Konstruktion optimaler stabiler numerischer Filter

Hans v. Storch

21.9.77

VORBEMERKUNG

Seit dem Experiment von PHILLIPS (1959) ist bekannt, daß bei der Simulation des Verhaltens der Atmosphäre mithilfe von Differenzenapproximationen zur Sicherstellung der Stabilität diejenigen kleinskaligen Prozesses, deren Scale an der Untergrenze ,der Auflösungsfähigkeit des Modells liegt, gedämpft oder ganz herausgenommen ,werden müssen. Dies liegt zum einen am "aliasing"-Effekt, dem man allerdings durch geeignete Differenzenapproximationen begegnen kann (ARAKAWA, 1970), zum anderen an der Aufrauhung der meteorologischen Felder infolge von Inbalancen der Felder Randeffekten (ROECKNER, 1976) und der sich Die Notwendigkeit der Herausfilterung kleinräumiger Prozesse entsteht auch bei anderen numerischen Modellen, so etwa bei der Methode der Finiten Elemente (CULLEN, 1976) oder bei der pseudospectral method (MERILEES, 1974)

Dieser „Filtervorgang“ kann auf verschiedene Weise realisiert werden:

(1) Einfügung eines Diffusionstermes. Bei CULLEN(1976) und SHAPIRO (1971) werden verschiedene Ansätze augelistet.

(2) Chopping-method: Bei diesem Verfahren werden die Felder nach Orthogonalfunktionen entwickelt. Bei der anschließenden Rekombination werden diejenigen Anteile, die kleinskaligen Prozesses darstellen, nicht mitgeführt.
In Modellen, die die Kugel oder die Halbkugel als Geometrie unterlegt haben, kann dies etwa dadurch realisiert werden, daß man für jeden in der Rechnung mitgeführten Breitenkreis eine FOURIER-Analyse durchführt und bei der anschließenden Rekombination nur jene Wellen mit Wellenlängen, die größer gleich 4 Gitterabständen sind, berücksichtigt werden. (PHILLIPS , 1959; WILLIAMSON, 1974; MERILEES, 1974)

(3) Numerische Filterung: In diesem Verfahren wird jeder Wert eines Feldes durch ein gewichtetes Mittel seiner Nachbarn ersetzt (genaue Definition siehe unten). Der erste numerische Filter wurde 1958 von SHUMAN vorgestellt, der dann 1962 von WALLINGTON in einem Simulationsprogramm verwandt wurde. Eine wichtige Klasse, von numerischen Filtern wurde 1970 von SHAPIRO veröffentlicht. Diese SHAPIRO-Filter wurden u.a. von CULLEN (1976) und FRANCIS (1975) benutzt

Die Methoden (1) und (3) besitz den gegenüber der chopping-method drei Vorteile:

(a) die Verfahren sind weniger rechenzeitintensiv, d.h. ökonomisch günstiger: Liegen auf einem Breitenkreis m Gitterpunkte, so benötigen die Verfahren(1) und (3) 0(m) Multiplikationen, während (2) $0(m^2)$ braucht.

(b) Das Verfahren (2) wirkt nicht lokal im Gegensatz zu den beiden anderen, d.h. hat man ein in einem ansonsten glatten Feld eine räumliche begrenzte kleine Störung, so verändert (2) das ganze Feld, während (1) und (3) nur zu Änderungen in unmittelbarer Nähe der Störung führen.

(c) Für den Fall der Kugel und der zonalen FOURIER-Analyse gilt, da? nur zonale Störungen erfaßt werden, nicht aber meridionale Irregularitäten.

Die Methode der Diffusion hat den Vorteil, daß man sie als Parametrisierung des Energieflusses in den sub-grid -Bereich und Modellierung der Horizontaldiffusion der Atmosphäre interpretieren kann. Andererseits ist sie oft nicht selektiv genug, d.h. beeinflußt in zu starkem Maße auch die großräumigen Prozesse. Ein weiterer Nachteil besteht darin, daß der Diffusionsterm als fester Bestandteil der beschreibenden Differentialgleichungen zu jedem Zeitschritt berechnet werden muß.

Dagegen kann die Methode der numerischen Filterung beliebig selektiv gestaltet werden (SHAPIRO, 1970) auf Kosten der Ökonomie und braucht nur ab und zu und nicht zu jedem Zeitschritt durchgeführt zu werden.

In der Diskussion, welche der Methoden (1) und (3) besser ist, ist aber noch nicht das letzte Wort gesprochen.

Bei der numerischen Filterung muß sichergestellt werden, dass sie die Konvergenz des Differenzenverfahrens nicht beeinträchtigt. Dies wird dann der Fall sein, wenn der Filteroperator eine konsistente und stabile Differenzenapproximation der Identität ist. Die Konsistenz ist dabei unproblematisch; anders sieht es bei der Stabilität aus: diese liegt genau dann vor, wenn keine Wellenkomponente verstärkt wird. Solche Filter sind in der Vergangenheit z.T. unsystematisch gefunden worden: SHUMAN (1958)t, SHAPIRO (1970, 1975)

In dieser Arbeit wird nun eine Methode entwickelt, bei vorgegebenem Rechenumfang "optimale" stabile numerische Filter zu konstruieren. Dies geschieht, indem die Konstruktionsaufgabe als (nichtlineare) Optimierungsaufgabe mit linearen Restriktionen aufgefaßt wird. Diese Optimierungsaufgabe kann dann mit bekannten Verfahren der angewandten Mathematik näherungsweise gelöst werden.

Schon früher ist die Konstruktionsaufgabe als Problem der Approximation, also der Optimierung ohne Restriktionen, aufgefaßt worden. Die so konstruierten Filter, die im Übrigen auch für andere Zwecke gedacht waren, sind leider meist nicht stabil, d.h. es gibt ein "Overshooting". Die früheste mir bekannte Veröffentlichung stammt von BLECK (1965). Unabhängig davon wurde dieser Ansatz 1967 von GALLI und RANDI mit einer nachfolgenden Korrektur von ZELEI (1971) in der geophysikalischen Literatur veröffentlicht.

MATHEMATISCHE PRÄZISIERUNG
==========

Definition "Numerischer Filter nach trigonometrischen Funktionen"

Wir betrachten' eine, Zerlegung des Intervall $(0,2\pi)$ in M (M gerade) äquidistante Teile der Länge $\Delta x := 2\pi/M$ und die Menge D der Funktionen, die auf dem Gitternetz $\{i\Delta x; i=1,..,M\}$ erklärt, sind.
Es sei $m \in \mathfrak{I}$ und $a_o,... \in \mathfrak{R}$. Dann heißt die Abbildung

[1] $$T:D \longrightarrow D;\ (T(f))(x) := a_o f(x) + \sum_{j=1}^{m} a_j \left[f(x+j\Delta x)+f(x-j\Delta x)\right]$$

numerische Filter nach trigonometrischen Funktionen.
Die Zahl m heißt Trägerlänge, die Zahlen $a_o,\cdots a_m$ Filtergewichte

Die Anwendung der numerischen Filter ergibt sich aus dem folgenden Sachverhalt:
Bekanntlich kann man jede Funktion $f \in D$ als Fourier-Summe darstellen:

$$f(x) = ß_o + \sum_{j=1}^{M/2} ß_j^s \sin(jx) + ß_j^c \cos(jx)$$

mit $\beta_{M/2}^S = 0$. Zu jedem numerischen Filter T gibt es nun Zahlen $\alpha(n)$, n=0,..M/2, M/2, sodaß gilt

$$(T(f))(x) = \alpha(0) ß_o + \sum_{j=1}^{M/2} \alpha(j)\left[ß_j^s \sin(jx) + ß_j^c \cos(jx)\right]$$

Die Ursache dafür besteht darin, daß, für alle reellen Zahlen $n \in \mathfrak{R}$, die Funktion $x \to \exp(inx)$ Eigenfunktion von T zum Eigenwert

[2] $$\alpha(n) = a_o + 2\sum_{j=1}^{m} a_j \cos(nj\tfrac{2\pi}{M})$$

ist. Diese "Eigenfunktionseigenschaft" bedeutet, daß der Filterprozess die Frequenz einer Welle nicht beeinflußt. Nur die Amplitude der Komponente mit Wellenzahl n wird um den Faktor $\alpha(n)$ verstärkt oder, vermindert.
Die Funktion α heißt response-function. Jeder Filter ist durch seine response-function eindeutig bestimmt.

Die Filtermethode innerhalb eines Simulationsprogramms läuft folgendermaßen ab:

Der Übergang vom Zeitpunkt t zum nächsten Rechenzeit punkt t+ Δt werde durch das Differenzenschema C (Δt) realisiert, d.h. die vorherzusagende Größe u (in der Regel ein Vektor, der die drei Geschwindigkeitskomponenten, das Geopotential, die Temperatur etc. enthält) ergibt sich aus der Rechenvorschrift

$$u(t+\Delta t) = C(\Delta t)\, u(t)$$

Dabei sei $C(\Delta t)$ eine konsistente und stabile Differenzanapproximation der den physikalischen Prozess beschreibenden Differentialgleichung. Sofern nicht gefiltert wird, beschreibt sich der Vorhersageprozeß durch die sukzessive Hintereinanderausführung der Vorschrift $C(\Delta t)$:

$$\prod C(\Delta t)$$

Beim Arbeiten mit numerischen Filtern wird alle paar Zeitschritte-,(etwa alle r) einmal der Filteroparator angewandt, d.h. man hat als Vorhersageprozeß

[3] $$\prod (T \circ C^r)$$

wobei der Kreis in $(T,o\ C^r)$ sagt, daß man auf das Anfangsfeld zunächst r mal den Differenzenoperator $C(\Delta t)$ und dann einmal den Filteroperator T anwenden soll.

Um die Konvergenz des neues Differenzenverfahrens sicherzustellen, muß $(T\ 0\ C^r)$ eine konsistente und stabile (bzgl. der L_2-oder Energienorm) Differenzenapproximation sein.[1])
Es gilt der

Satz $(T \circ C^r)$ ist konsistent, wenn gilt $\alpha\ (O) = 1$
$(T\ 0\ C^r)$ ist stabil, wenn gilt $\max\limits_{n=0,1..M/2} |\alpha(n)| \leq 1$

Beweis: T ist ein linearer beschränkter Operator, für dessen Norm gilt

$$||T|| = \max|\alpha(n)|$$

Demnach ist T stabil, wenn $|\alpha(n)|$ die 1 nicht übersteigt.

---

1) Zu den Begriffen "konsistent", "stabil" siehe etwa Kreiss/Oliger.

Zur Konsistenz: Da gilt

$$\|u(t+r\Delta t) - (T\circ C^r)u(t)\|$$
$$\leq \|u(t+r\Delta t) - T(u)(t+r\Delta t)\| + \|T(u(t+r\Delta t)) - T(C^r(u(t)))\|$$
$$\leq \|u(t+r\Delta t) - T(u)(t+r\Delta t)\| + \|T\|\cdot\|u(t+r\Delta t) - C^r(u(t))\|$$

und C^r konsistent ist, braucht nur noch gezeigt werden, daß gilt

$$\lim_{\Delta t\to 0} \frac{1}{\Delta t}\|u - Tu\| = \lim_{\Delta t\to 0} \frac{1}{\Delta t}\left[\int_0^{2\bar{u}} [u(x) - T(u)(x)]^2\,dx\right]^{1/2} = 0$$

Dazu nehmen wir an, dass gilt $\Delta t/\Delta x = \lambda$. Eine Taylorentwicklung liefert:

$$\|u-T(u)\| \leq \int_0^{2\bar{u}} \left[(1-a_o-2\sum_{j=1}^{m} a_j)u(x) + 2(\Delta x)^2 \sum_{j=1}^{m} j^2 a_j \max\left|\frac{\partial^2 u}{\partial x^2}\right|\right]^2 dx\Big]^{1/2}$$
$$\leq (1-\alpha(0))\|u\| + (\Delta x)^2 C$$

mit einer geeignet zu definierenden Konstanten C.
Gilt $\alpha(0) = 1$ bleibt:

$$\|u - T(u)\| \leq C\lambda^2 (\Delta t)^2$$ [Beweis Ende]

DIE FILTERKONSTRUKTIONSAUFGABE
==========

Die Filterkonstruktionsaufgabe besteht nun darin, Zahlen $a_o,..$ a_m zu finden, sodaß für die zugehörige response-function gilt:

[4] $\alpha(0) = 1$
[5] $0 \leq \alpha(n) \leq 1$ für $n=1,...M/2-1$
[6] $\alpha(M/2) = 0$
[7] $\sum_{n=1}^{M/2} (\alpha(n) - 1)^2 \stackrel{!}{=} \min$

Die Forderungen [4] bis [7]kann man so interpretieren. Die kleinstskalige Kaponente, also die $2\Delta x$-Welle, wird ganz herausgenommen ([6]), der Mittelwert bleibt unverändert ([4]), keine Welle darf verstärkt werden, keine Phase darf verändert werden ([5]). [7] stellt sicher, daß bei dem ganzen Prozeß möglichst

wenig verändert wird.
Ein Filter, der diesen Bedingungen genügt, wird als optimaler (wegen [7]), diskreter (weil er nur die ersten M/2 Frequenzen berücksichtigt) numerischer Filter nach trigonometrischen Funktionen bezeichnet.

DER FALL m=2

Die Aufgabe, einen optimalen Filter der Trägerlänge 2 zu konstruieren, ist recht einfach. Zunächst ist es möglich vermöge Der Restriktionen [4] und [6] die Variablen a_o und a_1 aus der Darstellung [2] der Response-Funktion αx zu eliminieren, sodaß man nur noch a_2 als. . zu bestimmende Variable zurück behält.

Die Restriktion [4] bedeutet:

$$a_o + 2\sum_{j=1}^{m} a_j = 1$$

und [6]

$$a_o + 2\sum_{j=1}^{m} (-1)^j a_j = 0$$

Dies sind zwei Gleichungen in a_o und a_1. Die Lösung des Systems lautet:

[8] $$a_o = 1/2 - \sum_{j=2}^{m} \left(1 + (-1)^j \right) a_j$$

[9] $$a_1 = 1/4 + \frac{1}{2}\sum_{j=2}^{m} \left((-1)^j - 1 \right) a_j$$

Man sieht, daß im Falle m =2 gilt

$a_o = 1/2 - 2a_2$

$a_1 = 1/4$

Setzt man die Ausdrücke [8] und [9]. in [2] ein, so erhält man

$$\alpha(n) = \frac{1}{2} + \frac{1}{2}\cos(n\Delta x) + \ldots$$

$$\ldots \sum_{j=2}^{m} \left[2\cos(nj\Delta x) + ((-1)^j - 1)\cos(n\Delta x) - \ldots \right.$$

$$\left. \ldots (1 + (-1)^j) \right] a_j$$

Für den Fall m = 2 erhalten wir:

[10] $$\alpha(n) = \frac{1}{2} + \frac{1}{2}\cos(n\Delta x) + \left[2\cos(2n\Delta x) - 2\right] a_2$$

$$\frac{-\frac{1}{2} - \frac{1}{2}\cos(n\,x)}{2\cos(2n\Delta x) - 2} \geqq a_2 \geqq \frac{\frac{1}{2} - \frac{1}{2}\cos(n\Delta x)}{\cos(2n\Delta x) - 2} \quad \text{für } n = 1, \ldots \frac{M}{2} - 1$$

(unter Berücksichtigung von $2\cos(2n\Delta x) - 2 < 0$ für die betrachteten Zahlen n)

Für den Fall M = 64 ergeben sich als Schranken für a_2:

untere Schranke: o.o62 65 = U
obere Schranke: o.o62 65 = O

Bis jetzt haben wir erreicht: Wählen wir a_o und a_1 gemäß [8] und [9] J und ist a_2 aus dem Intervall(untere Schranke, obere Schranke), so sind alle Restriktionen erfüllt. Es bleibt die Minimierungsaufgabe [7]:

$$[11]\quad \Phi(a) := \sum_{j=1}^{M/2-1} (\alpha(n) - 1)^2 \stackrel{!}{=} \min$$

Setzt man

$$f_n := -\frac{1}{2} + \frac{1}{2}\cos(n\Delta x)$$

$$g_n := 2\cos(2n\Delta x) - 2$$

so ist

$$[12]\quad \Phi(a) = \sum_{j=1}^{M/2-1} f_n^2 + 2f_n g_n a + g_n^2 a^2$$

$$= Ra^2 + Sa + T$$

wobei wir der Kürze halber a statt a_2 geschrieben haben. Die Größen R und S:

$$R := \sum g_n^2 \qquad S := 2\sum f_n g_n \qquad T := \sum f_n^2$$

Dies ist ein Parabel, die nach oben geöffnet ist ($R > \sigma$). Sie nimmt ihr Minimum auf dem Intervall (untere Schranke; obere Schranke) entweder an den Intervallenden an oder in ihrem Scheitelpunkt -S/2R an, sofern er zum zulässigen Intervall gehört.

Die Lösung der Aufgabe besteht demnach darin, daß man prüft, ob sich der Scheitelpunkt innerhalb des zulässigen Bereichs befindet. Ist dies der Fall, so setzt man

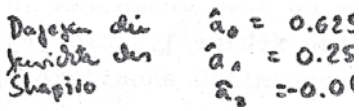
$$[13]\qquad a_2 := -\frac{1}{2}\,\frac{S}{R}$$

andernfalls prüft man, welcher der beiden Intervallenden den kleineren Wert Ø(a) liefert und definiert a_2 dann entweder als den Wert der oberen oder der unteren Intervallgrenze.

Im Falle M = 64 liegt der Scheitelpunkt bei-1.666 7, also außerhalb des zulässigen Intervalls.

Das Funktional 0 nimmt am unteren Intervallende den Wert-3.2560, am oberen den Wert den Wert 4.7633 an (abzüglich der Konstanten T). Als Koeffizienten erhalten wir so:

a_0 = o.625 30
a_1 = o.250 30
a_2 = o.062 30



Wir vergleichen nun diesen konstruierten Filter mit dem von SHAPIRO(1970,1975)angegebenen. Es zeigt sich, dass die respone-Funktion des konstruierten Filters überall oberhalb von der des entsprechenden SHAPIRO-Filters verläuft. Die Verbesserung beträgt allerdings nur bis zu 1%o. (siehe Tabelle)

Vergleich des konstruierten Filters mit dem entsprechenden SHAPIRO-Filter

N	ALPHA (N)	FEHLER	SHAPIRO	FEHLER	AL-SH
1	1.000000	.0000	.999994	.0000	.000
2	.999931	.0001	.999908	.0001	.000
3	.999587	.0004	.999536	.0005	.000
4	.998640	.0014	.998551	.0014	.000
5	.996648	.0034	.996514	.0035	.000
6	.993086	.0069	.992899	.0071	.000
7	.987362	.0126	.987119	.0129	.000
8	.978855	.0211	.978553	.0214	.000
9	.966943	.0331	.966583	.0334	.000
10	.951038	.0490	.950621	.0494	.000
11	.930614	.0694	.930145	.0699	.000
12	.905245	.0948	.904730	.0953	.001
13	.874629	.1254	.874076	.1259	.001
14	.838611	.1614	.838030	.1620	.001
15	.797204	.2028	.796607	.2034	.001
16	.750603	.2494	.750000	.2500	.001
17	.699187	.3008	.698590	.3014	.001
18	.643520	.3565	.642940	.3571	.001
19	.584344	.4157	.583791	.4162	.001
20	.522562	.4774	.522047	.4780	.001
21	.459217	.5408	.458748	.5413	.000
22	.395467	.6045	.395050	.6049	.000
23	.332550	.6674	.332190	.6678	.000
24	.271748	.7283	.271447	.7286	.000
25	.214351	.7856	.214108	.7859	.000
26	.161616	.8384	.161430	.8386	.000
27	.114727	.8853	.114593	.8854	.000
28	.074760	.9252	.074672	.9253	.000
29	.042647	.9574	.042596	.9574	.000
30	.019145	.9809	.019122	.9809	.000
31	.004815	.9952	.004809	.9952	.000

N Wellenzahl
Alpha(n) response-Funktion für die Wellenzahl N des - konstruierten Filterts
Fehler Abweichung von Alpha(n) von der 1
Shapiro response-Funktion für die Wellenzahl N des SHAPIRO-Filters
Fehler Abweichung des SHAPIRO-Filters von der 1
AL-SH Differenz Alpha(N)-SHAPIRO, d.h. Verbesserung gegenüber dem SHAPIRO-Filter

Der Fall $m \geq 3$

Die Aufgabe, einen optimalen Filter mit einer Trägerlänge von mindestens 3 zu konstruieren, ist nicht mehr so einfach. Wir gehen auf die Darstellung [4] bis [7] zurück und formulieren diese um in eine "quadratische Optimierungsaufgabe" mit "linearen Restriktionen"[1]); d.h. wir gehen über zu der Aufgabe, einen Vektor $\mathfrak{a} = (a_o, .. a_m)$ zu finden, der ein geeignet zu definierendes Funktional

[14] $z(\mathfrak{a}) := \mathfrak{a}' W \mathfrak{a} + C' \mathfrak{a}$ "Zielfunktion"

minimiert unter den "Restriktionen"

[15] $A \mathfrak{a} \leqq b$

Dabei sind A und W Matrizen; A ist i.A. nicht quadratisch; C und b sind Vektoren. Das "$\leqq$"-Zeichen in [15] ist komponentenweise gemeint: sind $\varphi = (x_o, .. x_m)$ und $\psi = (y_o, .. y_m)$ so soll $\varphi \leqq \psi$ bedeuten, daß für i=o,..m gilt: $x_i \leqq y_i$.
Der Strich ' bei $\mathfrak{a}$ und C in [14] soll andeuten, daß mit dem transponierten Vektor zu arbeiten ist.

Der Vorteil der Überführung von [4]-[7] in [14]- [15] liegt darin, daß die Angewandte Mathematik Verfahren zur näherungsweisen Lösung von [14]- [15] anbietet.

Wir beginnen jetzt mit der Konstruktion der Matrizen W und A und der Vektoren C und b.
Dazu setzen wir

$$\beta_{ni} := (\beta_{no}, .. \beta_{nm})' \quad \text{mit} \quad \beta_{ni} := \begin{cases} 1 & \text{falls } i = 0 \\ 2\cos(ni\Delta x) & \text{sonst} \end{cases}$$

Dann gilt $\alpha(n) = \sum_{j=o}^{m} \beta_{nj} a_j = \beta_n' \mathfrak{a}$.

und aus [7] wird

[16] $$\sum_{n=1}^{M/2} (\alpha(n)-1)^2 = \mathfrak{a}' \underbrace{\left(\sum_{n=1}^{M/2} \beta_n \beta_n' \right)}_{=: W} \mathfrak{a} \underbrace{-2 \sum_{n=1}^{M/2} \beta_n'}_{=: C'} \mathfrak{a} + 1$$

---

[1]) zum Themenkreis „Optimierung" siehe etwa COLLATZ/WETTERLING

Für die Nebenbedingungen [15] erhalten wir:

[4] $\alpha(0) = 1 \quad \Leftrightarrow \quad \beta_0' \alpha = 1$

Da die Restriktionen als Ungleichungen zu schreiben sind

[17] $\Leftrightarrow \quad \beta_0' \alpha \leq 1 \quad \text{und} \quad -\beta_0' \alpha \leq -1$

[6] $\alpha(M/2) = 0 \quad \Leftrightarrow \quad \beta_{M/2}' \alpha = 0$

[18] $\Leftrightarrow \quad \beta_{M/2}' \alpha \leq 0 \quad \text{und} \quad -\beta_{M/2}' \alpha \leq 0$

[5] $0 \leq \alpha(n) \leq 1 \quad \Leftrightarrow \quad \beta_n' \alpha \geq 0 \quad \text{und} \quad \beta_n' \alpha \leq 1$

[19] $\Leftrightarrow \quad -\beta_n' \alpha \leq 0 \quad \text{und} \quad \beta_n' \alpha \leq 1$

Die Restriktionen [17]-[19]lassen sich schreiben als

$$A \alpha \leq b$$

wenn man setzt:

$$A := \begin{bmatrix} \beta_{oo} & \beta_{o1} & \cdots & \beta_{om} \\ -\beta_{oo} & -\beta_{o1} & \cdots & -\beta_{om} \\ \beta_{1o} & \beta_{11} & \cdots & \beta_{1m} \\ -\beta_{1o} & -\beta_{11} & \cdots & -\beta_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{M/2-1,o} & \beta_{M/2-1,1} & \cdots & \beta_{M/2-1,m} \\ -\beta_{M/2-1,o} & -\beta_{M/2-1,1} & \cdots & -\beta_{M/2-1,m} \\ \beta_{M/2,o} & \beta_{M/2,1} & \cdots & \beta_{M/2,m} \\ -\beta_{M/2,o} & -\beta_{M/2,1} & & -\beta_{M/2,m} \end{bmatrix} \qquad b := \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Die Aufgabe wurde mit dem bei ECKHARDT angegebenen Verfahren von CRYER näherungsweise gelöst. Da dabei verwandte Programm wurde von Klaus Roleff, Institut für Angewandte Mathematik der Universität Hamburg, übernommen.
Die Durchführung der Rechnung nimmt relativ viel Rechenzeit in Anspruch. Dabei geht im Wesentlichen die Anzahl der Restriktionen, weniger die Anzahl der Variablen ein. Daher wurden bei den folgenden Beispielen nur jede 2. bzw. 4. Frequenz explizit in den Restriktionen mitgeführt. Es zeigte sich, dass die nicht berücksichtigten Frequenzen die Restriktionen trotzdem erfüllten.

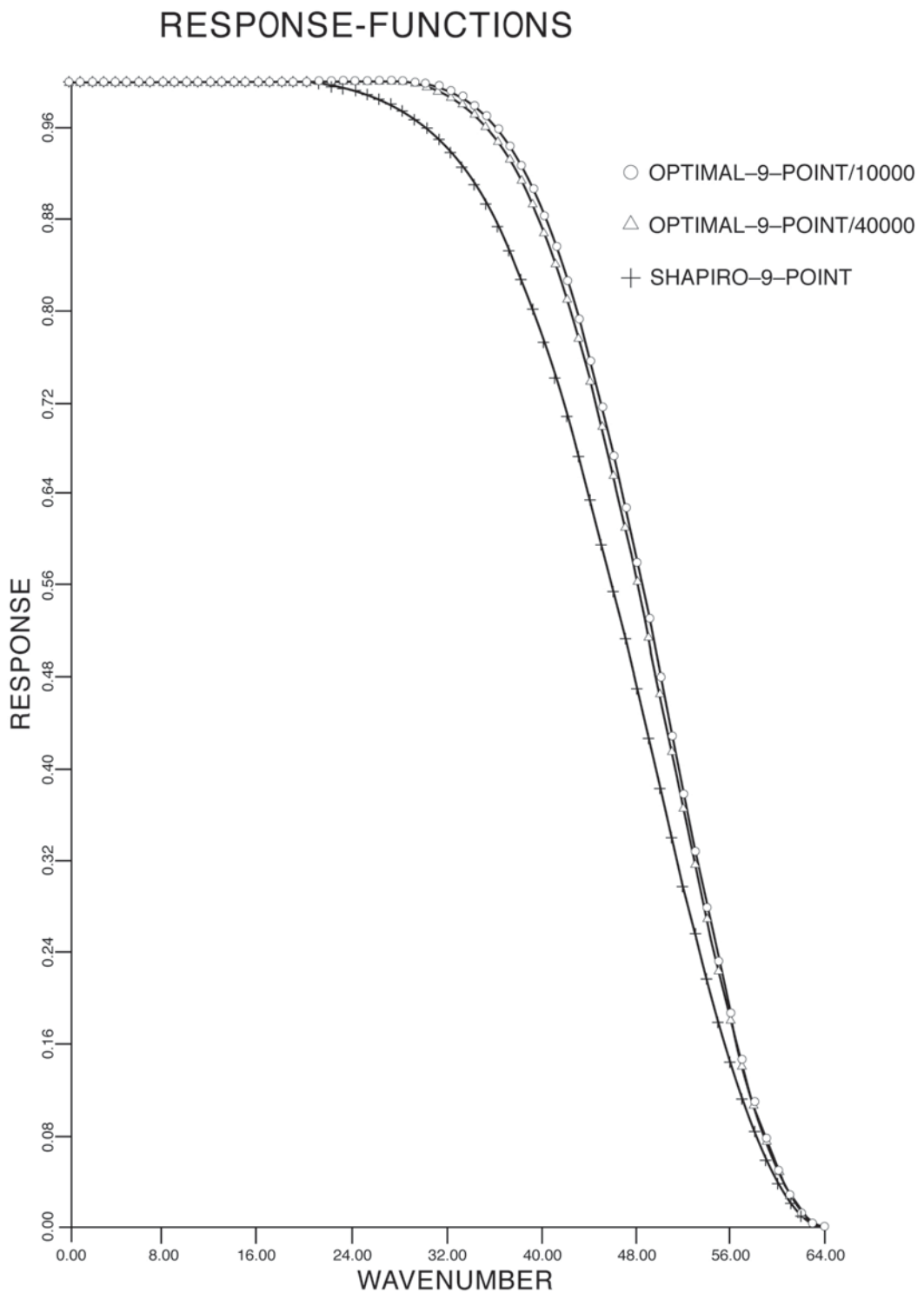
Wie schon erwähnt, ist das benutzte Verfahren ein Näherungsverfahren, liefert also i.A. nicht die optimale Lösung sondern nur eine Lösung in der Nähe des Optimums. Je länger man rechnet, d.h. je mehr Iterationen man rechnet, umso näher kommt man an die wahre Lösung.
Für den Fall M = 32 (zur Erinnerung: M gibt die Anzahl der Δx-Intervalle an) und m = 4 (also je 4 Nachbarn rechts und links werden zur Mittelbildung herangezogen) wurde eine Rechnung mit 10 000 und eine mit 40 000 Iterationen durchgeführt. Dabei ergaben sich als Gewichte (die Gewichte des entsprechenden SHAPIRO-Filters wurden wieder dazu geschrieben)

Gewichte

n	10 000 Iterat.	40 000 Iterat.	Shapiro
0	o.767 194 62o	o.761 8o4 339	o.726 562 ooo
1	o.199 102 75o	o.202 o8o 218	o.218 750 ooo
2	-o.122 877 59o	-o.121 5o6 8o5	-o.1o9 375 ooo
3	o.o5o 55o o13	o.047 788 626	o.o31 250 ooo
4	-o.o11 o63 128	-o.oo9 526 418	-o.oo3 906 2oo

Die response-Funktionen für die drei Filter
sind auf der folgenden Seite aufgetragen für 128 Stützpunkte. Im Falle der 10 000 Iterationen wird die 1 um höchstens o.ooo6 überschritten. Im Bereich der Wellenzahlen 0 bis 16 ist die Maximalabweichung von der 1 o.oo14.
Im Falle der 40 000 Iterationen wird die 1 um maximal o.oooo2 überschritten. Im Bereich der Wellenzahlen 0 bis 16 ist die Maximalabweichung von 1 o.ooo95.

Aus der Zeichnung sieht man, daß die konstruierten "optimalen" besser sind als der entsprechende von Shapiro.

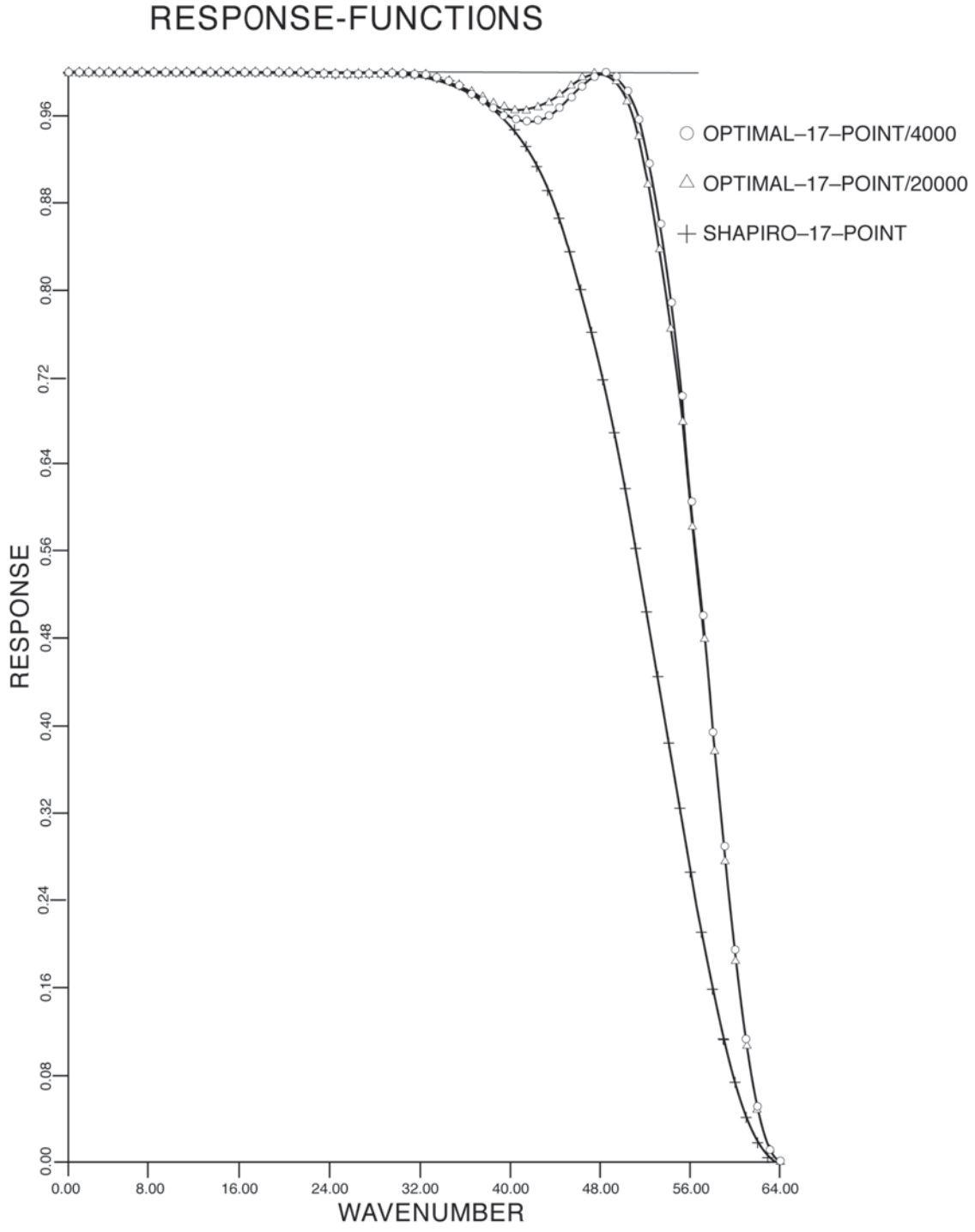

RESPONSE-FUNCTIONS
OPTIMAL–9–POINT/10000
OPTIMAL–9–POINT/40000
SHAPIRO–9–POINT
RESPONSE
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
0.00
8.00
16.00
24.00
32.00
40.00
48.00
56.00
64.00
WAVENUMBER

Der Fall M = 32 und m = 8 ergibt folgende Resultate

Gewichte

n	4000 Iterationen	20 000 Iterationen	Shapiro
0	o.882 64o 853	o.88o 859 662	o.8o3 619 o
1	o.1o9 746 171	o.112 137 85o	o.174 56o 5
2	-o.o93 246 934	-o.o96 o15 222	-o.122 192 3
3	o.o76 584 o36	o.o78 539 667	o.o66 65o 3
4	-o.o62 636 148	-o.o62 748 978	-o.o27 77o 9
5	o.o48 378 397	o.o46 398 499	o.oo8 544 9
6	-o.o31 591 336	-o.o28 619 579	-o.oo1 831
7	o.o15 094 762	o.o12 77o 449	o.ooo 244 1
8	-o.oo4 043 787	-o.oo3 199 785	-o.ooo o15 2

Im Falle der 4 000 Iterationen übersteigt die response-Funktion die 1 um maximal o.ooo15. Im Bereich der Wellenzahl 0 bis 16 ist die Maximalabweichung von der 1 nicht größer als o.ooo8.
Im Falle der 20 000 Iterationen gilt: Maximales overshooting: o.ooo18, maximale Abweichung von der 1:im Bereich der Wellenzahlen 0 bis 16: o.ooo62.
Auf der folgenden Seite sind die response-Funktionen der Filter aufgetragen für 128 Stützstellen.
Auch hier schienen die optimalen Filter dem Shapiro-Filter überlegen. Allerdings soll noch eine Rechnung mit 60 000 Iterationen gemacht werden, da anzunehmen ist, daß der "Hubbel" im Wellenzahlenbereich um 40 herum noch erheblich abgeschwächt werden kann.


RESPONSE-FUNCTIONS
OPTIMAL–17–POINT/4000
OPTIMAL–17–POINT/20000
SHAPIRO–17–POINT
RESPONSE
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
0.00
8.00
16.00
24.00
32.00
40.00
48.00
56.00
64.00
WAVENUMBER

Arakawa, A.
"Numerical simulation of large-scale atmosphere motion"
in: SIAM-AMS Proceedings 2 "Numerical solution of field problems in continuum physics", 24 - 40 (1970)

Bleck, R.

"Lineare Approximationsmethoden zur Bestimmung ein- und zweidimensionaler numerischer Filter der dynamischen Meteorologie", Diplomarbeit, Berlin (1965)

Collatz, L. / Wetterling, W.

"Optimierungsaufgabe", Springer Verlag Berlin Heidelberg New York (1971)

Cullen, M.J.P.

"On the use of artifical smoothing in Galerkin and finite difference solutions in the primitive equations"
Quart. J.R.Meto. Soc., 102, 77 - 93 (1976)

Eckhardt, U.

"Quadratic programming by successive overrelaxation"
Ber. KFA Jülich 1064 (1974)

Francis, P.E.

"The use of a multipoint filter as a dissipative mechanism in a numerical model of the general circulation of the atmosphere"
Quart. J. R. Met. Soc. 101, 567 - 582 (1975)

Galli, M. / L Randi. P.

"On the design of optimum numerical filters with a prefixed response", Ann. Geofis. 20, 401 - 414 (1967) .

Kreiss, H. / Oliger, J.

"Methods for the approximate solution of time dependet problems", WMO-ICSU GARP-.Publications Series No. 10 (1973)

Merilees, P.E.

"Numerical experiments with the pseudospectral method in spherical coordinates" GARP-Report 7, 215 - 264 (1974)

Phillps N.A.

"An example of nonlinear computational instability"
in: The atmosphere and the sea in motion, B. Bolin (ed.). The Rockefeller institute press, 501 - 504 (1959)

Roeckner, E.

"The control of noise in a general circulation model" in: simulation of large-scale atmospheric processes, 191 - 193 (1976)—

Shapiro R.

"Smoothing, filtering and boundary effects", Rev Geoph.8 (1970)

"The use of a linear filtering as a parameterization of the atmospheric diffusion", J. atm. Sc. 58., 523 - 531 (1971)
"Linear Filtering"
Math. Com. 29. 1094 - 1097 (1975)

Shuman, F.
"Numerical methods in weather prediction" Month. Wea. Rev 85, 357 - 361 (1958)

Wallington, C.E.

"The use of smoothing or filtering operators in numerical forecast", Quart. J. R. Met. Soc. 88, 470 - 484 (1962)

Williamson, D.

"Use of spectral filtering in a global circulation model"
GARP-Report 7, 265 - 292 (1974)

Zelei, A.

"On the design of numerical filters"
Ann. Geofis. 24, 457 - 474 (1971)