nachkommastelle üblich – gewahrt werden.

Die Entwicklung der einzelnen Modelle wurde mittels Vorwärtsselektion durchgeführt. Zur Variablenselektion wurde die U-Methode (vergl. Abschnitt 4.1) eingesetzt, wobei der MAE als Fehlermaß diente. Er wurde anstelle des üblichen MSE verwendet, um große Vorhersagefehler nicht überproportional zu bestrafen. Es sollte vermieden werden, daß gelegentlich auftretende grobe Fehlprognosen nicht durch eine Vielzahl von Vorhersagen auszugleichen sind, die den Prognosefehler im Vergleich zur Klimareferenz leicht verringern. Abgebrochen wurde das Screening-Verfahren, wenn der MAE durch die Hinzunahme eines weiteren Prediktors nicht mindestens um 0,1°C (Niederschlag 0,15 Tage) reduziert wurde. Aufgrund des zur Verfügung stehenden Examinationskollektivs konnte auf die Durchführung einer „double Cross-Validation“ verzichtet werden.
Um Redundanzen zu vermeiden, wurde das angewendete Screening-Verfahren um einen zusätzlichen Schritt erweitert. Am Ende jeder Stufe, nachdem jeweils ein weiterer potentieller Prediktor dem Prediktoren-Kollektiv hinzugefügt wurde, wird der Pool der potentiellen Prediktoren reduziert. Sämtliche Gitterpunkte, die mit dem soeben ausgewählten signifikant korreliert sind, werden vom Verfahren ausgeschlossen. Dadurch wird erreicht, daß bereits bekannte Informationen nicht nochmals berücksichtigt werden können. Konkret wurden all diejenigen Gitterpunkte eliminiert, deren Korrelationskoeffizient die 95%ige Signifikanzschranke des „t-Tests“ (siehe z.B. *Taubenheim*, 1969) überschritt.

Basierend auf der mittleren monatlichen Bodendruckverteilung des jeweils vorangegangenen Kalendermonats (im Folgenden als Vormonat bzw. Basismonat bezeichnet) wurden die Regressionsmodelle für die elf folgenden Kalendermonate (Zielmonate) unabhängig voneinander entwickelt. Der Jahreszyklus wird dabei einmal durchlaufen, so daß jeder Kalendermonat genau einmal als Vormonat fungiert. Dadurch kommen 132 unterschiedliche Modelle zustande (12 Vormonate mal 11 Zielmonate). So dienen z.B. die Monatsmittel der Gitterpunktwerte des März als potentielle Prediktoren für elf unabhängige Regressionen, je eine für die Monate April des laufenden bis Februar des Folgejahres. Zusätzlich soll folgende Terminologie verwendet werden: Stammt eine Prognose aus dem unmittelbar vorangegangenen Monat (z.B. Juni vor Juli), so wird sie als 1-Monats-Prognose bezeichnet. Entsprechend würde eine Vorhersage der Novembertemperatur aus der Bodendruckverteilung des Februar als 9-Monats-Prognose bezeichnet werden. Diese Definition unterscheidet sich von der in der Literatur üblichen. Hier gilt zumeist nur jene Zeit als lead time, die zwischen dem Ende des Basiszeitraumes und dem Beginn des Vorhersagezeitraumes vergeht. Eine Prognose der Julitemperatur basierend auf Mittelwerten des Juni würde daher in der Regel mit der lead time 0 Monate gekennzeichnet.
Teilweise, bevorzugt in Tabellen, soll der Vormonat auch als t_0 bezeichnet werden. Der erste, unmittelbar folgende Zielmonat bzw. die 1-Monats-Prognose als t+1 usw. Wird der Pool der potentiellen Prediktoren noch durch Gitterpunktwerte ergänzt, die aus t_0 vorangehenden Monaten stammen (in diesem Kapitel nicht der Fall), so werden diese mit einem Minus gekennzeichnet (t-1, t-2 usw.).

7.4 Vorselektion der potentiellen Prediktoren

Um das bestehende Mißverhältnis zwischen der Anzahl der potentiellen Prediktoren und der Objektanzahl (389 Gitterpunkte bei nur 60 Jahren Entwicklungszeitraum) zu verringern, sollte eine Vorauswahl durchgeführt werden. Da die Entscheidung darüber, welche Gitterpunkte ausscheiden sollten, nicht an Hand der physikalischen Gesetze der Atmosphäre getroffen werden konnte, blieb nur die Möglichkeit, sich auf statistische Methoden zu verlassen. Dabei sollte die alleinige Verwendung klassischer Signifikanztests als Auswahlkriterium vermieden werden, um das Problem der „multiplicity“ berücksichtigen zu können (vergl. dazu Abschnitt 4.1). Eine geeignete Möglichkeit hierfür ist durch die Verwendung der „Binomialverteilung“ gegeben. Sie gibt die Wahrscheinlichkeit für die Häufigkeit x eines Ereignisses bei N unabhängigen Versuchen an, dessen Einzelwahrscheinlichkeit p ist:

$$W_N(x) = \binom{N}{x} p^x (1-p)^{N-x} . \tag{7.7}$$

Im Kontext multipler klassischer Signifikanztests können x als die Anzahl individueller Tests mit signifikantem Ausgang, p als Irrtumswahrscheinlichkeit und N als die Gesamtzahl der durchgeführten Tests (entsprechend der Anzahl der potentiellen Prediktoren) angesehen werden. Führte man die 389 Einzeltests mit einer Irrtumswahrscheinlichkeit von z.B. 1% durch, so läge die Wahrscheinlichkeit dafür, daß die Signifikanzschranke mindestens einmal überschritten wird, bei 98% [$W_N(0) = 0{,}02$].

Um Zufälligkeiten ausschließen zu können, wird nun aber gefordert, daß die absolute Häufigkeit des Überschreitens der Signifikanzschranke unabhängig von der Anzahl der potentiellen Prediktoren konstant bleibt. Um diese Bedingung zu realisieren, muß die Irrtumswahrscheinlichkeit p der Einzeltests modifiziert werden. Hierfür ist $W_N(0)$ gleich der gewünschten statistischen Sicherheit S zu setzen, da die Signifikanzschwelle – wie im Falle des einzelnen Tests – mit S prozentiger Wahrscheinlichkeit (überhaupt) nicht überschritten werden soll. Das daraus resultierende p entspricht dann der Irrtumswahrscheinlichkeit, mit welcher die Einzeltests durchzuführen sind. Im vorliegenden Fall ergibt sich nach Gleichung 7.7 für S = 95% der Wert p = 0,0001318, für S = 99% gar der Wert p = 0,0000258.
Als Kriterium zur Vorselektion wurde der einfache lineare Korrelationskoeffizient r verwendet. Zur Prüfung der Signifikanzen kann dann z.B. der „t-Test“ benutzt werden. Für vorgegebenes S können die Signifikanzschwellen auch direkt für den Korrelationskoeffizienten berechnet werden. Es gilt:

$$\left| r_{sig} \right| = \left[\frac{(n-2)}{t^2} + 1 \right]^{-0,5} . \tag{7.8}$$

Die hierfür benötigten t-Werte können Tabellen entnommen bzw. durch Computeralgorithmen berechnet werden. Sie sind nichts anderes als die exakten Signifikanzschranken des „t-Tests“ bei vorgegebener Irrtumswahrscheinlichkeit. Mit n=60 erhält man für die Signifikanzschwellen des Korrelationskoeffizienten für S=95% den Wert r=0,474 und für S=99% den Wert r=0,515. Diese Resultate stimmen mit einer von *Enke* (1988) vorgestellten Approximation sehr gut überein.
Eine weniger stringente Alternative stellt die Forderung dar, daß die (theoretische) relative Häufigkeit des Überschreitens der Signifikanzschranke unverändert bleibt: Führte man 389 Einzeltests mit einer Irrtumswahrscheinlichkeit von 1% durch, so würde theoretisch in etwa vier Fällen ein signifikantes Testergebnis vorgetäuscht. Nach Gleichung 7.7 beträgt die tatsächliche Wahrscheinlichkeit, daß dieses Ereignis maximal viermal eintritt, jedoch lediglich 65%, da gilt:

$$\sum_{i=0}^{4} W_N(i) = 0,6504 . \tag{7.9}$$

Entsprechend der Forderung soll nun aber eine 99%ige Wahrscheinlichkeit gegeben sein, d.h. die rechte Seite der Gleichung 7.9 ist durch 0,99 zu ersetzen. Diese Bedingung ist für eine Irrtumswahrscheinlichkeit der Einzeltests von p ≈ 0,00325 erfüllt.
Aufgrund der Vielzahl sowie der hohen Interkorrelationen der zur Verfügung stehenden potentiellen Prediktoren wurde die Gefahr, dem Zufall aufzusitzen, höher eingeschätzt als jene, ungewollt bedeutsame Informationen zu eliminieren. Daher wurde zur Vorauswahl das schärfere Kriterium verwendet.

7.5 Ergebnisse

In den Tabellen 7.1 und 7.2 ist für sämtliche Kombinationen der Vor- und Zielmonate die Anzahl der Gitterpunkte aufgeführt, welche die zur Vorselektion verwendeten Kriterien (Abschnitt 7.4) erfüllen. Für beide untersuchten Zielgrößen weist die Anzahl der potentiellen Prediktoren, welche die Signifikanzschranken überschreiten, trotz der insgesamt 132 Realisierungen auf Überzufälligkeit hin. Hierbei ist jedoch der Einfluß der Interkorrelationen meteorologischer Feldverteilungen noch nicht berücksichtigt. Eine Betrachtung der geographischen Verteilung der signifikanten Gitterpunkte ergibt, daß diese ausnahmslos benachbart sind, wenn sich bei einer bestimmten Kombination von Vor- und Zielmonat gleich mehrere als signifikant erweisen. So sind z.B. im Falle der 9-Monats-Prognose des Niederschlags basierend auf der Druckverteilung des Dezembers die Gitterpunkte 70N/75W, 70N/70W sowie 70N/65W betroffen. Nun ist es aber vorstellbar, daß eine vorhandene Korrelation zwischen den Daten zweier Örtlichkeiten zu einer Steigerung der Wahrscheinlichkeit einer irrtümlichen Ablehnung der Nullhypothese am einen Ort führt, wenn ein solcher Fehler am anderen Ort bereits begangen worden ist. Dies hat zur Folge, daß derartige Fehler erster Art innerhalb einer Feldverteilung in der Regel räumlich gehäuft auftreten (*Wilks*, 1995). In Anbetracht dieser Tatsache ist es sinnvoll, zur Beantwortung der Frage nach der Überzufälligkeit der Testergebnisse lediglich die Anzahl der Monatskombinationen, für welche überhaupt signifikante Gitterpunkte ermittelt wurden, zu Rate zu ziehen: Im Falle der Temperatur als Prediktand werden die Signifikanzschwellen insgesamt sechs- bzw. einmal, beim Niederschlag vier- bzw. zweimal überschritten. Diese Häufigkeiten liegen stets in der Größenordnung der theoretischen Erwartungen. Inwieweit ein Risiko besteht, daß die Modellbildung auf Zufälligkeiten basiert, ist daher an Hand der Ergebnisse der Vorselektion nicht endgültig zu entschieden. In den Fällen, in denen kein potentieller Prediktor die Kriterien der Vorselektion erfüllt, ist es als erhöht zu bezeichnen. Andererseits ist auf Grund der Schärfe der Kriterien nicht auszuschließen, daß hier relevante Informationen verloren gegangen sind.

t_0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11
Januar	-	-	-	-	-	-	-	-	-	-	-
Februar	5/2	2/0	-	-	-	-	-	-	-	-	-
März	-	-	-	-	-	-	-	-	-	-	-
April	-	-	-	-	-	-	-	-	-	-	-
Mai	-	-	-	-	-	-	-	1/0	-	-	-
Juni	-	-	-	-	1/0	-	-	-	-	-	-
Juli	-	3/0	-	-	-	-	-	-	-	-	-
August	-	-	-	-	-	-	-	-	-	-	-
September	-	-	-	-	-	-	-	-	-	-	-
Oktober	-	-	-	-	-	-	-	1/0	-	-	-
November	-	-	-	-	-	-	-	-	-	-	-
Dezember	-	-	-	-	-	-	-	-	-	-	-

Tabelle 7.1: Anzahl der Gitterpunkte, welche die an die Anzahl der potentiellen Prediktoren angepaßten Signifikanzschwellen überschreiten. Dabei bezieht sich der erste Wert auf 95%ige, der zweite Wert auf 99%ige statistische Sicherheit. Prediktand ist die Temperatur.

t_0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11
Januar	-	-	-	-	-	-	-	-	-	-	-
Februar	-	-	-	-	-	-	-	-	-	-	-
März	-	-	-	3/0	-	-	-	-	-	-	-
April	-	-	-	-	-	-	-	-	-	-	-
Mai	-	-	-	-	-	12/7	-	-	-	-	-
Juni	-	-	-	-	-	-	-	-	-	-	-
Juli	-	-	-	-	-	-	-	-	1/1	-	-
August	-	-	-	-	-	-	-	-	-	-	-
September	-	-	-	-	-	-	-	-	-	-	-
Oktober	-	-	-	-	-	-	-	-	-	-	-
November	-	-	-	-	-	-	-	-	-	-	-
Dezember	-	-	-	-	-	-	-	-	3/0	-	-

Tabelle 7.2: wie Tabelle 7.1, jedoch mit dem Niederschlag als Prediktand.

Als Konsequenz der Benachbarung und der damit verbundenen hohen Korrelation der vorläufig ausgewählten Gitterpunkte mußte gänzlich auf die Vorselektion verzichtet werden. Auf Grund des eingeführten Zwischenschrittes zur Vermeidung von Redundanzen (vergl. Abschnitt 7.3) wäre die Modellbildung zumeist nach der ersten Stufe abgebrochen worden, da die übrigen potentiellen Prediktoren aussortiert worden wären. Wie sich herausstellte, wurde jeweils einer dieser Gitterpunkte von den Regressionsmodellen dennoch berücksichtigt. Mit nur einer Ausnahme (beim Niederschlag März vor Juli wurde keiner der drei Punkte überhaupt ausgewählt) wurden sie in der ersten Stufe selektiert.

In der jeweils oberen Zeile der Tabelle 7.3 sind die RV-Werte der Temperaturprognosen im Examinationskollektiv für die einzelnen Monatskombinationen angegeben. Obwohl diese Werte nach dem „F-Test“ zum Vergleich der Varianzen zweier Stichproben (siehe z.B. *Taubenheim*, 1969) nicht statistisch signifikant von Null verschieden sind (für eine Irrtumswahrscheinlichkeit von 5% liegen die Signifikanzschranken bei RV ≈ 0,45 bzw. RV ≈ -0,8), geben sie Hinweise zur Stabilität der Regressionsmodelle. In der überwiegenden Mehrheit der Fälle konnte die Vorhersageleistung gegenüber der Klimareferenzprognose nicht gesteigert werden. Häufig liegt der RV-Wert deutlich unter Null, teilweise wird die Signifikanzschranke fast erreicht (z.B. im Falle der 9-Monats-Prognose aus dem Juli). In insgesamt 23 Fällen werden aber auch positive RV-Werte erzielt, wobei der Höchstwert bei 20% (September vor Dezember) liegt. Die entsprechenden Monatskombinationen sind in Tabelle 7.3 durch Kursivdruck gekennzeichnet. Es stellt sich daher die Frage, ob es sich hierbei um reale Vorhersagbarkeiten oder lediglich um Zufälle handelt. Zu ihrer Klärung kann möglicherweise die Theorie einen Teil beitragen: Ein stabiles Vorhersagemodell liegt mit hoher Wahrscheinlichkeit dann vor, wenn sowohl der durch Cross-Validation abgeschätzte RV-Wert im Entwicklungskollektiv als auch der RV-Wert im Examinationskollektiv positiv sind. Auf Grund des Abfalls der Vorhersageleistung am unabhängigen Kollektiv ist der Wert im Entwicklungszeitraum dabei in der Regel höher. Da hier auf eine „double Cross-Validation“ verzichtet wurde, ist gar eine deutliche Differenz beider Werte zu erwarten. Tatsächlich ist jedoch festzustellen, daß der RV-Wert im Entwicklungskollektiv zumeist kleiner ausfiel, wenn er im Examinationskollektiv größer Null ist. In den übrigen Monatskombinationen kommt dies praktisch nicht vor. In diesen Fällen ist die Wahrscheinlichkeit zufällig guter

Übereinstimmungen daher hoch. Die Tatsache, daß die Modellbildung sehr häufig schon nach der ersten Stufe abgebrochen wurde und nie mehr als zwei Prediktoren zur Anwendung kamen (siehe Tabelle 7.4), bekräftigt dies zusätzlich. Entsprechen die Ergebnisse den Erwartungen (fett gedruckte Werte der Tabelle 7.3), so liegt die Reduktion der Varianz unter 5%. Einzige Ausnahme sind die 1-Monats-Prognosen des März. Diesen Ergebnissen zufolge könnte eine erfolgreiche Vorhersage an Hand der Druckverteilung im Februar möglich sein.

t_0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11
Januar	*0,01* *-0,02*	*0,04* *-0,02*	-0,23 0,11	-0,57 0,10	-0,36 0,11	*0,03* *0,01*	-0,09 0,03	-0,51 0,01	-0,13 0,01	-0,28 0,09	-0,26 0,06
Februar	***0,09*** ***0,22***	-0,02 0,18	-0,19 -0,01	-0,16 0,02	-0,06 0,05	-0,17 0,03	-0,22 0,06	-0,11 0,02	*0,06* *0,02*	-0,08 0,04	-0,20 0,09
März	-0,27 -0,01	*0,02* *-0,01*	-0,04 -0,02	-0,07 -0,04	-0,39 0,03	*0,05* *0,01*	-0,14 0,00	-0,11 0,00	-0,25 0,03	-0,03 0,00	-0,08 -0,02
April	-0,21 0,07	-0,21 -0,01	***0,02*** ***0,03***	-0,05 0,23	-0,08 0,06	***0,04*** ***0,07***	-0,16 0,13	0,00 0,01	-0,06 0,13	-0,06 -0,02	-0,22 0,06
Mai	-0,41 0,14	-0,02 0,08	-0,16 0,06	-0,04 -0,01	-0,27 0,08	-0,28 0,04	-0,01 0,05	***0,01*** ***0,17***	0,00 -0,01	-0,02 0,21	-0,11 0,01
Juni	*0,10* *0,04*	-0,03 0,03	-0,45 0,00	-0,21 -0,02	-0,20 0,17	-0,14 0,12	-0,01 0,09	0,00 -0,06	-0,05 0,11	-0,18 0,11	-0,21 0,06
Juli	*0,15* *0,08*	0,00 0,18	-0,18 0,00	-0,15 0,08	-0,05 0,18	-0,13 -0,02	0,00 -0,04	-0,15 0,15	-0,68 0,08	-0,03 0,02	-0,03 0,01
August	-0,07 -0,01	-0,24 0,04	*0,15* *0,08*	-0,24 0,08	-0,14 -0,01	-0,09 -0,01	-0,39 0,10	-0,29 0,07	-0,10 0,03	-0,17 0,05	*0,04* *-0,02*
September	-0,16 0,01	-0,15 -0,02	*0,20* *0,11*	-0,19 -0,03	*0,16* *0,00*	*0,08* *-0,01*	-0,05 0,03	-0,34 0,23	-0,53 0,11	***0,01*** ***0,02***	*0,11* *0,04*
Oktober	-0,03 0,10	-0,19 0,09	-0,24 0,15	*0,03* *-0,03*	-0,09 0,21	-0,62 0,19	0,00 0,05	-0,37 0,19	-0,08 0,12	-0,20 0,00	-0,36 0,10
November	-0,02 0,06	-0,15 0,08	-0,06 -0,02	*0,05* *0,00*	-0,20 0,00	-0,09 0,02	-0,14 0,03	-0,14 -0,01	-0,27 0,01	-0,09 -0,03	-0,12 0,12
Dezember	-0,02 0,07	-0,34 0,00	*0,05* *0,03*	-0,45 0,07	*0,03* *0,00*	-0,11 0,07	-0,21 0,07	-0,05 -0,01	-0,16 0,00	-0,25 0,04	-0,16 0,04

Tabelle 7.3: RV-Werte der Temperaturprognosen im Examinations- und Entwicklungskollektiv (untere Zeile). Positive Werte im Examinationskollektiv sind kursiv gedruckt. Ist zusätzlich der RV-Wert im Entwicklungskollektiv entsprechend der theoretischen Erwartung höher, so ist dies durch Fettdruck gekennzeichnet.

Von fast größerem Interesse als die Ergebnisse der einzelnen Monatskombinationen ist das Abschneiden der Methodik im dauerhaften operationellen Einsatz. Hier gilt es zu entscheiden, ob stets die aktuellste oder die „beste“ Prognose verwendet werden soll. Letztere Verfahrensweise hat den Nachteil, daß die Vorhersage ggf. erst unmittelbar vor Beginn des Vorhersagezeitraumes zur Verfügung steht, wie im Falle des März, dessen Prognose auf Basis der Februarzirkulation mit 0,22 den höchsten RV-Wert im Entwicklungskollektiv aufweist (vergl. den jeweils unteren Wert der bei t_0=Februar/t+1 beginnenden und bei t_0=Januar/t+2 abbrechenden Diagonale der Tabelle 7.3, die bei t_0=Dezember/t+3 fortzuführen ist). Andererseits bedingt die stete Benutzung der aktuellsten Vorhersage u.U. einen deutlichen Informationsverlust, wenn die einzelnen Modelle stabil sind. Denkbar ist auch die Kombination beider Möglichkeiten, indem z.B. solange eine aktuelle Prognose verwendet wird, bis der Zeitpunkt der „Besten“ erreicht ist.

t_0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11
1	40/60W	30/130E	30/20E	25/100W	40/5W	40/95W	40/150W	60/40E	30/5W	70/85W	20/85W
2	40/20E	40/20E	30/105E	45/10W	40/125E	30/85W	60/5W	20/80W	25/115E	70/40E 30/175W	50/170E
3	70/0	70/80W	60/165E	25/70W	30/40E	65/175E	40/20E	30/160W	50/120W	65/35E	35/5W
4	55/40W	35/35E	55/35E	40/10E 30/20W	70/40E	40/15W	30/20W	55/95W	35/5E 35/120W	30/160E	35/35E
5	60/5E	70/55W	35/5E	55/110W	55/105W	35/140E	30/115E	50/15E	25/115E	35/150E 60/125W	50/105W
6	40/170E	45/95W	30/80E	25/90W	45/85W	65/0	30/40E 55/20W	45/95W	70/40E	45/75W	40/155W
7	60/10E	40/160W	55/30E	55/135W	30/105E 40/150E	50/95W	70/10E	35/165E 60/35W	30/70E	70/40E	30/115W
8	40/35W	40/85W	65/35E	30/110W	55/85W	35/35W	30/115W 35/70W	35/40W	30/105E	60/65W	35/35E
9	45/10W	30/180	35/70W	35/160E	35/80W	50/140E	70/60W	30/105E 50/180	30/40E	55/45W	30/5W
10	40/35W	30/175W 30/95W	60/65W 45/120W	60/60W	70/20W 60/165E	30/70E 30/105E	30/30E	65/55W	30/35W	60/175E	30/120E
11	45/85W	50/140W 35/150E	50/165W	50/20W	40/160W	30/165E	50/5E	30/140E	70/155W	25/70W	25/50W
12	70/160W 25/65W	60/125W	70/170W	65/125W	35/170W	35/10E	40/155E	35/110W	55/15W	25/80W	70/20E

Tabelle 7.4: Von den Regressionsmodellen zur Vorhersage der monatlichen Mitteltemperatur selektierte Gitterpunkte der nördlichen Halbkugel in der jeweiligen Reihenfolge der Auswahl.

Vorlaufzeit	t+1	t+2	t+3	t+4	t+5	t+6	t+x
Kalenderjahr	-0,03	-0,14	-0,07	-0,14	-0,08	-0,10	-0,14
SS_{MSE}	-0,03	-0,03	-0,08	-0,03	-0,04	-0,07	0,02
Kalenderjahr	-0,03	-0,06	-0,04	-0,07	-0,05	-0,04	-0,05
SS_{MAE}							
Kalenderjahr	-0,05	-0,11	-0,08	-0,16	-0,13	-0,10	-0,14
SS_{MSEstd}	-0,04	-0,04	-0,06	-0,04	-0,04	-0,06	-0,01
Frühling	-0,06	0,02	-0,07	-0,22	-0,08	-0,11	-0,16
SS_{MSE}	-0,13	-0,09	-0,05	-0,08	-0,17	-0,22	-0,20
Sommer	-0,02	-0,07	-0,04	-0,09	-0,21	-0,05	-0,15
SS_{MSE}	-0,05	-0,06	-0,14	-0,02	-0,03	-0,06	0,07
Herbst	-0,08	-0,13	-0,11	-0,14	-0,19	-0,09	-0,12
SS_{MSE}	-0,23	-0,25	-0,24	-0,21	-0,22	-0,19	-0,13
Winter	-0,01	-0,22	-0,07	-0,13	-0,01	-0,12	-0,12
SS_{MSE}	0,29	0,30	0,11	0,18	0,27	0,21	0,33

Tabelle 7.5: Skill score der Temperaturprognosen mittels multipler linearer Regression in Abhängigkeit von der Jahres- und der Vorlaufzeit. Für das Kalenderjahr wurde die SS auf Basis des MSE (RV-Wert), des MSE bei standardisierten Vorhersagen und Beobachtungen sowie des MAE berechnet. Die unteren Werte geben den jeweiligen Bias der Prognosen an.

	Jan	Feb	März	April	Mai	Juni	Juli	Aug	Sep	Okt	Nov	Dez
EW	-----	-0,08	-0,06	0,15	-0,03	0,10	0,00	-0,02	0,11	0,10	0,03	-0,12
EW*	-0,11	-0,04	-0,02	0,16	0,01	0,12	0,03	-0,02	0,11	0,06	0,07	-----
EX	0,23	0,17	0,12	-0,30	0,06	-0,20	0,01	0,03	-0,22	-0,20	-0,07	0,24

Tabelle 7.6: Bias der Klimareferenzprognose im Entwicklungs- (EW) und Examinationskollektiv (EX). Die Kennzeichnung mit einem Stern symbolisiert jene Fälle, in denen an nur 59 Jahren entwickelt wurde.

In Tabelle 7.5 ist die Vorhersageleistung (obere Werte) der 1- bis 6-Monats-Prognosen sowie der (nach den Ergebnissen der einzelnen Monatskombinationen im Entwicklungskollektiv) „besten“ Prognose angegeben, die mit t+x bezeichnet ist. Im Falle des gesamten Kalenderjahres beruhen diese Werte auf insgesamt 360 Vorhersagen (30 Jahre à 12 Monate), bei den einzelnen Jahreszeiten entsprechend nur auf 90 Fällen. Hier basiert die Zuordnung der Prognosen ausschließlich auf dem jeweiligen Zielmonat, d.h. eine Vorhersage des September gehört unabhängig von der Vorlaufzeit stets dem Kollektiv der Herbstprognosen an.
Mit Ausnahme der 2-Monats-Prognosen im Frühling liegen die RV-Werte sämtlich unter Null, d.h. die Vorhersage der jeweiligen Klimamittelwerte wäre vorzuziehen gewesen. Entgegen der meteorologischen Erwartung ist keine Steigerung der Vorhersagegüte mit abnehmendem *lag* auszumachen. Lediglich die 1-Monats-Prognose hebt sich leicht ab. Auch die SS der einzelnen Jahreszeiten unterscheiden sich kaum voneinander. Schließlich ist festzustellen, daß die Verwendung der „besten“ Prognose langfristig keine Vorteile mit sich bringt, eher das Gegenteil ist der Fall. Dies ist als ein weiteres Zeichen der Instabilität zu bewerten, da sich die im Entwicklungskollektiv erzielten Resultate im Examinationskollektiv nicht bewahrheitet haben.
Die Vermutung, ein derart schlechtes Abschneiden sei lediglich die Folge einer Überbewertung einiger weniger markanter Fehlprognosen, konnte durch die Berechnung der skill score basierend auf dem MAE (zweite Zeile der Tabelle 7.5) im wesentlichen entkräftet werden. Zwar nähern sich die Werte der Null geringfügig an, eine positive Vorhersageleistung wird jedoch ebenfalls nicht erzielt. Ähnliches gilt im Falle der möglichen „Umverteilung der großen Streuungen im Winter auf die übrigen Monate und umgekehrt“ (*Röder*, 2001). Um falsche Schlußfolgerungen hierdurch ausschließen zu können, wurde die Berechnung der RV-Werte für das Kalenderjahr mit standardisierten Vorhersagen und Beobachtungen wiederholt. Die daraus resultierenden Ergebnisse (dritte Zeile der Tabelle 7.5) unterscheiden sich kaum von denen ohne Standardisierung.
Bei der Beurteilung der systematischen Fehler (untere Werte der Tabelle 7.5) ist der Bias der Klimareferenzprognose zu berücksichtigen. Da sämtliche Modelle ohne Ausnahme biasfrei entwickelt wurden, beziehen sie sich auf den Mittelwert des Prediktanden im Entwicklungskollektiv. Dieser ist mit dem Bias der Klimaprognose identisch und in der Regel von Null verschieden (vergl. Abschnitt 7.3). Welchen Wert er tatsächlich hat, hängt dabei nicht nur vom Zielmonat, sondern auch vom jeweiligen Vormonat ab: Gehören Vor- und Zielmonat nicht dem selben Kalenderjahr an, so konnte nur an Hand von 59 Fällen entwickelt werden; die Druckdaten des Jahres 1909 dienen dann als Prediktoren für die Temperatur des Jahres 1910 usw. Liegt dagegen kein Jahreswechsel zwischen Vor- und Zielmonat, so umfaßt das Entwicklungskollektiv 60 Fälle. Dementsprechend unterscheiden sich auch die in Tabelle 7.6 angegebenen Mittelwerte der Temperatur voneinander. In der obersten Zeile sind sie für den 60-jährigen Zeitraum 1909 bis 1968, in der mittleren Zeile hingegen für den 59-jährigen Zeitraum 1910 bis 1968 vermerkt. Für den Januar bzw. Dezember ist jeweils nur ein Wert relevant, da deren Vormonate nie bzw. stets im selben Kalenderjahr liegen.
Der Erwartungswert der Vorhersagen im Examinationskollektiv stimmt nun – bei gleicher Verteilung der Einflußgrößen – mit dem Mittelwert des Prediktanden im Entwicklungskollektiv überein. Da hier jedoch ebenfalls ein von Null verschiedener Mittelwert vorhanden ist (allgemein von entgegengesetztem Vorzeichen und im Falle von 60 Entwicklungsjahren exakt von doppeltem Betrag; vergl. Tabelle 7.6), kommt es zu einer Abnormität des Erwartungswertes des mittleren Prognosefehlers $E[ME(Prog)_{EX}]$. Es gilt:

$$E\left[ME\left(\mathrm{Prog}\right)_{EX}\right]=\overline{T}_{EX}-\overline{T}_{EW}\ . \qquad (7.10)$$

Hierin stehen die Abkürzungen EW bzw. EX für Entwicklungs- bzw. Examinationskollektiv.
Entsprechend der Tabelle 7.6 ergibt sich nach Gleichung 7.10 für das Kalenderjahr ein Erwartungswert von ca. -0,03°C. Der tatsächliche Bias der verschiedenen Prognosen liegt zwischen –0,08° und +0,02°C, d.h. ein systematischer Fehler ist nicht gegeben. Ähnliches gilt für die vier Jahreszeiten. Lediglich die 3-Monats-Prognose der Wintermonate liegt um ca. 0,2°C unter dem Erwartungswert; gemessen am Entwicklungszeitraum sagt sie demnach zu warm vorher.
Die Auswirkungen auf die RV-Werte sind mit Sicherheit gering, da die Temperaturmittel im Entwicklungskollektiv im Vergleich zu den entsprechenden Varianzen vernachlässigbar klein sind. Die Abweichungen im Examinationskollektiv tragen nicht zur weiteren Verfälschung der RV-Werte bei, da sie sowohl die Vorhersagen als auch die Klimareferenzprognose tangieren.

Ein etwas anderes Bild präsentiert sich bei Betrachtung der vollständigen gemeinsamen Verteilungen der Vorhersagen und Beobachtungen in Form von „conditional quantile plots“ (siehe dazu auch Abschnitt 5.2). In den Abbildungen 7.1 bis 7.6 sowie 5.1 (1-Monats-Prognose) sind sie für das Kalenderjahr dargestellt. Zwar lassen es sämtliche Prognosen an Schärfe vermissen, doch weist lediglich die 4-Monats-Prognose keine (positive) Auflösung auf. D.h. im Mittel folgt auf eine wärmere Vorhersage auch wärmere Witterung. Dabei ist die Auflösung mit Steigungen der gewichteten Median-Trends kleiner 0,5 jedoch allgemein gering, so daß die Verläßlichkeit als unbefriedigend zu bezeichnen ist. Um die Verläßlichkeit zu steigern, müßten konservativere Vorhersagen gemacht werden. Da dies aber die ohnehin schon dürftige Schärfe weiter mindern würde, ist festzustellen, daß die Vorhersagen wohl hauptsächlich von wissenschaftlichem Wert sein dürften. Ob auch potentielle Verbraucher noch von ihnen profitieren könnten ist fragwürdig.
Während sich für die Jahreszeiten im allgemeinen ein ähnliches Bild ergibt, sticht eine Prognose jedoch deutlich hervor: Die 1-Monats-Prognosen des Sommerquartals (Abb. 7.7) weisen mit einer Steigung des gewichteten Median-Trends von 0,998 nicht nur eine gute Auflösung, sondern auch eine annehmbare Verläßlichkeit auf. Im Gegensatz zu den anderen Prognosen sind sogar die Verläufe des 0,2ten und 0,8ten Quantils verhältnismäßig gut an die 1:1-Diagonale angepaßt. Die Schärfe der Vorhersagen ist jedoch auch in diesem Fall als unzureichend zu bewerten. Dennoch vermittelt Abbildung 7.7 den Eindruck, daß zu dieser Jahreszeit ein realer Zusammenhang zwischen der unmittelbaren Vorzirkulation und der darauffolgenden Berliner Monatsmitteltemperatur besteht. Die multiple lineare Regression ist dabei offenbar nur in der Lage, diesen qualitativ zu erfassen. Die erzielte Genauigkeit kann hingegen nicht befriedigen.
Grundsätzlich bekräftigen die „conditional quantile plots“ den Eindruck, daß durch die prinzipielle Verwendung der 1-Monats-Prognose zu allen Jahreszeiten die bestmögliche Güte der Vorhersagen erzielt werden kann.

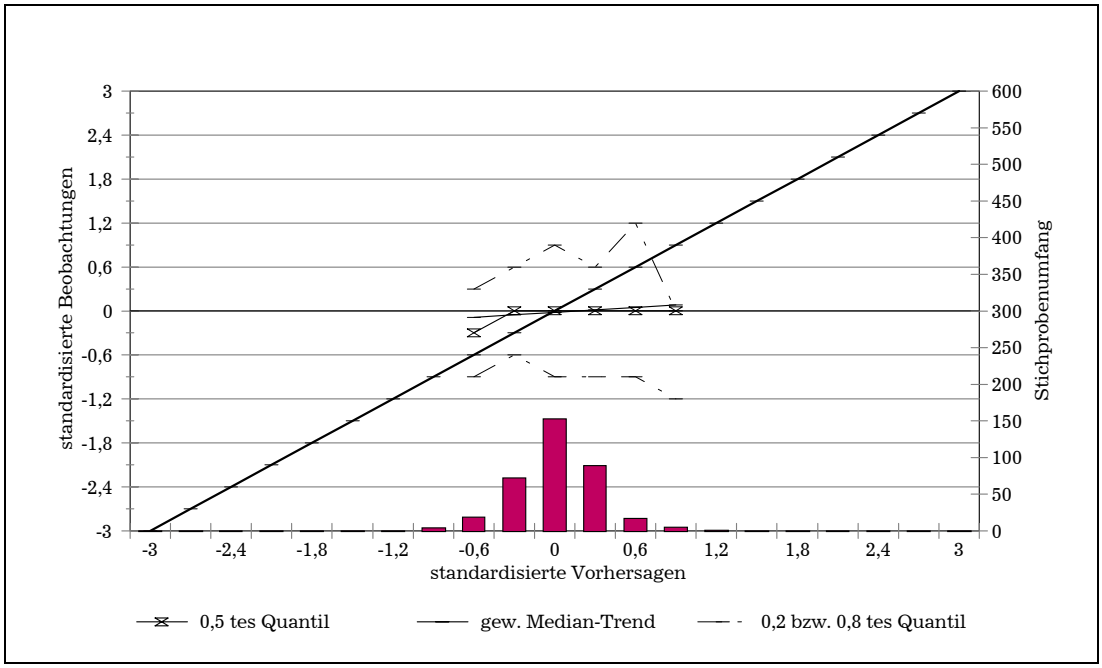

Abbildung 7.1: „Conditional quantile plot“ der 2-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

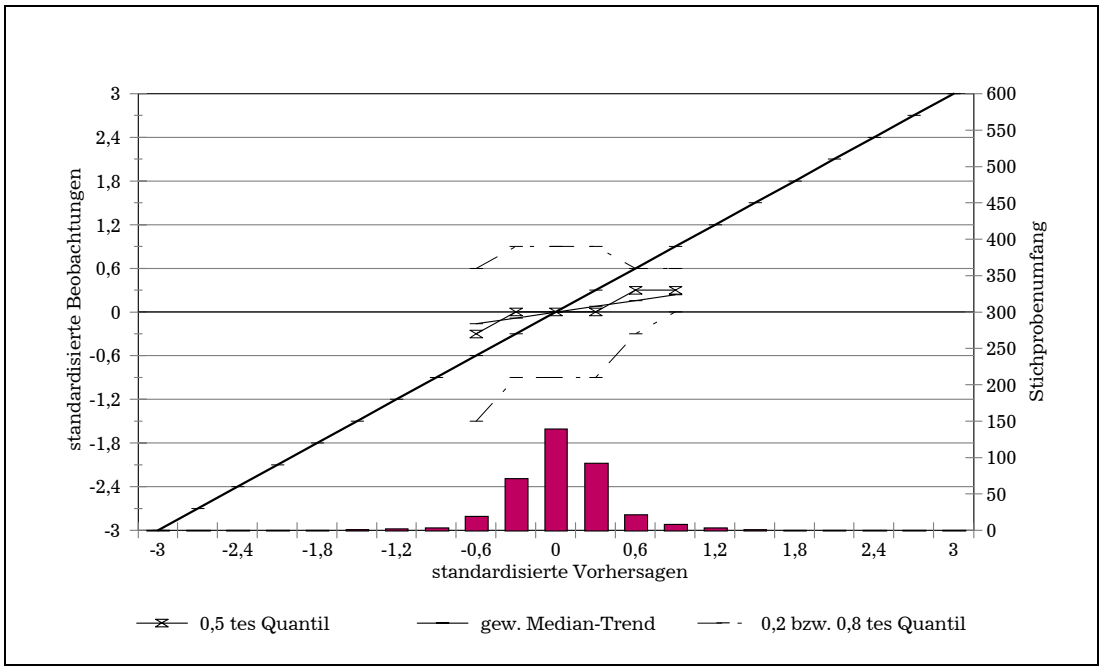

Abbildung 7.2: „Conditional quantile plot“ der 3-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

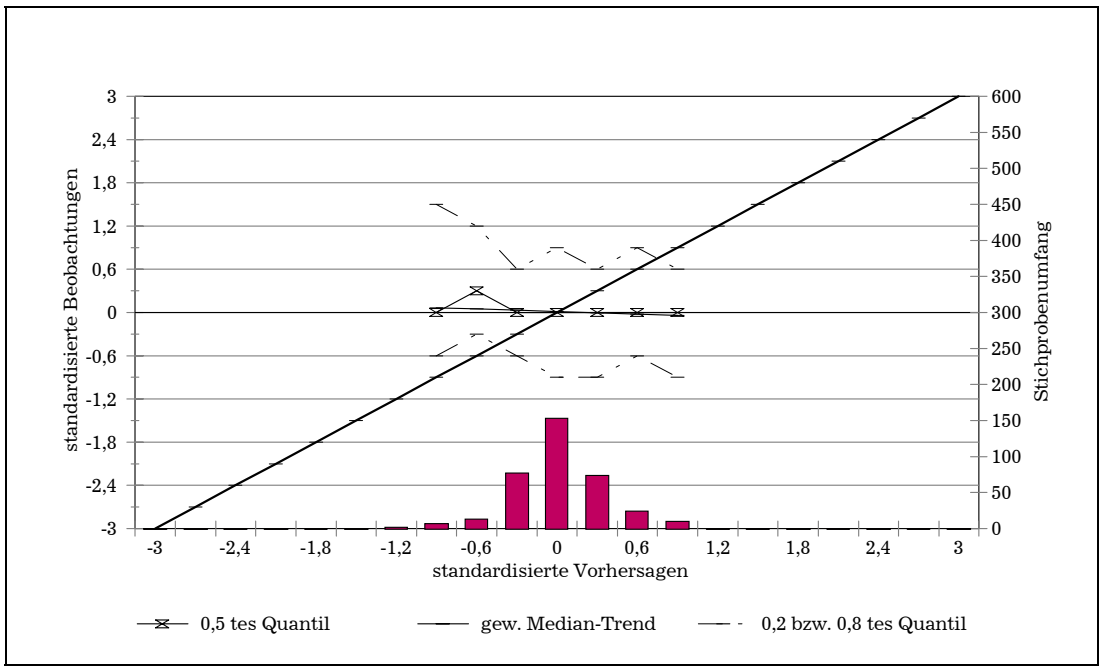

Abbildung 7.3: „Conditional quantile plot" der 4-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

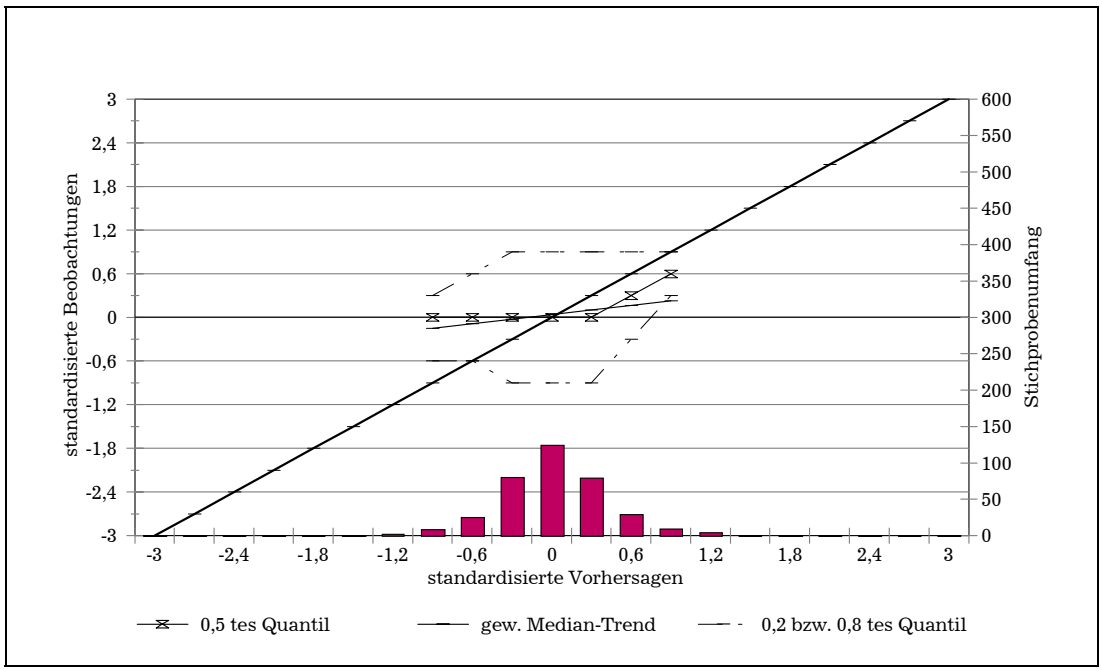

Abbildung 7.4: „Conditional quantile plot" der 5-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

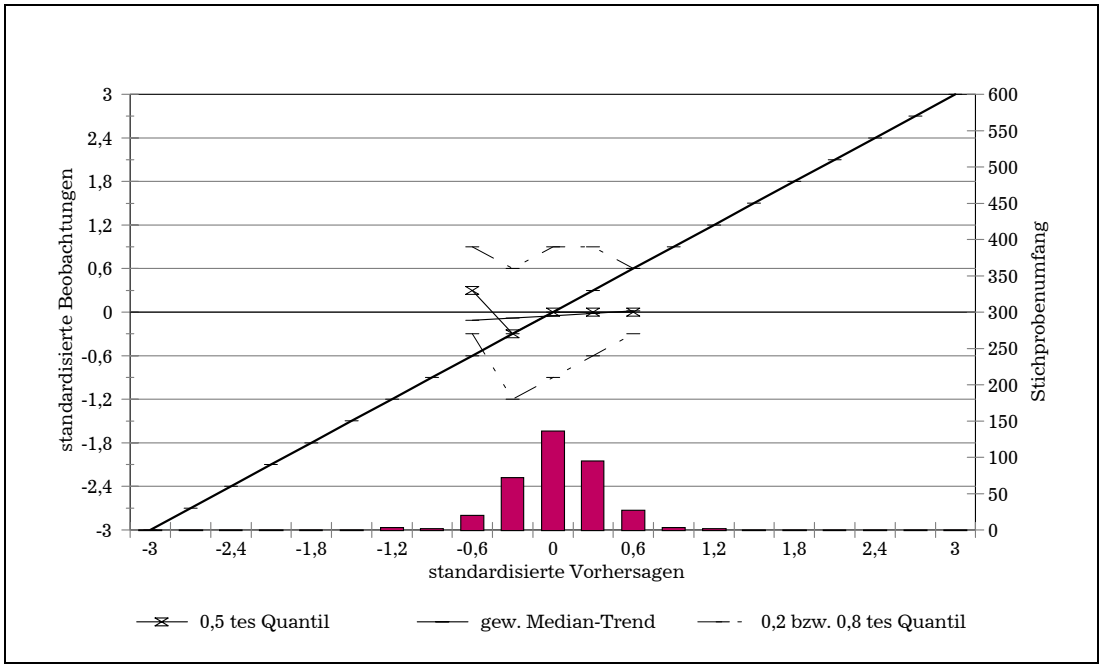

Abbildung 7.5: „Conditional quantile plot“ der 6-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

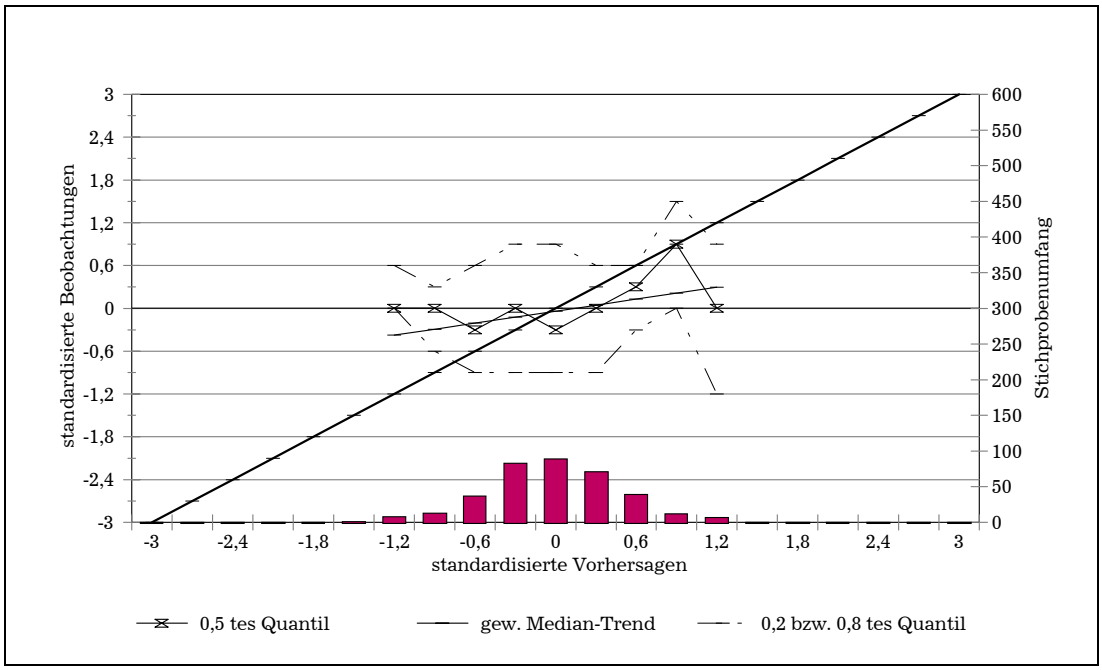

Abbildung 7.6: „Conditional quantile plot“ der nach den Ergebnissen im Entwicklungskollektiv „besten“ Prognosen (variable Vorlaufzeit, siehe Text) der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969-1998.

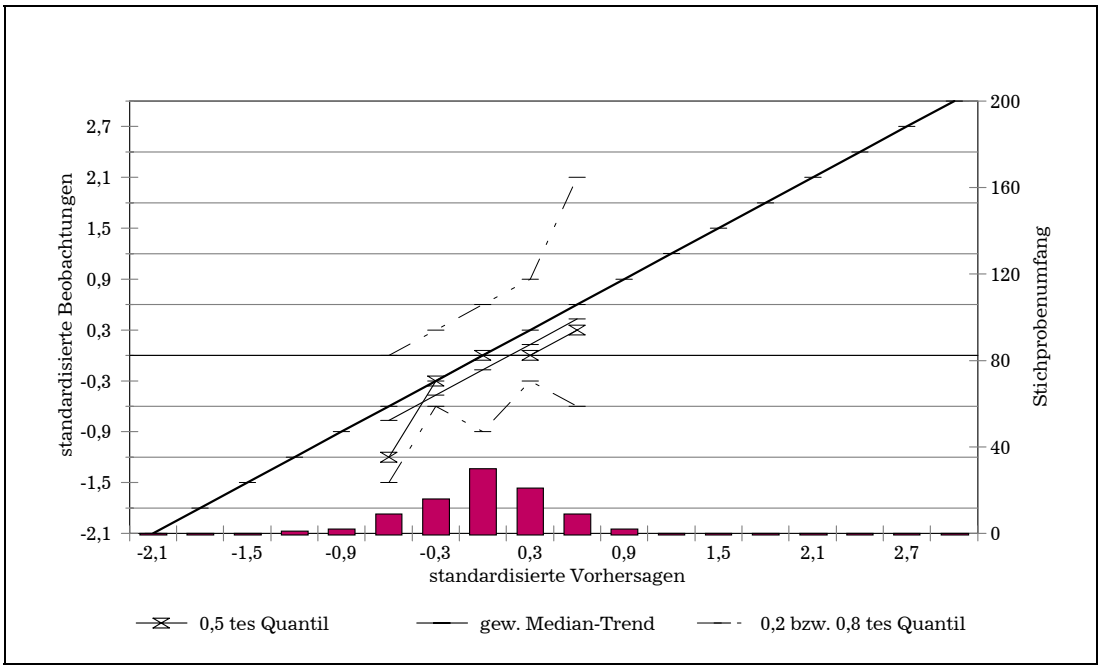

Abbildung 7.7: „Conditional quantile plot" der 1-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für die Sommermonate des Zeitraumes 1969-1998.

t_0	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11
Januar	-0,67 2	-0,24 1	-0,04 1	-0,09 1	-0,60 2	0,02 2	-0,07 1	0,07 1	-0,13 1	0,22 1	-0,16 1
Februar	-0,14 2	-0,22 1	-0,28 1	0,05 1	-0,25 1	-0,31 1	0,07 1	0,00 1	0,09 1	-0,12 1	0,02 2
März	0,02 1	0,15 1	-0,14 1	-0,28 1	-0,37 2	-0,04 1	-0,23 1	-0,08 1	-0,17 1	0,02 1	0,08 1
April	0,01 1	-0,06 1	0,08 2	-0,10 1	-0,18 1	0,06 2	-0,10 1	-0,19 2	-0,28 2	-0,18 1	-0,21 2
Mai	-0,45 2	0,01 1	-0,10 1	-0,11 1	-0,43 3	-1,81 4	-0,11 1	-0,07 1	-0,05 2	-0,29 2	-0,21 1
Juni	-0,12 1	-0,02 1	0,00 1	0,13 1	-0,19 1	-0,01 0,12	-0,02 1	-0,26 1	0,10 2	-0,17 1	-0,10 1
Juli	-0,03 2	0,11 1	-0,56 3	-0,09 1	0,02 1	-0,31 3	-0,12 1	-0,28 2	-0,31 1	-0,07 1	-0,66 2
August	-0,01 1	-0,17 1	-0,20 1	-0,30 2	-0,28 1	0,01 1	-0,30 2	-0,05 1	-0,12 1	-0,06 2	-0,46 4
September	-0,26 1	-0,55 2	-0,29 2	-0,45 3	-0,16 1	-0,04 3	-0,23 1	-0,10 1	-0,07 2	-0,16 2	-0,03 1
Oktober	-0,39 2	0,09 2	-0,28 2	-0,16 3	-0,04 2	-0,05 1	-0,33 1	-0,22 2	-0,17 1	-0,27 1	-0,16 1
November	-0,26 1	-0,34 2	-0,27 1	-0,08 2	-0,06 1	-0,11 1	-0,26 2	-0,06 1	-0,31 1	-0,15 3	-0,02 1
Dezember	-0,54 2	-0,14 1	-0,10 1	0,05 1	-0,11 1	-0,06 2	-0,16 1	-0,15 2	-0,33 1	-0,18 1	-0,38 1

Tabelle 7.7: RV-Werte der Niederschlagsvorhersagen im Examinationskollektiv und Anzahl der verwendeten Prediktoren (untere Zeile).

Die Ergebnisse im Falle des Niederschlages als Prediktand sind mit denen der Temperatur vergleichbar, wobei das Niveau der Vorhersageleistung allgemein noch etwas niedriger ist. In Tabelle 7.7 sind die RV-Werte zusammen mit der jeweiligen Anzahl der verwendeten Prediktoren (untere Zahl) für die 132 Monatskombinationen aufgeführt. Zwar wurde wesentlich häufiger mehr als nur ein Prediktor verwendet und die maximale Anzahl liegt bei vier statt lediglich zwei, jedoch erwiesen sich insbesondere diese Fälle als äußerst instabil. Die 6-Monats-Prognose des November (t_0 ist daher der Mai) erzielt im Examinationskollektiv gar einen RV-Wert von –1,81. Im Entwicklungskollektiv wurde dagegen der Wert 0,48 erreicht. Auch im Falle positiver RV-Werte im Examinationskollektiv verhält es sich wie bei der Temperatur, d.h. die im Entwicklungskollektiv ohne „double Cross-Validation“ ermittelte Vorhersageleistung liegt zumeist darunter.
Da bereits die Stabilität der Modelle zur Temperaturvorhersage strengen Kriterien nicht standzuhalten vermag, bleibt nur die Schlußfolgerung, daß langfristige Niederschlagsvorhersagen mittels multipler linearer Regression höchstens im Einzelfall von Nutzen sein können. Die in Tabelle 7.8 dargestellten Resultate des dauerhaften operationellen Einsatzes untermauern dies.

t+1	t+2	t+3	t+4	t+5	t+6	t+x
-0,24	-0,11	-0,17	-0,13	-0,21	-0,17	-0,37

Tabelle 7.8: Skill score (RV-Werte) der Niederschlagsprognosen mittels multipler linearer Regression in Abhängigkeit von der Vorlaufzeit für das gesamte Kalenderjahr.