Kapiel 6

Verifikation

Verifikationsmaße werden vorwiegend zur Verifikation von numerischen Wettervorhersagenmodellen (NWP, Numerical Weather Prediction) verwendet. Im Folgenden werden die bekanntesten Verifikationsmaße beschrieben und diskutiert. Prinzipiell wird zwischen zwei Arten von Verifikationsmaßen unterschieden, und zwar Verifikationsmaße für:

1. kategorische Größen

2. kontinuierliche Größen.

6.1 Verifikationsmaße für kategorische Größen

Wenn man sich für den einfachsten Fall einer richtigen Niederschlagsprognose interessiert, dann können Verifikationsmaße für kategorische Parameter verwendet werden. Diese beruhen auf Kontingenztabellen (Abb. 6.1), welche eine Häufigkeitsverteilung von *Estimates* und Beobachtungen darstellen. Mit Hilfe dieser Kontingenztabelle können eine Vielzahl an Genauigkeitsmaße (*accuracy measures*) und Qualitätsmaße (*Skill Scores*) berechnet werden (z.B. Wilks (1995); Dorninger (2000); Legates und McCabe Jr. (1999).)

Genauigkeitsmaße

Interessiert man sich für den einfachsten Fall einer Bestimmung der Vorhersagegüte von ja/nein Ereignissen, so wird im Allgemeinen die

- Trefferrate (*Hit Rate, Accuracy or Proportion Correct*)verwendet

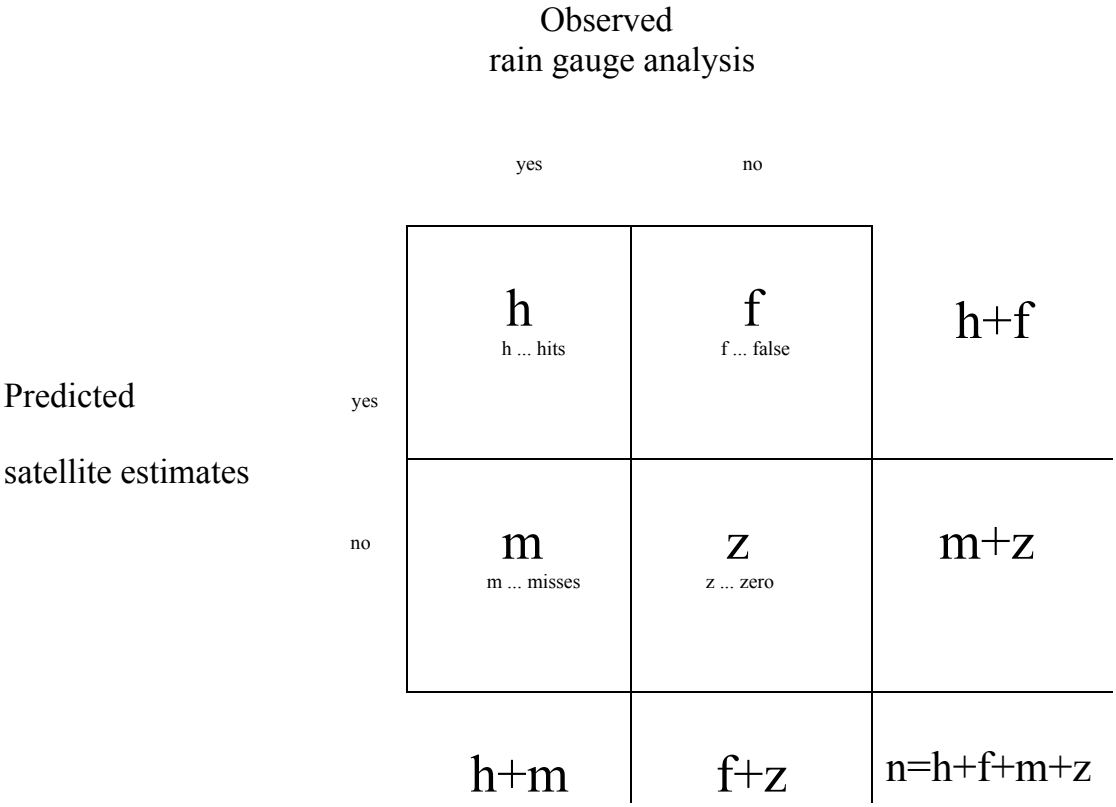

Abb. 6.1: Kontingenztabelle zur Berechnung von Verifikationsmaßzahlen.

Die Trefferrate beschreibt das Verhältnis einer korrekten Vorhersage *hits+zero* zur Gesamtzahl der Beobachtungen *n=hits+false+misses+zero.*

$$\text{HR} = \text{correct estimates / total estimates} = \frac{h + z}{n} \qquad (6.1)$$

Dieses Genauigkeitsmaß hängt sehr von der normalerweise dominierenden Anzahl an positiven Nicht-Ereignissen (*zero*) ab. Vor allem in trockenen Gebieten oder an niederschlagsarmen Tagen kann es zu einer fälschlicherweise guten Trefferquote kommen, wenn pauschal für das gesamte Gebiet kein Niederschlag angenommen wird.

- Entdeckungswahrscheinlichkeit (*Probability of Detection oder Prefigurance*):

Die Entdeckungswahrscheinlichkeit misst den Erfolg einer richtigen Vorhersage, wenn diese auch tatsächlich eintritt. POD ist das Verhältnis aus positiven Ereignissen (*hits*) zur Anzahl an positiven Beobachtungen (*hits + misses*), oder anders ausgedrückt, der prozentuelle Anteil der Ereignisse, die vorhergesagt wurden.

$$\text{P OD =correct rain estimates / rain observation } = \frac{h}{h + m} \qquad (6.2)$$

Die *Probability of Detection* kann Werte von 0 bis 1 annehmen, wobei 1 eine perfekte Prognose darstellt. Gelegentlich kann es vorkommen, dass die *Probability of Detection* eine gute Vorhersage vortäuscht, und zwar dann, wenn die Anzahl der missglückten Ereignisse *misses* die der positiven Ereignisse *hits* übersteigt, d. h. wenn im Allgemeinen die Zahl der Niederschlagsereignisse überschätzt wurde. Um diese Fehlinterpolation zu vermeiden wird zusätzlich die False Alarm Ratio (Glg. 6.3) eingesetzt.

- Die Rate Falschen Alarms (*False Alarm Ratio*):

Die Rate Falschen Alarms misst das Verhältnis, wo Niederschlag vorhergesagt wurde (*false*), aber keiner stattgefunden hat (*hits + false*) und kann mit anderen Worten als Verhältnis der falschen Ereignissen zur Gesamtzahl der positiven Nicht-Beobachtungen beschrieben werden.

$$\text{FAR = false rain estimates / rain estimates} = \frac{f}{h + f} \qquad (6.3)$$

Die Rate falschen Alarms kann Werte von 0 bis 1 annehmen, wobei 0 einer perfekten Prognose entspricht.

- Die Entdeckungswahrscheinlichkeit des falschen Alarms (*Probability of False Detection*):

Die Entdeckungswahrscheinlichkeit des falschen Alarms misst den Erfolg oder Misserfolg einer falschen Vorhersage. *Probability of False Detection* ist das Verhältnis aus falschen Ereignissen (*false*) zur Anzahl der positiven Nicht-Beobachtung (*false + zero*).

$$\text{POFD =false rain estimate / non rain observation} = \frac{f}{f + z} \qquad (6.4)$$

Die *Probability of False Detection* kann Werte von 0 bis 1 annehmen, wobei 0 eine perfekte Prognose darstellt.

- Kritischer Erfolgsindex (*Critical Success Index oder Threat Score*):

Der Kritische Erfolgsindex ist ein Gütemaß, das sowohl falsche Ereignisse *false alarms* als auch missglückte Ereignisse *missed events* berücksichtigt, aber den positiven Nicht-Ereignissen *zero* keine Bedeutung beimisst. In der Tat sind in vielen Anwendungen Nicht-Ereignisse *zero* von untergeordneter Bedeutung.

$$\text{CSI} = \text{correct rain estimates} \ / \ (\text{rain estimates} + \text{observations}) \ = \frac{h}{h + f + m} \quad (6.5)$$

Der *Critical Success Index* erreicht Werte von 0 bis 1, wobei Werte von 1 bedeuten, dass alle positiven Vorhersagen auch wirklich beobachtet wurden. Der *Critical Success Index* ist den positven Ereignissen *hits* direkt, den falschen Ereignissen *false* und den missglückten Ereignisse *misses* indirekt proportional (Schaefer, 1990). Mit einer kleinen algebraischen Manipulation kann man den *Critical Success Index* mit Hilfe der *Probability of Detection* und False Alarm Ratio auch als

$$\text{CSI} = \left[(POD)^{-1} + (1 - FAR)^{-1} - 1 \right]^{-1} \quad (6.6)$$

schreiben. Hier zeigt sich, dass der *Critical Success Index* in Abhängigkeit von der *Probability of Detection* oder *False Alarm Ratio* in einem hochnichtlinearen Zusammenhang steht, der einen falschen Eindruck entstehen lassen könnte, dass der *Critical Success Index* undefiniert sein könnte, wenn die *Probability of Detection* 1 oder die *False Alarm Ratio* 0 ist. Dies ist aber nicht der Fall, da für ein Nullwerden der *Probability of Detection* oder ein Einswerden der *False Alarm Ratio* die positiven Ereignisse hits 0 sein müssten, und dabei würde auch der *Critical Success Index* 0 werden. Der *Critical Success Index* kann auch als Trefferrate ohne Berücksichtigung der Nicht-Ereignisse *zero* interpretiert werden.

- BIAS Score (*Bias*):

Der Bias Score misst das Verhältnis von abgeschätzten und beobachteten Daten, ohne etwas über deren Genauigkeit oder deren räumlichen Verteilung zu sagen. Der Bias Score ist das Verhältnis aus positiver Vorhersage (*hits* + *false*) zu positiven Beobachtung (*hits* + *misses*).

$$\text{Bias} = \text{rain estimates / rain observations} = \frac{h + f}{h + m} \qquad (6.7)$$

Der Bias nimmt Werte von Null bis ∞ an. Wenn der vorhergesagte Niederschlag gleich dem beobachteten Niederschlag ist, so hat der Bias den Wert Eins. Bias-Werte unter Eins werden als "unterschätzt", über Eins als, "überschätzt" bezeichnet.

- Chancenverhältnis (*Odds Ratio*, Θ)

Die *Odds Ratio*, die vorwiegend in der Medizin zum Einsatz kommt, ist definiert als,

$$\Theta = \frac{hz}{fm} \quad (6.8)$$

Die *Odds Ratio* vergleicht die Qualität einer Vorhersage, indem die Wahrscheinlichkeit einer guten Treffervorhersage (*Probability of Detection* (POD)) mit der einer falschen Treffervorhersage *Probability of False Detection* (POFD)) verglichen wird, das zur einer weiteren Schreibweise des Chancenverhältnisses führt.

$$\Theta = \frac{POD}{1 - POD}\left(\frac{POFD}{1 - POFD}\right)^{-1} \quad (6.9)$$

Die *Odds Ratio* ist größer als 1, wenn die Werte der *Probability of Detection* die der *Probability of False Detection* übersteigen. Beobachtungen und Vorhersagen sind voneinander unabhängig, wenn die *Odds Ratio* 1 wird. Wenn die *Odds Ratio* von 1 abgezogen wird, entspricht dies der gewichteten Differenz zwischen der *Probability of Detection* (Glg. 6.2) und der *Probability of False Detection* (Glg. 6.4):

$$\Theta - 1 = \frac{POD - POFD}{POFD\ (1 - POD)} \quad (6.10)$$

Beim Betrachten von Glg. 6.10 erkennt man, dass bei Einswerden der *Odds Ratio* die *Probability of Detection* gleich der *Probability of False Detection* ist. Die *Probability of Detection* und die *False Alarm Ratio* können auf Signifikanz getestet werden, für die *True Skill Statistic* besteht die Möglichkeit der Angabe einer Standardabweichung, die bei Stephenson (2000) nachgelesen werden können.

Skill Scores

Mit Hilfe von *Skill Scores* werden Vorhersagen mit Referenzprognosen, die auf Zufall, Persistenz oder klimatologischen Mitteln aufbauen, verglichen. *Skill Scores* entfernen die auf Zufall aufbauenden Elemente der Vorhersagen. Wenn eine Vorhersage nur auf Zufall aufgebaut ist, gibt es eine nicht geringe Menge an Zahlen c, die zufällig die richtige Vorhersagen produzieren könnten und können. Wenn die Fähigkeit oder Qualität einer Vorhersage von Interesse ist, muss die Verifikationsstatistik so aufgebaut oder modifiziert werden, dass genau diese Zufallsvorhersagen bekannt gemacht werden. Aus diesem Grund wurde eine Vielzahl an Qualitätmaßzahlen (*Skill Scores*) in den letzten Dekaden entwickelt.

Skill Scores basieren immer auf der Definition aus Glg. (6.11) und verwenden im Zähler oder Nenner, Kombinationen der aus der Kontingenztabelle (Abb. 6.1) bekannten Elemente. Abgeleitet werden *Skill Scores* aus der Beziehung

$$SS = \frac{A_{estimated} - A_{reference}}{A_{perfect} - A_{reference}} \quad (6.11)$$

wobei A ein Genauigkeitsmaß ist. Für A = HR (hit rate) gilt: $A_{perfect} = 1$.

Der am Häufigsten verwendete *Skill Scores* ist der *Heidke Skill Score*.

- *Heidke Skill Score*

Der *Heidke Skill Score* verwendet als Grundgenauigkeitsmaß die Trefferrate (Glg.6.1), die durch Verwendung von Zufallsvorhersagen entsteht, und folgt der allgemeinen Definition des Skill Scores (Glg. 6.11). Dabei ist die korrekte Vorhersage durch,

$$[(h + f)(h + m)] /n = [(h + f)(h + m)] /n \qquad (6.12)$$

und die korrekte "Nicht-" Vorhersage durch,

$$[(f + z)(m + z)] /n = [(f + z)(m + z)] /n \qquad (6.13)$$

gegeben. Wenn man Glg. 6.12 und Glg. 6.13 in die allgemeine Definition des *Skill Scores* (Glg. 6.11) einsetzt, erhält man:

$$\text{HSS} = \frac{(h + z)/n - [(h + f)(h + m) + (f + z)(m + z)]/n}{n - [(h + f)(h + m) + (f + z)(m + z)]/n} \qquad (6.14)$$

Durch Ausmultiplikation und Vereinfachung von Glg. 6.14 erhält man die Gleichung für den *Heidke Skill Score.*

$$\text{HSS} = \frac{hz - fm}{[(h + m)(m + z) + (h + f)(f + z)]/2} \qquad (6.15)$$

Der *Heidke Skill Score* ist definiert im Intervall von -1 bis 1. Es bedeutet 1 eine perfekte Vorhersage, 0 eine Vorhersage die gleich der Referenzprognose und negative Werte bedeuten, dass die Vorhersage schlechter als die Referenzprognose sind.

- *True Skill Statistics (oder Hanssen and Kuipers Score*):

Der *Hanssen und Kuipers Score* ist ähnlich dem Heidke Skill Score aufgebaut. Die *True Skill Statistics* verwendet als Referenzprognose auch die Trefferrate, nur daß die Wahrscheinlichkeit einer korrekten Vorhersage und die Wahrscheinlichkeit einer korrekten "Nicht-" Vorhersage im Nenner frei von systematischen Abweichungen (*unbiased*) sein soll (vgl. Wilks (1995)). Die Wahrscheinlichkeit einer korrekten Vorhersage ohne systematischer Abweichung lässt sich als

$$[(h + m)(h + m)] /n = (h + m)^2 /n \qquad (6.16)$$

schreiben. Die Wahrscheinlichkeit einer korrekten "Nicht-" Vorhersage ohne systematischer Abweichung als

$$[(f + z)(f + z)] / n = (f + z)^2 / n \qquad (6.17)$$

Mit Glg. 6.16 und Glg. 6.17 und der allgemeinen Gleichung für *Skill Scores* (Glg..11) lässt sich die *True Skill Statistics* als

$$T\,SS = \frac{(h + z) - [(h + f)(h + m) + (f + z)(m + z)] / n}{n - [h + m)^2 + (f + z)^2] / n} \qquad (6.18)$$

schreiben, und durch Ausmultiplizieren und Vereinfachen der Ausdrücke ergibt sich die Gleichung für die *True Skill Statistics* als

$$T\,SS = \frac{hz - fm}{(h + m)(f + z)} \qquad (6.19)$$

Durch erweitern von Glg. 6.19 mit 0 = hf - hf folgt

$$T\,SS = \frac{h}{h + m} - \frac{f}{f + z} \qquad (6.20)$$

das bedeutet, dass der *Hanssen und Kuipers Score* auch als

$$T\,SS = POD - POFD \qquad (6.21)$$

geschrieben werden kann. Die *True Skill Statistics* kann Werte von -1 bis 1 erreichen, wobei wiederum Werte von 1 eine perfekte Prognose, 0 eine Prognose gleich der Referenzprognose und ein negativer Wert eine schlechtere Vorhersage als die Referenzprognose darstellt. Der *Hanssen-Kuipers Score* ist das einzige Verifikationsmaß das frei von systematischen Abweichungen ist (vgl. z.B. Tab. 6.1 auf Seite 58).

- *Equitable Threat Score (oder Gilbert Skill Score*):

Der *Equitable Threat Score* ist eine Modifikation des *Critical Success Index* (Glg.6.5), das die Anzahl der positiven Ereignisse h, die nur auf Zufall basieren, eliminiert.Wollen wir nun noch einige Abkürzungen mit Hilfe der Kontingenztabelle einführen:

p = h + f soll die Anzahl der positiven Vorhersagen sein

e = h + m soll die Anzahl der positiven Beobachtungen sein

Damit lässt sich der *Critical Success Index* (Glg. 6.5) schreiben als

$$\text{CSI} = \frac{h}{p + e - h} \qquad (6.22)$$

Um den CSI soweit zu modifizieren, dass die vom Zufall herrührenden Vorhersagen c berücksichtigt werden, werden diese von den positiven Ereignissen h abgezogen (h-c), was zu

$$\text{ET S} = \frac{h - c}{h - c + m + f} = \frac{h - c}{p + e - h - c} \qquad (6.23)$$

führt. Dabei ist c nichts anderes als die Anzahl der positiven Beobachtungen e (h+m) multipliziert mit der Anzahl der positiven Vorhersagen p (h + f).

$$\text{c} = \frac{pe}{n} = \frac{(h + m)(h + f)}{h + f + m + z} \qquad (6.24)$$

Eingesetzt in Glg. (6.23) folgt daraus der *Equitable Threat Score* als:

$$\text{ET S} = \frac{[h - p(e/n)]}{[(p + e - h) - p(e/n)]} \qquad (6.25)$$

und mit Termen der Kontingenztabelle zu:

$$\text{ET S} = \frac{hz - fm}{(f + m)n + (hz - fm)} \quad (6.26)$$

Aus Glg. 6.26 ist ersichtlich, dass das Minimum des ETS dann eintritt, wenn (1) keine positiven Ereignisse oder "Nicht-" Ereignisse gemacht wurden (h = z = 0) und (2) wenn die Anzahl der missglückten Ereignisse *misses* gleich der Anzahl der falschen Vorhersagen (m + z) ist. Der *Equitable Threat Score* reicht von -1/3 bis 1.

Der ETS erreicht Werte von Null, wenn entweder ein Ereignis immer vorhergesagt wurde (p = n), oder wenn ein Ereignis immer auftritt (e = n). Wenn extrem seltene Ereignisse auftreten (e/n), strebt der ETS gegen den CSI.

$$\lim_{e/n \to 0} ETS = \frac{h}{p + e - h} = CSI$$

In der Literatur (z.B. Ebert und McBride (1997)) wird öfters darauf eingegangen, welcher *Skill Score* nun am fähigsten ist, die Güte einer Prognose korrekt zu quantifizieren. Schaefer (1990) zeigte, dass der ETS mit dem HSS durch ET S = HSS/(2- HSS) verknüpft ist, wobei der Unterschied in diesen zwei *Skill Scores* darin besteht, dass für den ETS in erster Linie die korrekten Vorhersagen von Ereignissen von Belang sind, und der HSS auch die korrekte Vorhersage von Nicht-Ereignissen beachtet. Die TSS gibt sowohl Ereignissen als auch Nicht-Ereignissen ein entsprechendes Gewicht. Stanski et al. (1989) und Doswell III et al. (1990) kritisierten, dass die *True Skill Statistics* zu sehr von der *Probability of Detection* abhänge, und dem Zufolge, um die Genauigkeit einer Vorhersage von seltenen Ereignissen zu bestimmen, weniger geeignet sei als der *Heidke Skill Score*.

Dennoch kann diese Eigenschaft als eine positive angesehen werden, da ein korrektprognostiziertes kein-Regen-Ereignis oder ein Regen-Ereignis einen größeren Einfluss auf die *True Skill Statistics* hat, wenn das Ereignis ein seltenes ist (Wilks, 1995). Der *Hanssen-Kuipers Score* hat aber den Vorteil, dass er von der Verteilung von Ereignissen und Nicht-Ereignissen unabhängig ist (Woodcock, 1976; Ebert und McBride, 1997). Ein weiterer Grund, warum der *Hanssen und Kuipers Score* als *True Skill Statistics* bezeichnet wird, ist der, dass er davon unabhängig ist, ob es sich um ein trockenes oder um ein feuchtes Regime handelt. Die anderen *Skill Scores* (*Heidke Skill Score*, *Equitable Threat Score*) gelten eigentlich nur für die momentane Verteilung des Niederschlags, da die Referenzvorhersage der Zufall ist, und dieser von der beobachteten Niederschlagsverteilung abhängt (Ebert und McBride, 1997).

Um diesen Sachverhalt näher darzustellen wurde von Ebert und McBride (1997) ein kleines Experiment gemacht: Wenn die Verteilung von Ereignissen und Nicht-Ereignissen bekannt sind, dann sollte die Fähigkeit des Modells nur eine Funktion dieser Genauigkeits- maße sein, und sollte nicht von der Häufigkeitsverteilung bzw. von der Niederschlagsmengeabhängen. In

Abb. 6.2 sind fünf Genauigkeitsmaße, die *Hitrate* (HR), der *Critical Success Index* (CSI), der *Equitable Threat Score* (ETS), der *Heidke Skill Score* (HSS) und die *True Skill Statistics* (TSS) als Funktion der Niederschlagsverteilung geplottet. Dabei kann man beobachten, dass die HR linear mit zunehmenden Anteil an Niederschlag bei trockenem Regime abnimmt, und alle anderen Scores mit Ausnahme der *True Skill Statistics*, die konstant beleibt, und dem CSI, der zunimmt. Daraus ist leicht ersichtlich, dass der CSI die Vorhersage in Regionen oder Monaten mit hoher Niederschlagshäufigkeit eine zu große Qualität beigemessen wird.

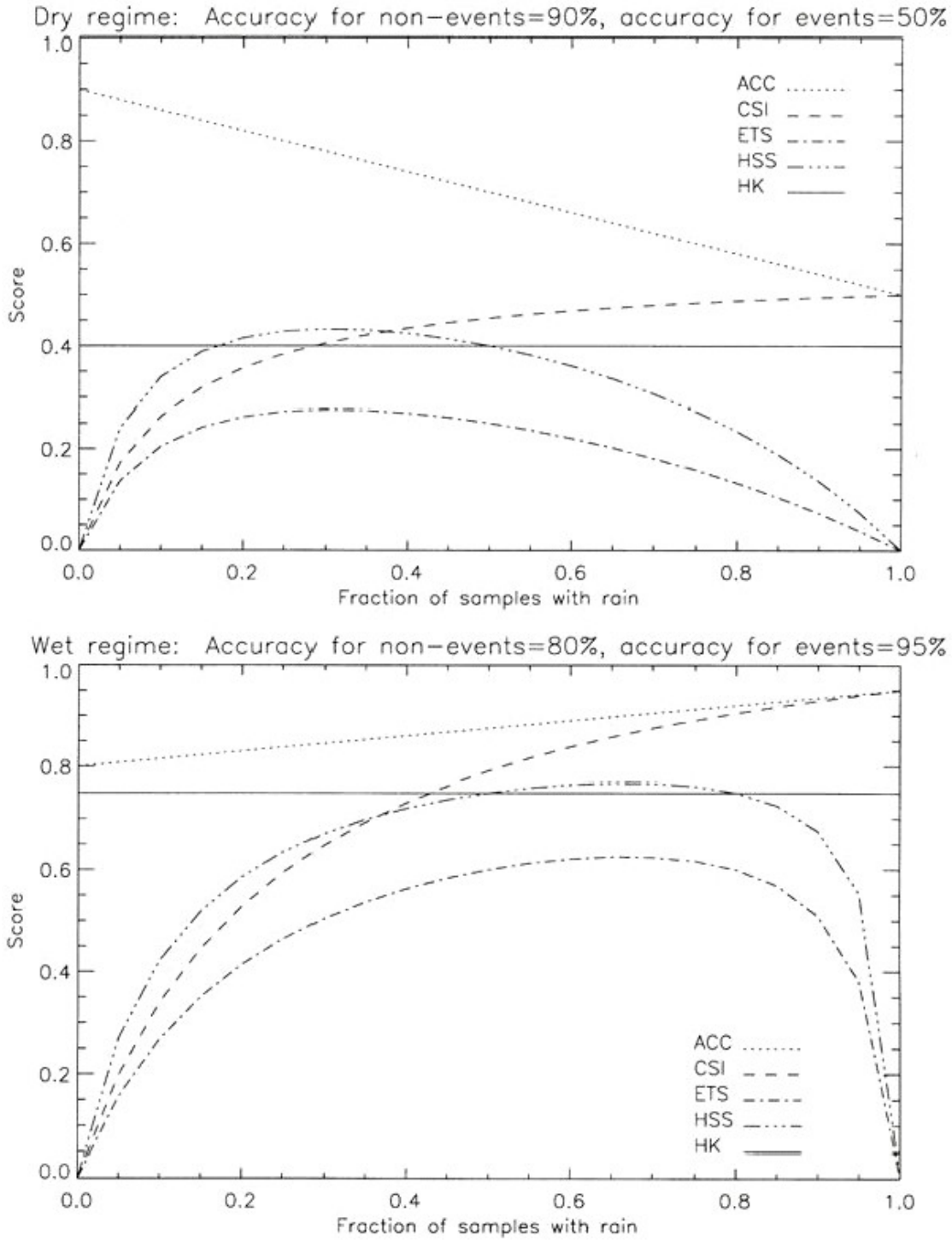

Abb. 6.2: Fünf Genauigkeitsmaße: Die *Hitrate* (HR=ACC), der *Critical Success Index* (CSI), der *Equitable Threat Score* (ETS), der *Heidke Skill Score* (HSS) und die *True Skill Statistics* (TSS) als eine Funktion für ein (a) trockenes Regime, und für ein (b)feuchtes Regime (Ebert und McBride 1997).

Elemente der Kontingenztabelle				Skill Scores			Accuracy
h hit	f false	m miss	z zero	Heidke Skill Scores (HSS)	True Skill Stati-stics (TSS)	Equitable Skill Score (ETS)	Hitrate (HR)
150	0	50	0	0.000	_	0.000	0.750
135	10	45	10	0.141	0.250	0.076	0.725
120	20	40	20	0.211	0.250	0.118	0.700
105	30	35	30	0.244	0.250	0.139	0.675
90	40	30	40	0.255	0.250	0.146	0.650
75	50	25	50	0.250	0.250	0.143	0.625
60	60	20	60	0.231	0.250	0.130	0.600
45	70	15	70	0.198	0.250	0.110	0.575
30	80	10	80	0.151	0.250	0.082	0.550
15	90	5	90	0.087	0.250	0.045	0.525
0	100	0	100	0.000	_	0.000	0.500

Tab. 6.1: Variation unterschiedlicher *Skill Scores*, bei 75% Vorhersage von Ereignissen und 50% Vorhersage von Nicht-Ereignissen mit variierender Verteilung (abgeändert nach Woodcock (1976)).

Tabelle 6.1 erklärt denselben Sachverhalt für die konstanten Werte von *accuracy for events* =75% und *accuracy for non-events*=50%. Damit kann gezeigt werden, wie die verschiedenen *Skill Scores* von der Verteilung der Elemente aus den Kontingenztabellen abhängen.

In diesem Beispiel wurde eine 75%-prozentige Trefferrate bei Ereignissen und eine 50%-prozentige Trefferrate bei Nicht-Ereignissen vorgegeben. Man erkennt die lineare Abnahme der Trefferrate mit abnehmenden h *(hits*) und zunehmenden z (zero). Nur bei einer ausgeglichenen Verteilung der f (*false*) und der z (*zero*) (f = z = 50) sind der *Heidke Skill Score* (HSS) und die *True Skill Statistics* (TSS) identisch, daher unabhängig von der Häufigkeitsverteilung der ja/nein Ereignisse. In dieser Arbeit wurde Regen/kein Regen Ereignisse, oder anders ausgedrückt feuchte und trockene Perioden betrachtet.

In den Tabellen 6.2 und 6.3 sind die Genauigkeitsmaße und die *Skill Scores* mit ihren unterschiedlichen Bezeichnungen, wie sie in der Literatur auftreten, zusammengefasst.

Genauigkeitsmaße (Accuracy Measures)					
Hit rate(HR)	Percent correct	Proportion correct	Ratio test		Glg. (6.1)
Critical success index (CSI)	Threat score	Prefigurance	Ratio of verification	Gilbert Skill Score[2]	Glg. (6.5)

Tab. 6.2: Übersicht über die unterschiedlichen Bezeichnungen von Genauigkeitsmaßn .

Skill Scores				
True skill statistics (TSS)	Hanssen-Kuipers discriminant	Kuipers' performance index	Peirce Skill Score	Glg. (6.19)
Equitable threat score (ETS)	Gilbert skill Score			Glg. (6.26)

Tab. 6.3: Übersicht über die unterschiedlichen Bezeichnungen von *Skill Scores* .

In dieser Arbeit wird die *True Skill Statistics* als Maß für die Abschätzung der Genauigkeit der EZMW-Niederschlagsprognose verwendet. Sie wurde deswegen ausgewählt, da die Fähigkeit der Niederschlagsereignisse korrekt abzuschätzen, unabhängig von der unterschiedlichen Niederschlagsverteilung der Beobachtungen, von Interesse ist.

6.2 Verifikationsmaße für kontinuierliche Größen

Mit kontinuierlichen Größen wird die Fähigkeit der Vorhersage (der zu verifizierende Größen)korrekte Lagen und Betrag des Niederschlags wiederzugeben, gemessen. In allen folgenden Gleichungen bedeuten f_i die i-Werte der vorhergesagten (abgeschätzten) Daten, die O_i die der beobachteten Daten.

- Mittlerer Fehler (*Mean Error, Bias, ME*):

Der mittlere Fehler zeigt die mittlere Richtung der Abweichung von den Beobachtungen, aber nicht dessen Betrag. Ein positiver mittlerer Fehler zeigt, dass die vorhergesagten (abgeschätzten) Werte die beobachteten Werte im Mittel um diesen Betrag überschätzen, wobei ein negativer mittlerer Fehler einer Unterschätzung der beobachteten Werte im Mittel gleichkommt.

$$\text{M E} = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) \qquad (6.28)$$

- Mittlerer absoluter Fehler: (*Mean absolute error (MAE)*):

Der mittlere absolute Fehler ist ein lineares Maß, das das Mittle deren Beträge aller Fehler wiedergibt, aber nicht dessen Richtung.

$$\text{M AE} = \frac{1}{N}\sum_{i=1}^{N}|(F_i - O_i)| \qquad (6.29)$$

- Wurzel aus dem mittleren quadratischen Fehler (*Root mean square error RMSE*):

Der *Root Mean Square Error* ist ebenfalls ein lineares Maß, welches den mittleren Betrag des Fehlers wiedergibt, das aber mit dem Quadrat des Fehlers gewichtet ist.Wie der mittlere Fehler gibt der RMS die Richtung des Fehlers nicht an. Wenn man RMSE mit dem mittleren absoluten Fehler vergleicht, so bekommen größere Fehler ein größeres Gewicht im quadratischen Mittel. Deswegen ist RMSE ein Maß das große Fehler hervorhebt.

$$\text{RM SE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \qquad (6.30)$$

- Produktmoment-Korrelationskoeffizient (Pearson-Korrelation *r*):

Unter der Voraussetzung, dass zwischen zwei normalverteilten Größen ein linearer Zusammenhang besteht, kann der Pearson-Korrelationskoeffizient angewandt werden (Schönwiese, 2000). Der Korrelationskoeffizient gibt eine Auskunft über Stärke und Richtung eines Zusammenhangs und kann Werte von -1 bis +1 erreichen. Ist r = 1, so bedeutet dies, dass die zwei Größen in einem 100 prozentigem linearen Zusammenhang stehen. Bei negativen Werten von r spricht man von einer negativen Korrelation, d.h., dass zu kleinen x-Werten große y-Werte gehören. Der Pearson-Korrelationskoeffizient wird als

$$r = \frac{\sum_{i=1}^{N}(F_i - \overline{F})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \overline{F}^2)}\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}} \quad (6.31)$$

geschrieben. Da Tagessummen des Niederschlags in der Regel rechtsschief verteilt sind, ist die Verwendung des Pearson-Korrelationskoeffizienten nach strengen statistischen Maßstäben nicht zulässig. Für monatliche und jährliche Niederschlagssummen, die annähernd normalverteilt sind, ist die Anwendung des linearen Korrelationskoeffizienten angezeigt.

- Rangkorrelation (*Spearman Correlation, Rank-order correlation rs*):

Die Stärke des Zusammenhangs zwischen nicht- normalverteilten Größen (z.B. tägliche Niederschlagsfelder) kann mit den parameterfreien Rangordnungskorrelationskoeffizienten nach SPEARMAN berechnet werden. Die Grundidee der Rangkorrelation besteht darin, dass jede Beobachtung O_i und jede Vorhersage f_i durch dessen Rang in der Gesamtheit aller Beobachtungen O und Vorhersagen f ersetzt werden. Daraus folgt eine genau definierte Verteilungsfunktion der Ränge. Dabei werden allen Beobachtungswerten O_i mit unterschiedliche Werten, jedem Wert genau ein Rang zugeordnet und zu allen unterschiedlichen Vorhersagewerten (zu verifizierenden Größen) f_i, jedem Wert genau ein Rang zugeordnet.Weisen einige der Beobachtungswerte O_i oder einige der Vorhersagewerte f_i gleiche Werte auf, spricht man von Bindungen. Nun sollen R_i die Rangzahlen aller Beobachtung O_i und S_i alle Rangzahlen der Vorhersagen f_i sein. Damit lässt sich der Rangordnungskoeffizient als linearer Korrelationskoeffizient der Ränge definieren, und zwar als,

$$r_s = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{N}(S_i - \bar{S})^2}} \tag{6.32}$$

Für die numerische Berechnung des Rangkorrelationskoeffizienten wird eine Definition des Rangkorrelationkoeffzienten mit Hilfe eines anderen parameterfreien Korrelationskoeffizienten, den so genannten *sum squared difference of ranks* (Press et al., 1994), verwendet, der definiert ist als

$$\mathrm{D} = \sum_{i=1}^{N}(R_i - S_i)^2 \tag{6.33}$$

Damit lässt sich der Rangkorrelationskoeffizient, wenn keine Bindungen vorhanden sind, als exaktes Verhältnis zwischen D und r_s schreiben als

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} \tag{6.34}$$

Glg. 6.34 lässt sich nur dann anwenden, wenn weniger als 1/5 ranggleiche Bindungen vorhanden sind. Treten in den Reihen mehr als 1/5 ranggleiche Bindungen auf, muss eine Korrektur durchgeführt werden, die zur Definition des korrigierten Rangordnungskoeffizienten führt, der definiert ist als

$$r_s = \frac{1 - \frac{6}{N^3 - N}\left[D + \frac{1}{12}\sum_k (f_k^3 - f_k) + \frac{1}{12}\sum_m (g_m^3 - g_m)\right]}{\left[1 - \frac{\sum_f (f_k^3 - f_k)}{N^3 - N}\right]^{\frac{1}{2}}\left[1 - \frac{\sum_m (g_m^3 - g_m)}{N^3 - N}\right]^{\frac{1}{2}}} \tag{6.35}$$

Dabei sind die f_k die Anzahl der Bindungen der k-ten Gruppe innerhalb der R_i und die g_m die der S_i. Man beachte, dass bei Einswerden von f_k und g_m keine Bindungen vorhanden sind und Glg. 6.35 dabei zu Glg. 6.34 wird. Die Signifikanz eines Rangkorrelationskoeffzientenwertes $r_s = 0$ kann mit Hilfe des t-Tests auf Signifikanz abgeschätzt werden (vgl. Press et al. (1994)). Die Prüfung auf Signifikanz kann für die beiden Korrelationskoeffizienten nach PEARSON und SPEARMAN auch mit Hilfe der kritischenWerte in Tab. 6.3 (r_s für den Korrelationskoeffizienten nach SPEARMAN, r^* für den nach PEARSON) bestimmt werden. Dabei muss die Größe des Stichprobenumfangs

bekannt sein. Zum Beispiel nehme man an, dass ein Stichprobenumfang von n = 30 und eine Korrelation nach SPEARMAN von $r_s = 0.45$ gegeben sei. Man prüft dann, ob die Korrelation größer als der kritische Wert r aus Tab. 6.3 ist. Nimmt man eine Irrtumswahrscheinlichkeit von z.B α= 0.05 (5%) an, dann ist der kritische Wert $r_s^* = 0.364$. Dann ist die Korrelation zwischen den beiden Reihen statistisch mit 95%iger Wahrscheinlichkeitabgesichert. Auf ähnliche Weise wird für den Korrelationskoeffizienten nach PEARSON vorgegangen.

- Bestimmtheitsmaß (*Coefficient of Determination* R^2*):*

Der *Coefficient of Determination* ist das Quadrat des Produktmoment-Korrelationskoeffizienten nach Pearson und ist die durch das Modell erklärte Varianz. R^2 ist

	α (zweiseitig)			
n	0.10	0.05	0.02	0.01
5	0.900	-	-	-
6	0.829	0.866	0.943	-
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.818
12	0.497	0.591	0.703	0.780
13	0.475	0.566	0.673	0.745
14	0.457	0.545	0.646	0.716
15	0.441	0.525	0.623	0.689
16	0.425	0.507	0.601	0.666
17	0.412	0.490	0.582	0.645
18	0.399	0.476	0.564	0.625
19	0.388	0.462	0.549	0.608
20	0.377	0.450	0.534	0.591
21	0.368	0.438	0.521	0.576
22	0.359	0.428	0.508	0.562
23	0.351	0.418	0.496	0.549
24	0.343	0.409	0.485	0.537
25	0.336	0.400	0.475	0.562
26	0.329	0.392	0.465	0.515
27	0.323	0.385	0.456	0.505
28	0.317	0.377	0.448	0.496
29	0.311	0.370	0.440	0.487
30	0.305	0.364	0.432	0.478
31	0.300	0.358	0.425	0.470
32	0.295	0.352	0.418	0.463
33	0.291	0.347	0.411	0.455
34	0.286	0.341	0.405	0.448
35	0.282	0.336	0.399	0.442
36	0.278	0.331	0.393	0.435
37	0.274	0.327	0.388	0.429
38	0.270	0.322	0.382	0.424
39	0.267	0.318	0.377	0.418
40	0.263	0.314	0.373	0.413
	0.05	0.025	0.01	0.005
	$\frac{\alpha}{2}$ (einseitig)			

	α (zweiseitig)			
n	0.10	0.05	0.02	0.01
3	0.988	0.997	0.995	0.999
4	0.900	0.950	0.980	0.990
5	0.805	0.878	0.934	0.959
6	0.729	0.811	0.882	0.917
7	0.669	0.754	0.833	0.874
8	0.622	0.707	0.789	0.834
9	0.582	0.666	0.750	0.798
10	0.549	0.632	0.716	0.765
11	0.521	0.602	0.685	0.735
12	0.497	0.576	0. 658	0.708
13	0.476	0.553	0.634	0.684
14	0.458	0.532	0.612	0.661
15	0.441	0.514	0.592	0.641
16	0.426	0.497	0.574	0.623
17	0.412	0.482	0.558	0.606
18	0.400	0.468	0.542	0.590
19	0.389	0.456	0.528	0.575
20	0.378	0.444	0.516	0.561
21	0.369	0.433	0.503	0.549
22	0.360	0.423	0.492	0.537
23	0.352	0.413	0.482	0.526
24	0.344	0.404	0.472	0.515
25	0.337	0.396	0.462	0.505
26	0.330	0.388	0.453	0.496
27	0.323	0.381	0.445	0.487
28	0.317	0.374	0.437	0.479
29	0.311	0.367	0.430	0.471
30	0.306	0.361	0.423	0.463
31	0.301	0.355	0.416	0.456
32	0.296	0.349	0.409	0.449
33	0.291	0.344	0.403	0.442
34	0.287	0.339	0.397	0.436
35	0.283	0.334	0.392	0.430
36	0.279	0.329	0.386	0.424
37	0.275	0.325	0.381	0.418
38	0.271	0.320	0.376	0.413
39	0.267	0.316	0.371	0.408
40	0.264	0.312	0.366	0.403
	0.05	0.025	0.01	0.005
	$\frac{\alpha}{2}$ (einseitig)			

Abb. 6.3: Tabelle mit kritischen Werten r für den Korrelationskoeffizienten nach SPEARMAN (links) und mit kritischen Werte r für den Korrelationskoeffizienten nach PEARSON (rechts) zur Prüfung der Signifikanz mit verschiedenen Irrtumswahrscheinlichkeiten α.

definiert zwischen 0 und 1, wobei ein höherer Wert bessere Übereinstimmung in den Daten bedeutet und wird als

$$R^2 = \left\{ \frac{\sum_{i=1}^{N}(O_i - \overline{O})(F_i - \overline{F})}{\left[\sum_{i=1}^{N}(O_i - \overline{O})^2\right]^{0.5}\left[\sum_{i=1}^{N}(F_i - \overline{F})^2\right]^{0.5}} \right\}^2 \quad (6.36)$$

geschrieben (Legates und McCabe Jr., 1999).

- *Coefficient of Efficiency* E:

$$\mathrm{E} = 1 - \frac{\sum_{i=1}^{N}(O_i - F_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \quad (6.37)$$

Der *Coefficient of Efficiency* E ist im Intervall von - ∞ bis 1 definiert. Demnach ist die *Coefficient of Efficiency* nichts anderes als 1 minus dem Verhältnis des mittleren quadratischen Fehlers (MSE),

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2 \quad (6.38)$$

zur Varianz der beobachteten Daten. Zum Beispiel wird der *Coefficient of Efficiency* E = 0, wenn das Quadrat der Differenzen zwischen Vorhersage und Beobachtung die gleiche Variabilität in den Beobachtungen aufweisen. E wird mit anwachsender Variabilität der Beobachtungen negativ. Werte von E = 0 deuten darauf hin, dass das Mittel der Beobachtungen in der Vorhersage gleich gut wie das Modell ist, negative Werte von E zeigen eine bessere Vorhersage durch das Mittel der Beobachtungen. Die *Coefficient of Efficiency* E ist sensitiv auf große Fehler, da diese mit dem Quadrat in die Berechnung von E eingehen (Legates und McCabe Jr., 1999).

- Index of Agreement d:

Der Index of Agreement ist im Intervall von 0 bis 1 definiert und ist definiert als,

$$\mathrm{d} = 1 - \frac{\sum_{i=1}^{N}(O_i - F_i)^2}{\sum_{i=1}^{N}\left(\left|F_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \quad (6.39)$$

und kann mit Hilfe des "potentiellen Fehlers" (PE),

$$\mathrm{PE} = \sum_{i=1}^{N} \left(\left| F_i - \overline{O} \right| + \left| O_i - \overline{O} \right| \right)^2 \qquad (6.40)$$

als Verhältnis zwischen mittleren quadratischen Fehler (MSE) und potentiellen Fehler multipliziert mit N und von eins subtrahiert geschrieben werden.

$$\mathrm{d} = 1 - N \frac{MSE}{PE} \qquad (6.41)$$

So wie auch der *Coefficient of Efficiency* E ist der Index of Agreement d sensitiv Gegen über großen Fehlern in Beobachtungs- und Vorhersagedaten. Nach Legates und McCabe Jr. (1999) sollen daher zur Verifikation mehrere Scores herangezogen werden.