Positive und konservative Verfahren höherer Ordnung

Dissertation
zur
Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

im Fachbereich 17 Mathematik
der Universität Kassel

vorgelegt von
Philipp Andrea Zardo
aus
Kassel

Gutachter: Prof. Dr. Andreas Meister
Prof. Dr. Hans Burchard

Tag der Disputation: 26. Februar 2010

Inhaltsverzeichnis

Vorwort

Diese Arbeit habe ich während meiner Zeit als Wissenschaftlicher Bediensteter des Fachbereichs 17 Mathematik der Universität Kassel geschrieben.

Vor allem möchte ich Prof. Dr. Andreas Meister für die Betreuung danken, die mir an vielen Stellen immer wieder wichtige Impulse gab, und für sein persönliches Interesse an dem Thema und meiner Arbeit.

Prof. Dr. Hans Burchard danke ich sehr für sein Engagement und die konstruktive für mich sehr aufschlussreiche Kritik.

Sehr danke ich auch Dr. Joachim Benz für die Hilfe und Richtungsweisung im für mich völlig neuen Feld der ökologischen Modellierung sowie seinem Einsatz für das gemeinsame Projekt.

Zu nicht minderem Dank bin ich den Mitgliedern der Arbeitsgruppe verpflichtet für die unzähligen inhaltlichen Diskussionen und das sehr angenehme Umfeld. Beides hatte einen wichtigen Einfluss auf die Arbeit. Speziell hervorheben möchte ich Sigrun, die jeden Streit, in dem es um Erkenntnis ging, angenommen und bis zum Schluss geführt hat.

Abschließend gilt mein ganz besonderer Dank Katrin, die mir den Rücken frei gehalten und für die Lesbarkeit der Arbeit unschätzbare Dienste geleistet hat, und Livia, ohne die die Arbeit sicher noch nicht fertig wäre.

Einleitung

In der vorliegenden Arbeit sollen numerische Verfahren für Differentialgleichungen untersucht und weiterentwickelt werden, deren spezielle Anwendungsfelder in den Bereichen der ökologischen Seenmodelle und chemischen Reaktionen liegen. Beiden Feldern ist es gemeinsam, dass die beschriebenen Größen zumeist positiv sind und die Differentialgleichungssysteme in Teilen oder auch vollständig Masse erhaltende Reaktionen und Prozesse beschreiben.

Diese Eigenschaften, die Erhaltung der Positivität und die Erhaltung lokal konservativer Prozesse und globaler Konservativität, werden sich im Laufe der Arbeit als die bestimmenden Eigenschaften herausstellen, deren Respektierung von gängigen Verfahren für Differentialgleichungen in der konkreten Umsetzung nur eingeschränkt oder auch gar nicht garantiert werden.

In der Praxis sind häufig anzutreffende Vorgehensweisen zur Sicherstellung der Positivität der berechneten Näherungen z.B. das Abschneiden negativer Lösungen oder die Neuausführung mit geringeren Schrittweiten des vollzogenen Berechnungsschritts. Diese Vorgehensweisen zerstören entweder die Konservativität der Prozesse oder generieren im Grenzfall untragbaren Rechenaufwand.

Auf dem Gebiet der gewöhnlichen Differentialgleichungen ist in den letzten Jahren eine Serie von Veröffentlichungen [BDM03, BBKS07, BRBM07] zu Verfahren zu diesem Themenkomplex erschienen. Diese Arbeiten bauen mit unterschiedlichen Schwerpunkten eine Idee aus dem Buch [Pat80] von Patankar für ihre jeweiligen Verfahren aus, um die jeweils gewünschten Eigenschaften zu erhalten.

Im Gegensatz zur oben beschriebenen häufig vorgefundenen Praxis verwenden die erwähnten Verfahren problemangepasste Zeitschrittweitensteuerungen. Sie drosseln die verwendeten Zeitschritte je nach Verfahren mit einem Faktor für alle Reaktionen, mit einem Faktor für jede Reaktion oder mit einem Faktor für jede Gruppe von Reaktionen. Hierdurch garantieren sie die Positivität der Näherungen und gewährleisten eine (verfahrensabhängige) Form der Konservativität.

Im Kapitel 2 sind die wesentlichen theoretischen Ergebnisse zusammengetragen. Hier wird den oben beschriebenen Ansätzen eine gemeinsame formale Basis gegeben (Unterkapitel 2.1). Anschließend werden die theoretischen und praktischen Eigenschaften und Schwierigkeiten der Verfahren aus [BDM03, BBKS07] beschrieben (Unterkapitel 2.2 und 2.4.1). Abschließend werden zur Lösung der Schwierigkeiten alternative problemspezifische Strategien aufgezeigt und Verbesserungen vorgeschlagen. Konkret wird sowohl eine Variante der modifizierten Patankar Verfahren [BDM03] bewiesen, welche nicht mehr dem diesen Verfahren inhärenten Ordnungsmaximum von Zwei unterliegt (Unterkapitel 2.3), als auch eine Weiterentwicklung der modifizierten Patankar Verfahren, welche in der Lage ist, die von Bruggeman in [BBKS07] angeregte komplexere Definition der Konservativität zu genügen und steife Differentialgleichungen zu lösen (Unterkapitel 2.4.3).

Die gemachten theoretischen Aussagen werden anhand von praxisrelevanten und akademischen Beispielen aus den Bereichen der chemischen Reaktionen und der Ökologie illustriert und bestätigt. Die Vorstellung der Testfälle findet sich in Kapitel 1 und später auch in Unterkapitel 3.2. Die numerische Auswertung und Anwendung der Verfahren erfolgt in Kapitel 5 und Unterkapitel 6.1.

Ein zweiter Schwerpunkt der Arbeit besteht darin, das Verhalten dieser Verfahren im Kontext der Seenmodellierung zu untersuchen. Die Basis bildet das im Internet verfügbare „Null-D“ Modellierungswerkzeug „Ecobas“ (siehe [Ben09, AWB+96, Ben94, BGH99, BHG98, BHL01, HGB98, Str06, SBKB08]). Dies ist eine allgemeine Oberfläche. Sie wurde im Hinblick auf ökologische Anwendung konzipiert, ermöglicht es dem Benutzer aber grundsätzlich beliebige gewöhnliche Differentialgleichungssysteme zu definieren.

Wesentliche Optionen im Hinblick auf die ökologische Anwendung stellt Ecobas durch die angepassten Ausgaben von Graphiken der bestimmten Näherungen und der Dokumentation des aktuell betrachteten Systems inklusive aller verwendeten Rand- und Anfangsdaten mit Einheiten für alle Konstanten, Parameter und Differentialgleichungskomponenten bereit.

Ecobas löst die Systeme im Rahmen der gegebenen Möglichkeiten durch die Anwendung des im Internet auf [HP09] dokumentierten und frei verfügbaren ODE-Solver Pakets „LSODA“ als Blackbox, ohne weitere numerische Kenntnisse des Modellierers zu verlangen. Dies liefert für nicht steife Probleme sehr gute Ergebnisse. Für anspruchsvolle Differentialgleichungssysteme wie z.B. den Robertson Testfall oder das Orego Problem aus [Tes09] sind die Methoden als Black-

box nicht mehr in der Lage, zufriedenstellende Näherungen zu berechnen und liefern für diese Testfälle negative und damit unbrauchbare Ergebnisse. Wegen dieser Schwierigkeiten ist die Verwendung positivitätsgarantierender Verfahren von großem Interesse.

Ebenso ist es ein langfristiges Ziel, die Einschränkung von Ecobas auf gewöhnliche Differentialgleichungen zu beenden und eine Erweiterung auf spezielle Typen von partiellen Differentialgleichungssystemen zu erreichen, um eine größere physikalische Nähe der Modelle zu komplexen ökologischen Phänomenen zu erlauben.

Auf die direkte Implementierung eines Strömungslösers in Ecobas ist vorerst aber verzichtet worden. Bei einem solchen Ansatz wäre die Kontrolle der berechneten Ergebnisse und die direkte Einflussnahme auf Verfahrensdetails unnötig erschwert worden. Beides sind aber in der Entwicklung sehr wünschenswerte Arbeitsbedingungen.

Stattdessen wurde die im Bereich der Seenmodellierung für flache, nicht geschichtete Seen in natürlicher Weise auftretende, zweidimensionale Flachwassergleichung, beschrieben in Unterkapitel 3.1, mit einem realistischen ökologischen Modell für einen Phosphorzyklus gekoppelt (Unterkapitel 3.2). Das verwendete ökologische Modell ist eine Weiterentwicklung des Phosphorzyklus Modells von Hongping und Jianyi zum West Lake bei Hangzhou in China aus [HJ02].

Das so entstehende gemischt hyperbolisch parabolische System wurde mit einer Adaption des bewährten Strömungslösers Taucode (siehe z.B. [MS98, Mei94]) und den oben beschriebenen Zeitintegrationsverfahren gelöst. Der Strömungslöser wird in Kapitel 4 beschrieben.

Die numerischen Ergebnisse für das gesamte System aus Phosphorzyklus und zweidimensionaler Flachwassergleichung, die sich allerdings auf Grund mangelnder Messdaten auf strukturelle Testfälle beschränken, werden in Unterkapitel 6.3 präsentiert. Zusätzlich werden aber Vergleichsrechnungen für ein Dammbruchproblem aus [Tor01b] die Eignung des Strömungslösers für komplexe Strömungen demonstrieren. Die Ergebnisse finden sich in Unterkapitel 6.2. Ebenso werden für das Phosphormodell die Rechnungen mit den Ergebnissen des aktuellen Ecobas verglichen. Dies wird in Unterkapitel 6.1 vorgestellt.

Abschließend findet sich die Zusammenfassung und der Ausblick auf mögliche weitere Entwicklungsfelder in Kapitel 7.

Teil I

Über gewöhnliche Differentialgleichungen

Kapitel 1

Testfälle gewöhnlicher Differentialgleichungssysteme

Um die in Kapitel 2 vorgestellte Theorie anhand von Modellgleichungen motivieren zu können, werden im Folgenden einige autonome gewöhnliche Differentialgleichungssysteme vorgestellt. Um alle für die im Weiteren gemachten Untersuchungen benötigten Informationen zu den einzelnen Differentialgleichungssystemen an einer Stelle gesammelt zu haben, werden in diesem Kapitel Begriffe verwendet, die erst im Laufe des 2. Kapitels erklärt werden. Das Verständnis von nicht eingeführten Begriffen ist nicht notwendig und soll den Leser bei der ersten Lektüre nicht verwirren. Diese Art der Aufbereitung dient der Bündelung aller relevanten Informationen an einer Stelle in der Arbeit.

1.1 Ein einfacher linearer Fall

Das erste System (aus [BDM03]) ist ein akademischer Testfall, der überschaubar ist, da er eine analytische Lösung besitzt. Dieser wird genutzt, um die tatsächlichen numerischen Fehler der jeweiligen Verfahren exakt (im Rahmen der Rechengenauigkeit) zu bestimmen.

Beispiel 1.1 *Das System*

$$\boldsymbol{c}' = \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = \begin{pmatrix} c_2 - a\, c_1 \\ a\, c_1 - c_2 \end{pmatrix} \tag{1.1}$$

ist linear und beschreibt einen Masseaustausch zwischen c_1 und c_2. Die analytische

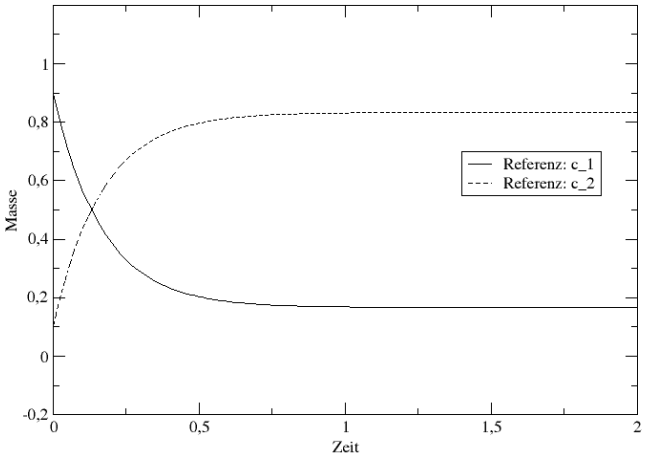

Abbildung 1.1: Die analytische Lösung zum System (1.1).

Lösung ist

$$c_1(t) = (1 + c\exp^{-(a+1)t})c_1^\infty$$

mit der asymptotischen Lösung für $t \to \infty$

$$c_1^\infty = \frac{c_1(0) + c_2(0)}{a+1} \text{ und } c = \frac{c_1(0)}{c_1^\infty} - 1.$$

Wegen der Konservativität nach Definition 2.36 und auch 2.12 ergibt sich

$$c_2(t) = c_1(0) + c_2(0) - c_1(t).$$

Für das Beispiel wird $a = 5$, $c_1(0) = 0.9$ *und* $c_2(0) = 0.1$ *gewählt, d.h. die zu konservierende Masse beträgt eins. Entsprechend sollte* $c_1(t) + c_2(t) = 1$ *für alle* t *gelten. Dies führt auf*

$$c_1^\infty = \frac{1}{6} \text{ und } c_2^\infty = \frac{5}{6}.$$

Eine graphische Darstellung der analytischen Lösung findet sich in Abbildung 1.1.

Zwei alternative Darstellungen von (1.1) sind gegeben durch

$$\boldsymbol{c}' = S_1\mathbf{r}_1 = S_2\mathbf{r}_2 \tag{1.2}$$

mit

$$S_1 = \begin{pmatrix} -a & 1 \\ a & -1 \end{pmatrix}, \ \mathbf{r}_1 = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \text{ und } S_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \ \mathbf{r}_2 = \begin{pmatrix} a\,c_1 \\ c_2 \end{pmatrix}.$$

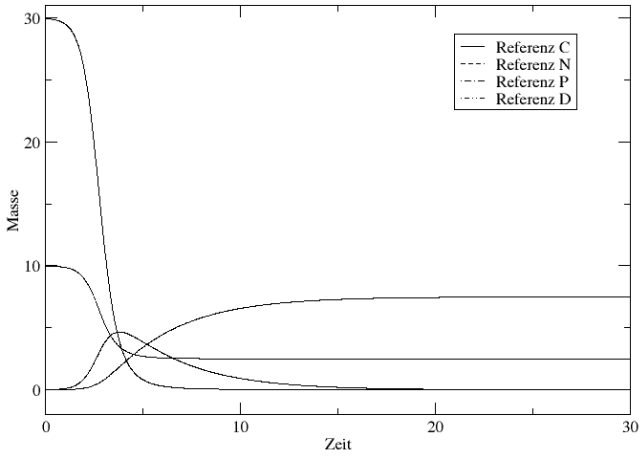

Abbildung 1.2: Die Referenzlösung zum System (1.3), erstellt mit dem Euler Verfahren und einer Schrittweite von $h = 10^{-5}$.

1.2 Ein einfaches biologisches Modell

Das zweite Beispiel (aus [BBKS07]) beschreibt ein einfaches biologisches Modell für die Nährstoffaufnahme und das Sterben von Phytoplankton. Eine wesentliche Eigenschaft ist die Konservativität auf atomarer Ebene nach Definition 2.36 aber nicht auf der Ebene der Masse nach Definition 2.12.

Beispiel 1.2 *Das System mit den Konstanten* $r_{\max}, K_C, K_N, a, b, e$

$$\boldsymbol{c}' = \begin{pmatrix} c_1' \\ c_2' \\ c_3' \\ c_4' \end{pmatrix} = \begin{pmatrix} C' \\ N' \\ P' \\ D' \end{pmatrix} = \begin{pmatrix} -a\, r_{\max} \dfrac{C}{K_C + C}\, \dfrac{N}{K_N + N}\, P \\ -b\, r_{\max} \dfrac{C}{K_C + C}\, \dfrac{N}{K_N + N}\, P \\ r_{\max} \dfrac{C}{K_C + C}\, \dfrac{N}{K_N + N}\, P - eP \\ eP \end{pmatrix} \tag{1.3}$$

beschreibt das Wachstum auf der Basis zweier Nährstoffe Stickstoff, N*, und Kohlenstoff,* C*, sowie das Absterben zu toter Masse,* D*, von Phytoplankton,* P*, in höheren Meeresschichten. Als Startwerte werden* $C(0) = 29.98$*,* $N(0) = 9.98$*,* $P(0) = 0.01$ *und* $D(0) = 0.01$ *verwendet. Die Konstanten werden wie folgt gewählt:* $r_{\max} = 100, K_C = 1000, K_N = 1, a = 4, b = 1, e = 0.3$*. Durch eine*

Einschränkung auf die drei Komponenten N, P und D und unter Vernachlässigung des Kohlenstoffs, C, würde man ein NPD konservatives Modell, bzw. eine Konservativität auf Ebene der Massen (Definition 2.12), erhalten. Da ein Ziel der Arbeit aber die Entwicklung von Verfahren ist, welche die komplexere Form der Konservativität, eben jene auf atomarer Ebene (Definition 2.36), erhalten, ist die genannte Form (1.3) gewählt worden.

Eine alternative Darstellung des Systems in der Form

$$\boldsymbol{c}' = S\boldsymbol{r}$$

wird in (2.27) detailliert ausgeführt. Mit der Matrix E wie in (2.28) und unter Berücksichtigung der gewählten Anfangsbedingungen und Konstanten für alle Zeiten t findet man die Invariant $E\,\mathbf{c}(t)$, die durch

$$E\,\mathbf{c}(t) = \begin{pmatrix} C(t) + 4P(t) + 4D(t) \\ N(t) + P(t) + D(t) \end{pmatrix} = \begin{pmatrix} 30.06 \\ 10 \end{pmatrix}$$

festgelegt ist.

Die Referenzlösung (Abbildung 1.2) wurde mit dem Euler Verfahren und einer festen Schrittweite von $h = 10^{-5}$ berechnet.

1.3 Das Orego Problem, Teil 1

Das dritte Beispiel ist ein sehr steifes, d.h. numerisch sehr anspruchsvolles, System aus [HWN02, Tes09] und wird in zwei Versionen betrachtet. Die ursprüngliche Version, hier zu erst beschrieben, hat lediglich positive Größen. Allerdings ist das System weder nach Definition 2.12 noch nach Definition 2.36 konservativ.

Für eine Beschreibung der zugrunde liegenden chemischen Zusammenhänge wird hier direkt die Quelle [Tes09] zitiert.

„The OREGO problem originates from the celebrated Belousov Zhabotinskii (BZ) reaction. When certain reactans, like bromous acid, bromide ion and cerium ion, are combined, they exhibit a chemical reaction which, after an induction period of inactivity, oscillates with change in structure and in color, from red to blue and viceversa. The color changes are caused by alternating oxidationreductions in which the cerium switches its oxidation state from Ce(III) to Ce(IV). Field, Koros and Noyes formulated the following model for the most important parts

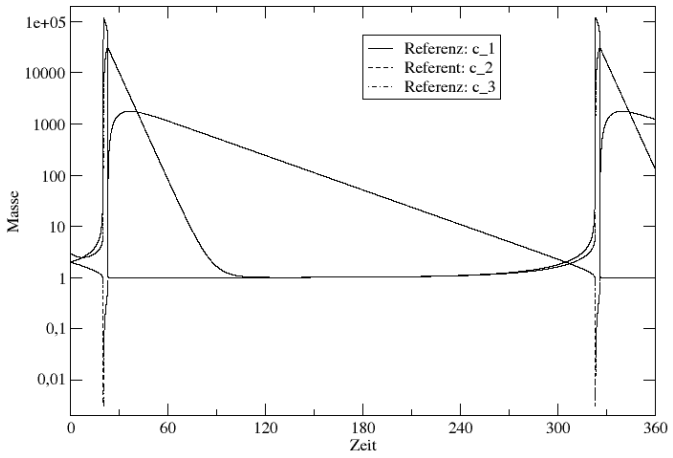

Abbildung 1.3: Die Referenzlösung zum System (1.4), erstellt mit dem vmPE Verfahren und einer Schrittweite von $h = 10^{-5}$.

of the kinetic mechanism that gives rice to oscillation in the BZ reaction. This mechanism can be summarized as three concurrent processes [Gra02]:

- the reduction of bromate (BrO_3^-) to bromine (Br) via the reducing agent bromide (Br^-). Bromomalonic acid (BrMA) is produced;
- the increase of hypobromous acid ($HBrO_2$) at an accelerating rate and the production of Ce(IV). Here we have a sudden change in color from red to blue;
- the reduction of Cerium catalyst Ce(IV) to Ce(III). Here we have a gradual change in color from blue to red.“

Beispiel 1.3 *Das System*

$$\boldsymbol{c}' = \begin{pmatrix} c_1' \\ c_2' \\ c_3' \end{pmatrix} = \begin{pmatrix} s(c_2 - c_1c_2 + c_1 - qc_1^2) \\ \frac{1}{s}(-c_2 - c_1c_2 + c_3) \\ w(c_1 - c_3) \end{pmatrix} \tag{1.4}$$

mit den Konstanten $s = 77.27, w = 0.161$ *und* $q = 8.375 \cdot 10^{-6}$ *beschreibt die Belousov Zhabotinskii Reaktion. Dabei entsprechen* c_1 $HBrO_2$, c_2 Br^- *und* c_3 *Ce(IV).*

Die Anfangswerte sind gegeben durch $c_1(0) = 1, c_2(0) = 2, c_3(0) = 3$. *Mit den Abkürzungen*

$$S = \begin{pmatrix} -s & s & 0 & -s & s \\ 0 & -\frac{1}{s} & \frac{1}{s} & -\frac{1}{s} & 0 \\ 0 & 0 & -w & 0 & w \end{pmatrix} \quad \textit{und } \mathbf{r} = \begin{pmatrix} qc_1^2 \\ c_2 \\ c_3 \\ c_1c_2 \\ c_1 \end{pmatrix} \tag{1.5}$$

ergibt sich $\boldsymbol{c}' = S\mathbf{r}$.

Als Referenz werden von den Autoren in [Tes09] die Massen der drei Komponenten c_1, c_2 und c_3 am Ende des Integrationsintervalls ($t_e = 360$) mit den nummerischen Werten

$$c_1^{\text{Bari}}(t_e) = 1.0008148, c_2^{\text{Bari}}(t_e) = 1228.1785, c_3^{\text{Bari}}(t_e) = 132.05549$$

angegeben. Um für weitere Zeitpunkte $0 < t < t_e$ Vergleichslösungen zur Verfügung zu haben, wurde mit dem vmPE Verfahren eine Referenzrechnung mit einer Schrittweite von $h = 10^{-5}$ durchgeführt (siehe Abbildung 1.3). Die korrespondierenden Werte der hier verwendeten Referenzlösung für den Zeitpunkt t_e sind

$$c_1(t_e) = 1.0008146, c_2(t_e) = 1228.2315, c_3(t_e) = 132.08593.$$

Man sieht eine Übereinstimmung von wenigstens vier Stellen, insgesamt ergibt sich ein gemittelter relativer Fehler

$$\frac{1}{3}\left(\frac{|c_1(t_e) - c_1^{\text{Bari}}(t_e)|}{c_1^{\text{Bari}}(t_e)} + \frac{|c_2(t_e) - c_2^{\text{Bari}}(t_e)|}{c_2^{\text{Bari}}(t_e)} + \frac{|c_3(t_e) - c_3^{\text{Bari}}(t_e)|}{c_3^{\text{Bari}}(t_e)}\right) = 9.1287 \cdot 10^{-5}.$$

Die erstellte Referenzlösung liefert also einen sehr hohen Grad an Übereinstimmung mit der Referenz der Autoren von [Tes09].

1.4 Das Orego Problem, Teil 2

Für theoretische Zwecke wird das System (1.4) um die Komponente c_4 erweitert. Diese Erweiterung wird so gewählt, dass das so entstehende System konservativ nach Definition 2.36 ist. Da die neue Größe c_4 keine physikalische Bedeutung hat, ist sie im Vorzeichen nicht festgelegt. Wie man später noch sieht, gilt sogar:

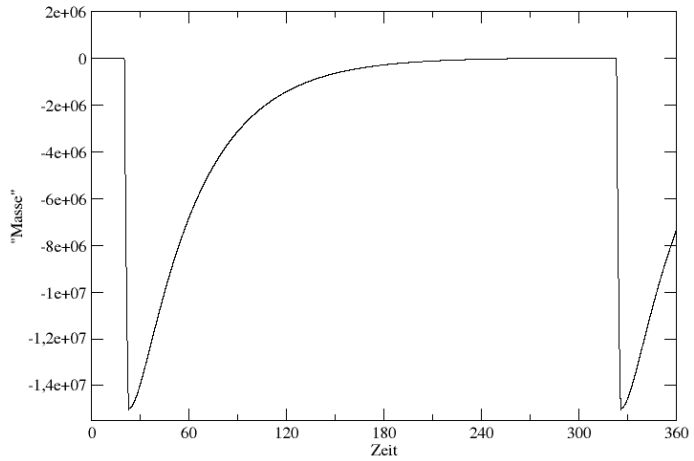

Abbildung 1.4: Die Referenzlösung zur Größe c_4 des Systems (1.6), erstellt mit dem vmPE Verfahren und einer Schrittweite von $h = 10^{-5}$.

Da durch die Erweiterung keine Veränderung in der verwendeten Schrittweite der numerischen Verfahren zu Problem (1.4) erzeugt werden soll, muss die vierte Größe uneingeschränkt im Vorzeichen sein.

Das so entstehende System ist eingeschränkt positiv nach Definition 2.4. In diesem Sinne ist dieses System das Anspruchsvollste der in diesem Kapitel vorgestellten.

Beispiel 1.4 *Verwendet man*

$$\boldsymbol{c}' = \begin{pmatrix} c_1' \\ c_2' \\ c_3' \\ c_4' \end{pmatrix} = \begin{pmatrix} s(c_2 - c_1c_2 + c_1 - qc_1^2) \\ \frac{1}{s}(-c_2 - c_1c_2 + c_3) \\ w(c_1 - c_3) \\ s(2c_1c_2 - 2c_1 + qc_1^2) \end{pmatrix} \tag{1.6}$$

mit den Konstanten und Anfangsbedingungen wie in (1.4), so erhält man ein konservatives System nach Definition 2.36. Als zusätzlicher Anfangswert wird $c_4(0) = 0$ *festgesetzt.*

Für die entsprechende Matrix

$$S = \begin{pmatrix} -s & s & 0 & -s & s \\ 0 & -\frac{1}{s} & \frac{1}{s} & -\frac{1}{s} & 0 \\ 0 & 0 & -w & 0 & w \\ s & 0 & 0 & 2s & -2s \end{pmatrix}$$

und den Vektor $\mathbf{r}$ *wie in (1.5) ergibt sich wiederum die alternative Darstellung*

$$\mathbf{c}' = S\,\mathbf{r}.$$

Mit dem Zeilenvektor $E = \left(1, s^2, \frac{s}{w}, 1\right)$ *sowie den Konstanten und den Anfangswerten für* c_1, c_2, c_3 *wie im Beispiel 1.3 ergibt sich für alle Zeiten* t *die Invariante*

$$E\,\mathbf{c}(t) = c_1 + s^2 c_2 + \frac{s}{w} c_3 + c_4 = 1 + 2s^2 + 3\frac{s}{w} \approx 13382.12. \tag{1.7}$$

Die Referenzlösung wurde wiederum mit dem vmPE Verfahren und einer festen Schrittweite von $h = 10^{-5}$ *erstellt.*

Da die Komponente c_4 *nicht Bestandteil des ursprünglichen Problems ist, gibt es hier keine Literaturwerte. Für* $t_e = 360$ *liefert die Konservativität aus (1.7) mit den Referenzwerten* $c_1^{Bari}(t_e)$, $c_2^{Bari}(t_e)$ *und* $c_3^{Bari}(t_e)$ *die Bedingung*

$$c_4(t_e) = -7383024.971.$$

Die Näherung der hier verwendeten Referenzlösung (Abbildung 1.4) ist $c_4(t_e) = -7383356$. *Also ergibt sich auch hier eine minimale Übereinstimmung von vier Stellen und ein relativer Fehler von* $4.4837 \cdot 10^{-5}$.

Kapitel 2

Numerische Verfahren für gewöhnliche Differentialgleichungen

In diesem Kapitel werden Verfahren und Verfahrensklassen für autonome gewöhnliche Differentialgleichungssysteme vorgestellt, die wesentliche anwendungsrelevante Nebenbedingungen (zusätzlich zur klassischen Fehlerordnung für ein gegebenes System von gewöhnlichen Differentialgleichungen) erfüllen. Es gab seit dem Jahr 2003 eine Reihe von Veröffentlichungen [BDM03, BBKS07, BRBM07] zum Thema der positiven und Masse erhaltenden Verfahren. Ein ausführlicher Vergleich der Verfahren aus [BDM03, BBKS07] findet sich in [Zar05].

Im Folgenden (Unterkapitel 2.1) wird ein Formalismus dargestellt, mit dem alle in den eben erwähnten Veröffentlichungen vorgestellten Verfahren bzw. Verfahrensklassen einen gemeinsamen Hintergrund bekommen. Anschließend (Unterkapitel 2.2) wird auf die modifizierten Patankar Verfahren aus [BDM03] und den ihnen eigenen Konservativitätsbegriff eingegangen. Diese werden beschrieben und ihre Einschränkungen aufgezeigt. Eine Modifikation des in Unterkapitel 2.2 vorgestellten modifizierten Euler Patankar Verfahrens von beliebiger Ordnung ist im Unterkapitel 2.3 zu finden.

Im letzten Unterkapitel (Unterkapitel 2.4) wird eine allgemeinere Definition der Konservativität präsentiert. Da die in [BBKS07] besprochenen Verfahren diese allgemeine Form der Konservativität respektieren, werden sie anschließend vorgestellt. Abschließend wird eine Erweiterung für die modifizierten Patankar Verfahren vorgeschlagen und diskutiert, welche ebenfalls diese allgemeinere Form der Konservativität erhält und steife Differentialgleichungen lösen kann.

2.1 Ordnungsanalyse für gestörte Runge Kutta Verfahren

Es erweist sich als zweckmäßig, in Problemstellungen der Biologie und Chemie die übliche allgemeine Form der rechten Seite eines Systems von gewöhnlichen autonomen Differentialgleichungen

$$c_i'(t) = f_i(\mathbf{c}(t)), i = 1, \ldots, N \tag{2.1}$$

zu präzisieren. Jedes System lässt sich alternativ zu (2.1) auch immer in der Form

$$\mathbf{c}'(t) = \mathbf{f}(\mathbf{c}(t)) = S\ \mathbf{r}(\mathbf{c}(t)) \tag{2.2}$$

darstellen. Vergleiche dazu die Darstellungen in (1.2) oder (1.5) zu den Systemen (1.1) respektive (1.4).

Definition 2.1 *Die zur Darstellung (2.2) gehörende Matrix $S \in \mathbb{R}^{N\times K}$ heißt* Stoichiometriematrix. *Sie beschreibt das Verhältnis der K Reaktionen zu den N Komponenten eines biochemischen Systems. Der Vektor $\boldsymbol{r}(\boldsymbol{c}) : G \to (\mathbb{R}^+)^K, G \subset \mathbb{R}^N$ heißt* Reaktionsvektor.

Für eine exemplarische Umsetzung der Begriffe der Stoichiometriematrix und des Reaktionsvektors siehe in Kapitel 1 und in der Einleitung zum Unterkapitel 2.4.

Da es als pathologisches Beispiel immer wieder von Interesse sein wird, wird folgende Terminologie verwendet.

Definition 2.2 *Das System $\mathbf{c}' = \mathbf{0}$ heißt* triviales System.

Folgendes wird in spätere Überlegungen immer wieder stillschweigend eingehen.

Bemerkung 2.3 *Weder die Matrix S noch der Vektor $\mathbf{r}$ aus der Darstellung (2.2) sind eindeutig. Speziell ist $\mathbf{r}$ beliebig und $S = \mathbf{0}$ eine Umsetzung des trivialen Systems. Ebenso wenig ist die Anzahl der Reaktionen K formal eindeutig.*

Von zentraler Bedeutung sind die folgenden Definitionen.

Definition 2.4 *Eine Größe c_i heißt* positiv, *wenn für $c_i(0) > 0$ und alle $t > 0$ gilt, dass $c_i(t) > 0$ ist. Ein System (2.2) heißt* positiv, *wenn alle Größen c_i positiv sind. Ein System heißt* eingeschränkt positiv, *wenn es sowohl positive als auch uneingeschränkte Größen c_i gibt.*

Ebenso von außerordentlicher Bedeutung ist der im weiteren Verlauf folgende und später allgemeiner bestimmte Begriff der Konservativität.

Definition 2.5 *Ein System (2.2) heißt* patankar-konservativ *genau dann, wenn für alle Reaktionen $r_k, k = 1, \ldots, K$ genau ein i und j existieren mit $S_{ik}, S_{jk} \neq 0$ und für diese beiden Elemente zudem $S_{ik} = -S_{jk}$ gilt.*

Dazu folgt ein Beispiel.

Beispiel 2.6 *1. Zuerst wird noch einmal der pathologische Fall betrachtet. Das triviale System ist patankar-konservativ. Wähle (in Ergänzung zu Bemerkung 2.3) z.B. $K = N$,*

$$S = \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix} \quad \textit{und } r_1 = r_2 = \ldots = r_N = 1.$$

2. Aus den expliziten Formen von S_1 und S_2 aus (1.2) wird ersichtlich, dass das System (1.1) patankar-konservativ ist.

3. Man erkennt auch, dass das System (1.4) mit der Darstellung (1.5) nicht patankar-konservativ ist.

Diese Form der Konservativität lässt sich physikalisch deuten.

Bemerkung 2.7 *In der Praxis wird die Patankar-Konservativität auch NPD Konservativität oder Massenkonservativität genannt, da sie von allen Komponenten des Systems verlangt, dass sie in derselben Einheit betrachtet werden.*

Um den Begriff Konservativität noch weitergehend zu motivieren, dient der folgende Satz.

Satz 2.8 *Gilt für ein System (2.2) mit stetig differenzierbarer Lösung* $\mathbf{c}$ *für alle $k = 1, \ldots, K$*

$$\sum_{i=1}^{N} S_{ik} = 0,$$

so folgt für die analytische Lösung

$$\sum_{i=0}^{N} (c_i(t+h) - c_i(t)) = 0$$

für $t, h > 0$.

Beweis: Mit dem Mittelwertsatz folgt für ein $\zeta \in [t, t+h]$

$$\frac{1}{h} \sum_{i=0}^{N} (c_i(t+h) - c_i(t)) = \sum_{i=0}^{N} c_i'(\zeta) = \sum_{i=0}^{N} \sum_{k=0}^{K} S_{ik} r_k(\zeta) = \sum_{k=0}^{K} r_k(\zeta) \underbrace{\sum_{i=0}^{N} S_{ik}}_{=0} = 0.$$

□

Die, durch diesen Satz definierte, Klasse von Problemen, lässt sich in patankar-konservative Form bringen.

Bemerkung 2.9 *Gilt für ein System*

$$\sum_{i=1}^{N} S_{ik} = 0$$

für alle k*, so kann man daraus ein patankar-konservatives System konstruieren. Alle Reaktionen* r_k*, für die die Bedingung aus Definition 2.5 nicht zutreffen, lassen sich additiv so zerlegen, dass die Bedingungen aus der Definition erfüllt sind. Dies erhöht die Anzahl* K *der Reaktionen. Zur Verdeutlichung kann das im späteren Kontext ausgeführte Beispiel 2.48 dienen.*

Wichtiger aber ist, dass dadurch eine konservative Behandlung durch die später vorgestellten verallgemeinerten modifizierten Patankar Verfahren (Unterkapitel 2.4.3) für konservative Systeme nach der auch noch folgenden Definition 2.36 verhindert wird.

An den angegebenen Stellen wird dies im Detail erklärt. Hier soll nur darauf hingewiesen werden, dass diese Systeme formal zwar gleich, aber in der möglichen numerischen Behandlung sehr wohl unterschiedlich sind.

Zur Vereinfachung von Verfahrens- und Eigenschaftenbeschreibungen bedient man sich der folgenden Definition.

Definition 2.10 *Für Einschrittverfahren wird die Notation von Henrici [Hen62]*

$$\boldsymbol{c}^{n+1} = \mathbf{c}^n + h\, \phi\left(h, \mathbf{c}^n, \mathbf{c}^{n+1}\right) \tag{2.3}$$

mit Verfahrensfunktion ϕ *und Zeitschritt* $h = t^{n+1} - t^n > 0$ *verwendet.*

Die Definitionen 2.4 und 2.5 für Systeme haben Entsprechungen für Verfahren.

Definition 2.11 *Ein Verfahren (2.3) heißt genau dann* positiv*, wenn, angewendet auf ein positives System (2.2), für alle* $n \in \mathbb{N}, h \in \mathbb{R}^+$ *und* $\boldsymbol{c}^n > 0$ *gilt, dass*

$$\boldsymbol{c}^{n+1} = \mathbf{c}^n + h\ \phi(h, \mathbf{c}^n, \mathbf{c}^{n+1}, t^n) > 0$$

ist.

Ein Verfahren heißt allgemein positiv*, wenn es, angewendet auf ein eingeschränkt positives System, die Positivität der positiven Größen bewahrt.*

Genauso grundlegend für die Gedanken zu den modifizierten Patankar Verfahren ist die Definition der entsprechenden Konservativität.

Definition 2.12 *Ein Verfahren (2.3) heißt* patankar-konservativ*, wenn, angewendet auf ein patankar-konservatives System (2.2), für alle* $n \in \mathbb{N}, h \in \mathbb{R}^+$ *und* $\boldsymbol{c}^n \in \mathbb{R}^N$

$$\sum_{i=1}^{N} \phi_i(h, \mathbf{c}^n, \mathbf{c}^{n+1}, t^n) \quad = \quad 0$$

gilt.

Folgendes sieht man leicht.

Bemerkung 2.13 *Für zwei zeitlich aufeinander folgende Näherungen eines patankar-konservativen Verfahrens zu einem patankar-konservativen System gilt*

$$\begin{aligned} \sum_{i=1}^{N} \left(c_i^{n+1} - c_i^n\right) &= \sum_{i=1}^{N} \left(c_i^n + h\ \phi_i(h, \mathbf{c}^n, \mathbf{c}^{n+1}, t^n) - c_i^n\right) \\ &= h \sum_{i=1}^{N} \phi_i(h, \mathbf{c}^n, \mathbf{c}^{n+1}, t^n) = 0. \end{aligned}$$

Um einen allgemeinen Grundstock an Verfahren für gewöhnliche Differentialgleichungen zu haben, wird weitere Notation benötigt.

Definition 2.14 *Ein allgemeines* r *stufiges Runge Kutta Verfahren für autonome Systeme (2.2) mit rechter Seite* $\mathbf{f}$ *bzw.* $S\mathbf{r}$ *hat die Form für* $v = 1, \ldots, r$ *und* $i = 1, \ldots, N$

$$s_i^v = c_i^n + h \sum_{w=1}^{r} a_{v,w} \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^w) \right) \tag{2.4}$$

mit der letztendlichen Näherung

$$c_i^{n+1} = c_i^n + h \sum_{v=1}^{r} b_v \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^v) \right) \tag{2.5}$$

und Konstanten $b_1, \ldots, b_r, a_{1,1}, \ldots, a_{r,r} \in \mathbb{R}$.

Dies führt auf eine allgemeinere Klasse von Verfahren mit noch zu klärenden Eigenschaften.

Definition 2.15 *Gegeben sei ein allgemeines r stufiges Runge Kutta Verfahren für autonome Systeme (2.2) mit rechter Seite* $\mathbf{f}$ *bzw.* $S\mathbf{r}$ *und Koeffizienten* $b_1, \ldots, b_r, a_{1,1}, \ldots, a_{r,r} \in \mathbb{R}$ *nach Definition 2.14 mit exakt Ordnung* $p \in \mathbb{N}$.

Für $v = 1, \ldots, r$ *und* $i = 1, \ldots, N$ *heißt ein Verfahren der Form*

$$\tilde{s}_i^v = c_i^n + h \sum_{w=1}^{r} a_{v,w} \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^w) \alpha_k^v \right) \tag{2.6}$$

und

$$\tilde{c}_i^{n+1} = c_i^n + h \sum_{v=1}^{r} b_v \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^v) \beta_k \right) \tag{2.7}$$

mit Größen $\alpha_1^1, \ldots, \alpha_K^1, \ldots, \alpha_1^r, \ldots, \alpha_K^r \in \mathbb{R}$ *und* $\beta_1, \ldots, \beta_K \in \mathbb{R}$ gestörtes Runge Kutta Verfahren.

Die Größen $\alpha_1^1, \ldots, \alpha_K^r \in \mathbb{R}$ *und* $\beta_1, \ldots, \beta_K \in \mathbb{R}$ *heißen* Störkoeffizienten *des gestörten Runge Kutta Verfahrens.*

Es bleibt anzumerken, dass sich die Exponentenschreibweise für die Störkoeffizienten $\alpha_1^1, \ldots, \alpha_K^r \in \mathbb{R}$ an der korrespondierenden Schreibweise für Zeitschritte $\mathbf{c}^n$ bzw. $\mathbf{c}^{n+1}$ und den Zwischenlösungen $\tilde{\mathbf{s}}^v$ orientiert.

Es kann davon ausgegangen werden, dass im Weiteren alle Exponenten eine Zuordnung zu einem Zeitschritt oder Zwischenergebnis bedeuten. Ausnahmen werden explizit im Text angegeben. Eine globale Ausnahme bilden die Exponenten zur Schrittweite h, diese meinen immer Potenzen.

Um in den kommenden Beweisen bestimmte Maxima aus $\mathbb{Z} \cup \{-\infty, \infty\}$ verwenden zu können, werden die folgenden Konventionen verwendet. Bei allen folgenden Grenzwertprozessen wird immer $h \to 0$ betrachtet.

$$a = \mathbb{O}(h^\infty)$$

bedeutet $a = \mathbb{O}(h^p)$ für alle $p \in \mathbb{Z}$. Analog dazu versteht sich

$$a = \mathbb{O}(h^{-\infty})$$

als die Aussage, für kein $p \in \mathbb{Z}$ gilt $a = \mathbb{O}(h^p)$.

Theorem 2.16 *Gegeben sei ein gestörtes Runge Kutta Verfahren. Dies sei angewendet auf ein System (2.2) mit ausreichend glattem* $\mathbf{r}$*. Die exakte Ordnung des zugrunde liegenden Runge Kutta Verfahrens sei* p*. Es gelten alle Bezeichnungen entsprechend den Definitionen 2.14 und 2.15. Sei* $u_1 \in \mathbb{Z} \cup \{-\infty, \infty\}$ *maximal gewählt, so dass für* $k = 1, \ldots, K$

$$\beta_k = 1 + \mathbb{O}\left(h^{u_1}\right) \tag{2.8}$$

gilt. Wähle u_2 *so, dass für* $v = 1, \ldots, r$

$$\tilde{\mathbf{s}}^v = \mathbf{s}^v + \mathbb{O}\left(h^{u_2}\right) \tag{2.9}$$

erfüllt und $u_2 \in \mathbb{Z} \cup \{-\infty, \infty\}$ *maximal ist.*

Sei $\tilde{p} := \min\{p, u_1, u_2, u_1 + u_2\}$*. Ist* $\tilde{p} \geq 1$*, dann ist* $\tilde{p}$ *die exakte Ordnung des untersuchten gestörten Runge Kutta Verfahrens. Ist* $\tilde{p} < 1$*, so ist das Verfahren nur zum trivialen System konsistent.*

Beweis: Es sei daran erinnert, dass für Ordnungsuntersuchungen immer die exakten Startwerte angenommen werden, d.h. $c_i^n = c_i(t^n)$.

Einsetzen von (2.8) und (2.9) in (2.7) liefert für $i = 1, \ldots, N$

$$\begin{aligned}
\tilde{c}_i^{n+1} &= c_i^n + h \sum_{v=1}^{r} b_v \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^v) \beta_k \right) \\
&= c_i^n + h \sum_{v=1}^{r} b_v \left[\sum_{k=1}^{K} \Big(S_{ik} r_k(\mathbf{s}^v) + \mathbb{O}(h^{u_2}) \Big) \Big(1 + \mathbb{O}(h^{u_1}) \Big) \right] \\
&\overset{u:=\min\{u_1,u_2,u_1+u_2\}}{=} c_i^n + h \sum_{v=1}^{r} b_v \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^v) \right) + \mathbb{O}(h^{u+1}) \\
&\overset{(2.5)}{=} c_i^{n+1} + \mathbb{O}(h^{u+1}) \\
&= c_i(t^{n+1}) + \mathbb{O}(h^{p+1}) + \mathbb{O}(h^{u+1}) \\
&\overset{\tilde{p}:=\min\{u,p\}}{=} c_i(t^{n+1}) + \mathbb{O}(h^{\tilde{p}+1}).
\end{aligned}$$

Gilt also $\tilde{p} \geq 1$, so hat das gestörte Runge Kutta Verfahren mindestens Ordnung $\tilde{p}$. Da die Größen p, u_1, u_2 maximal gewählt waren, ist die Ordnung genau $\tilde{p}$. Ist

$\tilde{p} < 1$ und das untersuchte System nicht das Triviale, so ist das Verfahren nicht konsistent zu diesem System. Ist das untersuchte System das Triviale, so folgt unabhängig vom verwendeten Verfahren direkt

$$\tilde{\mathbf{c}}^{n+1} = \mathbf{c}^{n+1} = \mathbf{c}^n$$

und es ist in diesem Spezialfall konsistent. □
Daraus erhält man direkt die folgende Charakteristik.

Korollar 2.17 *Es gelten die Bezeichnungen aus Theorem 2.16. Ein gestörtes Runge Kutta Verfahren ist genau dann ordnungserhaltend, wenn $u_1, u_2 \geq p$ gilt.*

Der folgende Satz zeigt einen Zusammenhang zwischen den Störkoeffizienten α_k^v sowie der „Ähnlichkeit " der neuen Zwischenlösungen $\tilde{\mathbf{s}}^v$ und den klassischen Runge Kutta Zwischenlösungen $\mathbf{s}^v$.

Satz 2.18 *Gegeben sei ein System (2.2) mit glattem* $\mathbf{r}$*. Wendet man ein explizites, r-stufiges, $r \geq 2$, gestörtes Runge Kutta Verfahren auf dieses System an, so gilt für $v = 1, \ldots, r$*

$$\tilde{\mathbf{s}}^v = \mathbf{s}^v + \mathbb{O}\left(h^{u_3+1}\right),$$

falls ein $u_3 \in \mathbb{N} \cup \{\infty\}$ existiert, so dass für $k = 1, \ldots, K$ und $v = 1, \ldots, r$

$$\alpha_k^v = 1 + \mathbb{O}\left(h^{u_3}\right) \tag{2.10}$$

gilt. Für $r = 1$ gilt die Aussage analog für

$$\tilde{\mathbf{c}}^{n+1} = \mathbf{c}^{n+1} + \mathbb{O}\left(h^{u_3+1}\right)$$

und

$$\beta_k = 1 + \mathbb{O}\left(h^{u_3}\right).$$

Beweis: Liegt dem untersuchten gestörten Runge Kutta Verfahren ein explizites Runge Kutta Verfahren zugrunde, so wird aus (2.6)

$$\tilde{s}_i^v = c_i^n + h \sum_{w=1}^{v-1} a_{v,w} \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^w) \alpha_k^v \right). \tag{2.11}$$

Für $v = 1$ gilt $\tilde{\mathbf{s}}^1 = \mathbf{s}^1 = \mathbf{c}^n$. Mit Induktion zeigt man das Weitere. Sei $v = 2$, dann erhält man

$$\begin{aligned}
\tilde{s}_i^2 &= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k \left(\tilde{\mathbf{s}}^1 \right) \alpha_k^2 \right) \\
&= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k \left(\mathbf{s}^1 \right) \left(1 + \mathbb{O}\left(h^{u_3} \right) \right) \right) \\
&= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k \left(\mathbf{s}^1 \right) \right) + \mathbb{O}\left(h^{u_3+1} \right) \\
&= s_i^2 + \mathbb{O}\left(h^{u_3+1} \right).
\end{aligned}$$

Sei die Aussage für v gezeigt und folgere nun die Aussage für $v+1$. Dazu sieht man

$$\begin{aligned}
\tilde{s}_i^{v+1} &= c_i^n + h \sum_{w=1}^{v} a_{v+1,w} \left(\sum_{k=1}^{K} S_{ik} r_k \left(\tilde{\mathbf{s}}^w \right) \alpha_k^{v+1} \right) \\
&= c_i^n + h \sum_{w=1}^{v} a_{v+1,w} \left(\sum_{k=1}^{K} \left(S_{ik} r_k \left(\mathbf{s}^w \right) + \mathbb{O}\left(h^{u_3+1} \right) \right) \left(1 + \mathbb{O}\left(h^{u_3} \right) \right) \right) \\
&= s_i^{v+1} + \mathbb{O}\left(h^{u_3+1} \right).
\end{aligned}$$

Der Induktionsanfang liefert nach einer Umbenennung ebenso die Aussage für $r = 1$. □

Wie man im Folgenden sieht, gilt die Gegenrichtung auch für implizite Verfahren.

Satz 2.19 *Gilt für ein gestörtes Runge Kutta Verfahren (sowohl explizit als auch implizit) angewendet auf ein System (2.2) mit glattem* $\mathbf{r}$ *für* $v = 1, \ldots, r$

$$\tilde{\mathbf{s}}^v = \mathbf{s}^v + \mathbb{O}\left(h^{u_3+1} \right)$$

mit einem $u_3 \in \mathbb{N} \cup \{\infty\}$*, welches maximal ist, so folgt für* $k = 1, \ldots, K$ *und* $v = 1, \ldots, r$

$$\alpha_k^v = 1 + \mathbb{O}\left(h^{u_3} \right).$$

Beweis: Sei nun $u_3 \in \mathbb{N} \cup \{\infty\}$ maximal, so dass für $v = 1, \ldots, r$

$$\tilde{\mathbf{s}}^v = \mathbf{s}^v + \mathbb{O}(h^{u_3+1}) \tag{2.12}$$

gilt. Da die Menge $\left\{ \alpha_k^v \middle| k \in \mathbb{N}_{\leq K}, v \in \mathbb{N}_{\leq r} \right\}$ endlich ist, existiert ein maximales $u_4 \in \mathbb{Z} \cup \{-\infty, \infty\}$, welches für alle $k = 1, \ldots, K$ und $v = 1, \ldots, r$

$$\alpha_k^v = 1 + \mathbb{O}(h^{u_4})$$

erfüllt. Nun bleibt $u_3 = u_4$ zu zeigen. Einsetzen in (2.6) für $v = 1, \ldots, r$ und $i = 1, \ldots, N$ liefert

$$\begin{aligned}
\tilde{s}_i^v &= c_i^n + h \sum_{w=1}^{r} a_{v,w} \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^w) \alpha_k^v \right) \\
&= c_i^n + h \sum_{w=1}^{r} a_{v,w} \left(\sum_{k=1}^{K} \left(S_{ik} r_k(\mathbf{s}^w) + \mathbb{O}(h^{u_3+1}) \right) (1 + \mathbb{O}(h^{u_4})) \right) \\
&\overset{u_5 := \min\{u_3+1, u_4\}}{=} c_i^n + h \sum_{w=1}^{r} a_{v,w} \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^w) \right) + \mathbb{O}(h^{u_5+1}) \\
&\overset{(2.4)}{=} \mathbf{s}^v + \mathbb{O}(h^{u_5+1}).
\end{aligned}$$

Unter Berücksichtigung der Maximalitätsforderung in (2.12) folgt $u_3 \geq u_5 = \min\{u_3 + 1, u_4\}$, woraus sich direkt $u_5 = u_4$ ergibt. Aus den Maximalitätsforderungen für jeweils u_3 und u_4 folgt nun aus (2.12) direkt $u_3 = u_4$, was den Beweis abschließt. □

2.2 Die Klasse der modifizierten Patankar Verfahren

Die Verfahren des modifizierten Patankar Typs sind eine konservative Weiterentwicklung der von Patankar in [Pat80] vorgestellten Modifikation zur Linearisierung von Senkentermen zur Generierung von positivitätserhaltenden Verfahren. Patankar entwickelte für die Anwendung in turbulenten Strömungen positive Modifikationen bestehender Verfahren.

Formal, d.h. nach dem weiter oben vorgestellten Vokabular, sind es gestörte explizite Runge Kutta Verfahren, deren Anwendung ein patankar-konservatives (Definition 2.5) System voraussetzt. Man hat also zu jeder Reaktion r_k genau eine Komponente c_i und c_j mit $S_{ik} \neq 0 \neq S_{jk}$ und $S_{ik} = -S_{jk}$. Ohne Einschränkung wird angenommen, dass $S_{ik} = -S_{jk} = 1$ gilt. Mit Hilfe dieser Indizes, $i(k)$ und $j(k)$, ist es möglich, die Störkoeffizienten α_k^v und β_k so zu wählen, dass die entstehenden Verfahren sowohl positiv als auch patankar-konservativ sind.

Definition 2.20 *Zu einem expliziten r-stufigen Runge Kutta Verfahren liefern*

die Wahlen von

$$\alpha_k^v = \frac{\tilde{s}_j^v}{\tilde{s}_j^{v-1}}, v \geq 2, \text{ und } \beta_k = \frac{\tilde{c}_j^{n+1}}{\tilde{s}_j^r} \tag{2.13}$$

die zugehörigen modifizierten Patankar Verfahren*, wobei zu jeder Reaktion r_k der Index j mit der Komponente korrespondiert, für die $S_{jk} = -1$ gilt.*

Jeweils ein Verfahren erster und zweiter Ordnung werden vorgestellt. Betrachtet man das Euler Verfahren angewendet auf ein System (2.2) für alle $i = 1, \ldots, N$, so ergibt sich

$$c_i^{n+1} = c_i^n + h \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}^n) \right). \tag{2.14}$$

Definition 2.21 *Gemäß (2.13) wird aus (2.14) das* modifizierte Patankar Euler Verfahren (mPE Verfahren)*. Für $i = 1, \ldots, N$ hat es die Form*

$$c_i^{n+1} = c_i^n + h \left(\sum_{k=1}^{K} S_{ik} r_k(\boldsymbol{c}^n) \frac{c_j^{n+1}}{c_j^n} \right). \tag{2.15}$$

Mit einigem Aufwand [BDM03, Zar05] lässt sich der folgende Satz zeigen.

Satz 2.22 *Das mPE Verfahren ist angewendet auf ein patankar-konservatives System mit positiven Anfangswerten und ausreichend glatter Funktion* $\mathbf{r}$ *für alle Größen positiv, patankar-konservativ und erster Ordnung.*

Satz 2.22 gibt Anlass zu folgender Generalvoraussetzung. Für die Anwendung des mPE Verfahrens wird immer ein patankar-konservatives System mit positiven Anfangswerten und glattem Reaktionsvektor $\mathbf{r}$ angenommen.

Es sei ausdrücklich auf folgenden wichtigen Gedanken hingewiesen.

Bemerkung 2.23 *Wird das mPE Verfahren auf ein eingeschränkt positives, patankar-konservatives System mit lediglich positiven Anfangswerten angewendet, werden auch die uneingeschränkten Größen immer positiv angenähert.*

Das mPE Verfahren „erhält" Positivität auch dort, wo sie nicht die Güte der Näherung verbessert. Gleichzeitig, was die folgenden Rechnungen vereinfacht und auch regelmäßig Anwendung findet, weiß man aber sicher, dass alle berechneten Näherungen immer positiv sind, d.h. $c_i^{n+1} > 0$ für alle $i = 1, \ldots, N$.

Aus den Sätzen 2.18 und 2.22 gewinnt man eine bedauerliche Kernerkenntnis über Verfahren vom modifizierten Patankar Typ.

Satz 2.24 *Für ein gestörtes explizites Runge Kutta Verfahren vom modifizierten Patankar Typ gibt es ein Ordnungsmaximum von zwei.*

Beweis: Bei einem expliziten Verfahren ist der erste „echte“ Schritt, die Berechnung von $\tilde{\mathbf{s}}^2$, strukturell immer ein modifizierter Patankar Euler Schritt. Entsprechend erhält man mit der generischen Form (2.13) und unter Verwendung der ersten Ordnung des mPE Verfahrens, also $u_3 = 1$,

$$\frac{\tilde{s}_i^2}{\tilde{s}_i^1} = \frac{\left(c_i^{n+1}\right)_{mPE}}{c_i^n} = 1 + \mathbb{O}\left(h\right).$$

Einsetzen in (2.6) liefert

$$\begin{aligned}
\tilde{s}_i^2 &= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k(\tilde{\mathbf{s}}^1) \alpha_k^2 \right) \\
&= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^1) \left(1 + \mathbb{O}\left(h\right)\right) \right) \\
&= c_i^n + h\, a_{2,1} \left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{s}^1) \right) + \mathbb{O}\left(h^2\right) \\
&= s_i^2 + \mathbb{O}\left(h^2\right).
\end{aligned}$$

Entsprechend der Notation aus Theorem 2.16 ergibt sich also $u_2 = 2$ und damit die maximale Ordnung von zwei für das gestörte Verfahren. □

Weiterhin kann man zeigen, dass sich die Näherungen des mPE Verfahrens durch eine formale Potenzreihe darstellen lassen.

Satz 2.25 *Für das mPE Verfahren (2.15) gilt mit $P \in \mathbb{N} \cup \{0\}$ beliebig und für $h \to 0$*

$$c_i^{n+1} = \sum_{\nu=0}^{P} h^\nu f_i^\nu(\boldsymbol{c}^n) + \mathbb{O}\left(h^{P+1}\right), \tag{2.16}$$

wobei

$$f_i^\nu(\boldsymbol{c}^n) = \sum_{k=1}^{K} S_{ik} r_k(\boldsymbol{c}^n) \frac{f_j^{\nu-1}(\boldsymbol{c}^n)}{c_j^n} \tag{2.17}$$

mit

$$\boldsymbol{f}^0(\boldsymbol{c}^n) = \boldsymbol{c}^n \tag{2.18}$$

ist. Auch hier gilt, die Exponenten ν zu den $\mathbf{f}^\nu$ sind Ordnungszuordnungen und keine Potenzen (im Gegensatz zu den ν bei der Schrittweite h).

Beweis: Hiervon überzeugt man sich mit Induktion. Für $P = 0$ gilt unter Verwendung der Ordnungsaussage von Satz 2.22

$$c_i^{n+1} = c_i^n + \mathbb{O}\left(h^1\right).$$

Sei die Aussage nun für P gezeigt, d.h. gälte

$$c_i^{n+1} = \sum_{\nu=0}^{P} h^\nu f_i^\nu(\mathbf{c}^n) + \mathbb{O}\left(h^{P+1}\right).$$

Daraus folgt

$$\frac{c_i^{n+1}}{c_i^n} = \sum_{\nu=0}^{P} h^\nu \frac{f_i^\nu(\mathbf{c}^n)}{c_i^n} + \mathbb{O}\left(h^{P+1}\right).$$

Einsetzen in (2.15) liefert

$$\begin{aligned}
c_i^{n+1} &= c_i^n + h\left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}^n)\left(\sum_{\nu=0}^{P} h^\nu \frac{f_j^\nu(\mathbf{c}^n)}{c_j^n} + \mathbb{O}\left(h^{P+1}\right)\right)\right) \\
&= c_i^n + \sum_{\nu=0}^{P} h^{\nu+1}\left(\sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}^n)\frac{f_j^\nu(\mathbf{c}^n)}{c_j^n}\right) + \mathbb{O}\left(h^{P+2}\right) \\
&= c_i^n + \sum_{\nu=0}^{P} h^{\nu+1} f_i^{\nu+1}(\mathbf{c}^n) + \mathbb{O}\left(h^{P+2}\right) \\
&= c_i^n + \sum_{\nu=1}^{P+1} h^\nu f_i^\nu(\mathbf{c}^n) + \mathbb{O}\left(h^{P+2}\right) \\
&= \sum_{\nu=0}^{P+1} h^\nu f_i^\nu(\mathbf{c}^n) + \mathbb{O}\left(h^{P+2}\right).
\end{aligned}$$

□

Es kann noch folgendes Detail festgehalten werden.

Korollar 2.26 *Es gilt* $\boldsymbol{f}^1(\boldsymbol{c}(t)) = \boldsymbol{f}(\boldsymbol{c}(t)) \left[= \boldsymbol{c}'(t)\right]$.

Beweis: Einsetzen liefert für $i = 1, \ldots, N$

$$\begin{aligned}
f_i^1(\mathbf{c}(t)) &= \sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}(t)) \frac{f_j^0(\mathbf{c}(t))}{c(t)_j} = \sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}(t)) \frac{c(t)_j}{c(t)_j} \\
&= \sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}(t)) \overset{(2.2)}{=} f_i(\mathbf{c}(t)).
\end{aligned}$$

□

Nun lässt sich noch ein für spätere Beweise nützliches Faktum festhalten.

Lemma 2.27 *Existieren für alle Reaktionen r_k und Größen c_i aus (2.2) Taylorentwicklungen bis mindestens der Ordnung P, dann existieren auch Entwicklungen für die Funktionen f_i^ν aus (2.17) des mPE Verfahrens bis zur Ordnung P.*

Beweis: Der Beweis funktioniert wiederum induktiv. Für $\nu = 0$ gilt sicherlich

$$\begin{aligned} f_i^0\left(\mathbf{c}\left(t^{n+1}\right)\right) &= c_i\left(t^{n+1}\right) \\ &= \sum_{p=0}^{P} \frac{h^p}{p!} c_i^{(p)}\left(t^n\right) + \mathbb{O}\left(h^{P+1}\right). \end{aligned}$$

Sei die Aussage nun für ν gezeigt. Einsetzen liefert

$$f_i^{\nu+1}\left(\mathbf{c}\left(t^{n+1}\right)\right) = \sum_{k=1}^{K} S_{ik} r_k\left(\mathbf{c}\left(t^{n+1}\right)\right) \frac{f_j^\nu\left(\mathbf{c}\left(t^{n+1}\right)\right)}{c_j\left(t^{n+1}\right)}.$$

Da für die f_i^ν nach Voraussetzung eine Entwicklung vorliegt und die S_{ik} konstant sind, bleiben nur noch die folgenden Terme

$$\begin{aligned} \frac{r_k\left(\mathbf{c}\left(t^{n+1}\right)\right)}{c_j\left(t^{n+1}\right)} &= \frac{\sum_{p=0}^{P} \frac{h^p}{p!} \frac{d^p}{dt^p} r_k\left(\mathbf{c}\left(t^n\right)\right) + \mathbb{O}\left(h^{P+1}\right)}{\sum_{p=0}^{P} \frac{h^p}{p!} c_j^{(p)}\left(t^n\right) + \mathbb{O}\left(h^{P+1}\right)} \\ &= \underbrace{\frac{1}{c_j\left(t^n\right) + \mathbb{O}\left(h\right)}}_{>0} \sum_{p=0}^{P} \frac{h^p}{p!} \frac{d^p}{dt^p} r_k\left(\mathbf{c}\left(t^n\right)\right) + \mathbb{O}\left(h^{P+1}\right) \end{aligned}$$

zu untersuchen. Es lässt sich also auch eine Entwicklung für die Brüche $\frac{r_k\left(\mathbf{c}\left(t^{n+1}\right)\right)}{c_j\left(t^{n+1}\right)}$ in h für $h \to 0$ angeben. □

Mit Lemma 2.27 und (2.16) erhält man direkt eine Entwicklung für die Verfahrensfunktion ϕ.

Korollar 2.28 *Sei $P \in \mathbb{N} \cup \{0\}$. Existieren Taylorentwicklungen für die Funktionen r_k und c_i bis zur Ordnung P, so ergibt sich aus den Darstellungen (2.16) und (2.3) durch Einsetzen*

$$\phi\left(h, \mathbf{c}^n, \mathbf{c}^{n+1}, t^n\right) = \sum_{\nu=1}^{P} h^{\nu-1} \mathbf{f}^\nu(\mathbf{c}^n) + \mathbb{O}\left(h^P\right). \tag{2.19}$$

Haben also die $\mathbf{f}^\nu$ eine Entwicklung bis wenigstens zur Ordnung $P - \nu$, so hat ϕ eine der Ordnung $P - 1$.

Dazu ist noch eine kurze Bemerkung angebracht.

Bemerkung 2.29 *Besitzen alle Funktionen* $\mathbf{f}^\nu$ *eine Entwicklung bis zur Ordnung* m*, so gilt* $m = P - \nu \geq P - 1$*, woraus sofort* $P \geq m + 1$ *folgt.*

Oder anders formuliert, existieren Entwicklungen für alle $\mathbf{f}^\nu$ *bis zur Ordnung* m*, so besitzt auch* ϕ *eine Entwicklung der Ordnung* m*.*

Für weitere theoretische Aussagen ist die kommende Definition essentiell.

Definition 2.30 *Man definiert mit (2.3) den lokalen Diskretisierungsfehler* l_i *im* $(n+1)$*-ten Schritt der* i*-ten Komponente eines Einschrittverfahrens durch*

$$l_i(n+1) = c_i\left(t^{n+1}\right) - \left(c_i\left(t^n\right) + h\,\phi_i\big(h,\, \boldsymbol{c}\left(t^n\right),\, \boldsymbol{c}\left(t^{n+1}\right),\, t^n\big)\right) \tag{2.20}$$

für $i = 1, \ldots, N$*. Beachte dabei, dass in dieser Notation die Abhängigkeit des Fehlers von der verwendeten Schrittweite* h *stillschweigend vorausgesetzt und nicht explizit notiert wird.*

Aus (2.18) und der Taylorentwicklung für $\mathbf{c}(t^{n+1})$ ergibt sich direkt eine allgemeine Darstellung des lokalen Fehlers für das mPE Verfahren.

Korollar 2.31 *Der lokale Fehler des* $(n+1)$*-ten Schrittes in der* i*-ten Komponente* $l_i(n+1)$ *des mPE Verfahrens zu einem Problem* $\boldsymbol{c}' = S\boldsymbol{r}(\boldsymbol{c})$ *mit Funktionen* $\boldsymbol{c}$ *und* $\mathbf{r}$*, für welche Taylorentwicklungen bis zur Ordnung* P *existieren, lautet*

$$\begin{aligned} & l_i(n+1) \\ &= \sum_{\nu=0}^{P} h^\nu \frac{c_i^{(\nu)}(t^n)}{\nu!} + \mathbb{O}\left(h^{P+1}\right) - \sum_{\nu=0}^{P} h^\nu f_i^\nu\big(\boldsymbol{c}(t^n)\big) + \mathbb{O}\left(h^{P+1}\right) \\ &\overset{\text{Korollar 2.26, (2.18)}}{=} \sum_{\nu=2}^{P} h^\nu \left(\frac{\boldsymbol{c}^{(\nu)}(t^n)}{\nu!} - f_i^\nu\big(\boldsymbol{c}(t^n)\big)\right) + \mathbb{O}\left(h^{P+1}\right). \end{aligned}$$

Damit sind die Untersuchungen zum Verfahren erster Ordnung vorläufig abgeschlossen. Als Verfahren zweiter Ordnung wird das Heun Verfahren modifiziert.

Das Heun Verfahren ist hier noch einmal in neuer Schreibweise (2.2) für $i = 1, \ldots, N$ angegeben:

$$s_i^2 \;=\; c_i^n + h\left(\sum_{k=1}^{K} S_{ik} r_k\left(\mathbf{c}^n\right)\right)$$

und für $i = 1, \ldots, N$

$$c_i^{n+1} = c_i^n + \frac{h}{2} \sum_{k=1}^{K} \left(S_{ik} r_k \left(\mathbf{c}^n \right) + S_{ik} r_k \left(\mathbf{s}^2 \right) \right).$$

Definition 2.32 *Daraus gewinnt man das modifizierte Patankar Heun Verfahren (mPH Verfahren) für* $i = 1, \ldots, N$

$$\begin{aligned} \tilde{s}_i^2 &= c_i^n + h \left(\sum_{k=1}^{K} S_{ik} r_k \left(\boldsymbol{c}^n \right) \frac{\tilde{s}_j^2}{c_j^n} \right) \\ c_i^{n+1} &= c_i^n + \frac{h}{2} \sum_{k=1}^{K} S_{ik} \left(r_k \left(\boldsymbol{c}^n \right) + r_k \left(\tilde{\boldsymbol{s}}^2 \right) \right) \frac{c_j^{n+1}}{\tilde{s}_j^2}. \end{aligned} \tag{2.21}$$

Dabei ist $j = j(k)$ *derjenige Index für den* $S_{jk} < 0$ *gilt. Dies ist nach Voraussetzung eindeutig, da es nur zwei Elemente* $S_{jk}, S_{lk} \neq 0$ *gibt und für diese ebenfalls* $S_{jk} = -S_{lk}$ *gilt.*

Unter der Voraussetzung von Satz 2.22 lassen sich mit weniger Aufwand die gewünschten Eigenschaften für das mPH Verfahren zeigen (ebenso wie die Beweise zum mPE Verfahren findet sich auch dies in [BDM03, Zar05]).

Satz 2.33 *Das mPH Verfahren ist angewendet auf ein patankar-konservatives System mit positiven Anfangswerten und glattem* $\mathbf{r}$ *positiv, patankar-konservativ und zweiter Ordnung.*

2.3 Modifizierte Patankar Verfahren beliebiger Ordnung

Satz 2.24 erzwingt auf der Suche nach positiven und patankar-konservativen Verfahren höherer Ordnung neue oder wenigstens erweiterte Ansätze als die generische Modifikation bekannter Runge Kutta Verfahren gemäß (2.13). Der folgende Abschnitt dient dem Beweis einer Weiterentwicklung der modifizierten Patankar Verfahren, welche in Abhängigkeit des Problems von beliebig hoher Ordnung ist. Die Idee, welche genutzt wird, geht auf Romberg zurück und wurde von Gragg z.B. in [Gra65] formuliert.

2.3.1 Die Extrapolationsidee

Das wesentliche Ergebnis findet sich im Buch von Plato [Pla04]. Dies wird an dieser Stelle nur angegeben.

Satz 2.34 *Sei eine Funktion $\mathcal{T}$ für diskrete Werte $h_j > 0$ bestimmbar und besitze eine asymptotische Entwicklung in h für $h \to 0$ mit Konstanten $\tau_i \in \mathbb{R}$, $\Omega \in \mathbb{N}, \kappa \in \{1, 2\}$, so dass*

$$\mathcal{T}(h) = \sum_{i=0}^{\Omega} h^{i\kappa}\tau_i + \mathbb{O}\left(h^{\kappa(\Omega+1)}\right) \tag{2.22}$$

gilt. Außerdem gebe es ein $H > 0$ und eine Folge n_j mit $1 < n_0$ und $n_j \leq n_{j+1}$ die $h_j = \frac{H}{n_j}$ definieren.

$$P_{k,\ldots,k+m}\left(h_j^{\kappa}\right) = \mathcal{T}(h_j) \qquad \textit{mit} \qquad j = k, \ldots, k+m$$

seien die eindeutig bestimmten Interpolationspolynome m-ten Grades. Bezeichne

$$T_{k,\ldots,k+m} := P_{k,\ldots,k+m}(0)$$

den Wert der, an der Stelle Null ausgewerteten, Interpolationspolynome.

Dann gilt die asymptotische Abschätzung für $H \to 0$ und $0 \leq m \leq \Omega - 1$

$$T_{k,\ldots,k+m} = \tau_0 + (-1)^m \frac{\tau_{m+1}}{\prod_{i=k}^{k+m} n_i^{\kappa}} H^{\kappa(m+1)} + \mathbb{O}\left(H^{\kappa(m+2)}\right).$$

Was bedeutet das? Die Funktion $\mathcal{T}$ beschreibt üblicherweise ein numerisches Verfahren. Der globale Fehler g im n-ten Schritt eines Verfahrens ist festgelegt durch

$$g(n) = c(t^n) - c^n \tag{2.23}$$

für $c(t^n)$ die gesuchte Lösung und $c^n = \mathcal{T}(h)$ die numerische Lösung, berechnet mit Schrittweite $h > 0$ jeweils im n-ten Schritt. Kann man für den globalen Fehler $g(n)$ eines gewöhnlichen Einschrittverfahrens (2.3) eine Darstellung gemäß

$$g(n) = -\sum_{i=1}^{\Omega} h^{i\kappa}\tau_i + \mathbb{O}\left(h^{\kappa(\Omega+1)}\right) \tag{2.24}$$

finden, ergibt sich die theoretische Form (2.22) durch die Wahl $\tau_0 = c(t^n)$. Beachte hierbei, $\kappa = 2$ bedeutet, dass der globale Fehler des Verfahrens nur gerade Potenzen umfasst, also alle Koeffizienten vor den ungeraden Potenzen Null sind. Dies

führt zu einer großen Effizienzsteigerung, da mit jeder zusätzlichen Näherungslösung $\mathcal{T}(h_j)$ die Ordnung der erreichten Extrapolation um zwei zunimmt statt der sonst möglichen Ordnungssteigerung um eins für eine weitere Näherungslösung $\mathcal{T}(h_j)$.

Berechnet man mit Ω unterschiedlichen Schrittweiten $h_j, j = 0, \ldots, \Omega - 1$ Näherungen für den Zeitpunkt t^n, also $\mathbf{c}^{n,j} = \mathcal{T}(h_j) \approx \mathbf{c}(t^n), j = 0, \ldots, \Omega - 1$, so lässt sich daraus eine Näherung der Ordnung $\kappa\ \Omega$ konstruieren. Dies ist im Wesentlichen unabhängig von der Ordnung des ursprünglichen Verfahrens.

Das Vorgehen ist wie folgt. Man bestimmt das Interpolationspolynom

$$P_{0,\ldots,\Omega-1}(h_j^\kappa) = \mathcal{T}(h_j)$$

und wertet es an der Stelle 0 aus. Bedenke hierbei, dass H, der eigentliche Schritt, unverändert bleibt. Es wird lediglich für die Zwischenschritte h_j der Grenzfall $h_j = 0$ explizit eingesetzt.

Dieser Wert ist die abschließende Näherung

$$\mathbf{c}^{n+1} = T_{0,\ldots,\Omega-1} = P_{0,\ldots,\Omega-1}(0),$$

der Ordnung $\kappa\ \Omega$. Eine umfassende Darstellung der Ideen findet sich zum Beispiel auch in [HNW91, HWN02]. Für die Ausführung einer automatischen Schrittweitensteuerung sei hier noch auf [Deu83, Deu85] verwiesen.

2.3.2 Die Fehlerentwicklung des globalen Fehlers für das modifizierte Patankar Euler Verfahren

Möchte man also Satz 2.34 anwenden, um die Tauglichkeit eines Verfahrens als Grundschema in einem Extrapolationsalgorithmus zu untersuchen, muss man den globalen Fehler dieses Verfahrens untersuchen. Gelingt es diesen für ein spezielles Verfahren in der Form (2.24) für ein $\Omega \in \mathbb{N}$ darzustellen, erhält man die Ordnung (eben $\kappa\ \Omega$), die durch Extrapolation mindestens zu erreichen ist.

Eine außerordentliche Effizienzsteigerung erhält man, wenn man als Grundverfahren einen Löser verwendet, dessen globaler Fehler ausschließlich in geraden Potenzen der verwendeten Schrittweite h Koeffizienten $\tau_i \neq 0$ hat,

$$g(n) = -\sum_{i=1}^{\Omega} h^{2\,i} \tau_i + \mathbb{O}\left(h^{2\,(\Omega+1)}\right).$$

Ein prominentes Beispiel hierfür ist die implizite Trapezregel [HWN02] oder, vorgestellt in [BD83], die linear implizite Mittelpunktsregel. Leider sind beide Methoden nicht positivitätserhaltend.

Jetzt ist eine asymptotische Entwicklung, wie in (2.24) gefordert, für den globalen Fehler $g_i(n+1)$ in h für $h \to 0$ für das mPE Verfahren zu zeigen. Dazu wird $g_i(n+1)$ auf den lokalen Fehler zurückgeführt; außerdem finden die Ergebnisse aus dem vorherigen Unterkapitel über die Verfahrensfunktion ϕ des mPE Verfahrens Anwendung. Eine Aufschlüsselung des globalen Fehlers liefert

$$\begin{aligned} g_i(n+1) &\overset{(2.23)}{=} c_i(t^{n+1}) - c_i^{n+1} \\ &\overset{(2.3,2.20)}{=} c_i(t^n) + h\,\phi_i\big(h, t^n, \mathbf{c}(t^n)\big) + l_i(n+1) - \Big(c_i^n + h\,\phi_i\big(h, t^n, \mathbf{c}^n\big)\Big) \\ &= c_i(t^n) - c_i^n + l_i(n+1) + h\,\Big(\phi_i\big(h, t^n, \mathbf{c}(t^n)\big) - \phi_i\big(h, t^n, \mathbf{c}^n\big)\Big) \\ &= g_i(n) + l_i(n+1) + \underbrace{h\,\Big(\phi_i\big(h, t^n, \mathbf{c}(t^n)\big) - \phi_i\big(h, t^n, \mathbf{c}^n\big)\Big)}_{=:\triangle\phi_i(n)}. \end{aligned} \tag{2.25}$$

Dieses Ergebnis wird im folgenden Satz aufgearbeitet.

Satz 2.35 *Es existiert eine Entwicklung für $g_i(n+1)$ bis zur Ordnung P, wenn r_k bzw. c_i aus (2.2) wenigstens $(P+1)$ mal stetig differenzierbar sind und man für $g_i(0)$ eine Entwicklung der Ordnung P hat.*

Beweis: Korollar 2.31 liefert die gewünschte Entwicklung bis zur Ordnung P für den lokalen Fehler $l_i(n+1)$.

Lemma 2.27 liefert eine Entwicklung der Potenzreihenfunktion $\mathbf{f}^\nu$ bis zur Ordnung P. Hieraus erhält man mit Korollar 2.28 eine Entwicklung für ϕ_i bis zur Ordnung P. Insgesamt ergibt sich also eine Entwicklung für $\triangle\phi_i(n)$ bis zur Ordnung $(P+1)$. Das ist mehr als man benötigt, schadet aber nicht. Die Existenz dieser Aufschlüsselung (2.25) ist unabhängig von n. Daher lässt sie sich wiederholt ausführen und liefert

$$g_i(n+1) = g_i(m_1) + \sum_{m_2=m_1}^{n} l_i(m_2+1) + \triangle\phi_i(m_2),\ m_1 = 0, \ldots, n.$$

Betrachtet man in dieser Darstellung $m_1 = 0$ und nutzt die nach Voraussetzung vorhandene Entwicklung für $g_i(0)$, so hat man sich davon überzeugt, dass für sämtliche Terme Entwicklungen bis mindestens Ordnung P existieren. Damit sind alle Terme behandelt und die Aussage ist gezeigt. □

Es sei angemerkt, dass für Fehleruntersuchungen üblicherweise $g_i(0) = 0$ betrachtet wird, d.h. man startet mit der exakten Lösung. In diesem Fall existiert natürlich eine Entwicklung beliebig hoher Ordnung.

Es ist also möglich für ein hinreichend glattes Problem durch Extrapolation der mit dem mPE Verfahren bestimmten Näherungen eine Näherung beliebig hoher Ordnung zu berechnen. Die auf diese Weise entstehenden extrapolierten modifizierten Patankar Euler Verfahren seien mit *emPE ord* γ bezeichnet, wobei $\gamma \in \mathbb{N}$ für die Ordnung steht. Auch die Schrittweitensteuerung aus [Deu83, Deu85] wurde implementiert. Die durch diese Variante bestimmten Näherungen werden mit *emPE auto* bezeichnet. Im weiteren Verlauf sind Näherungen der Ordnung zwei bis zwölf betrachtet.

2.4 Verfahren zum Erhalt von linearen Invarianten

Im Kontext chemischer Reaktionsgleichungen oder komplexerer Ökosystemmodelle ergeben sich aber weitere Schwierigkeiten. Dies sieht man leicht an den Beispielsystemen 1.2, 1.3 und 1.4. Im Gegensatz zu Beispiel 1.1 sind alle diese Systeme nicht patankar-konservativ. Damit sind die Verfahren vom modifizierten Patankar Typ nicht anwendbar.

Gleichwohl sind diese Systeme aber von Interesse, worauf Bruggeman et al. in [BBKS07] und Broekhuizen et al. in [BRBM07] hingewiesen haben. Weiterhin ist es möglich für diese Systeme eine alternative Form (statt der bereits besprochenen Patankar-Konservativität) der Konservativität aufzustellen, deren Gewährleistung zusammen mit der entscheidenden Positivität durch gängige Verfahren nur eingeschränkt oder gar nicht gegeben ist. Zu diesem Zweck wird später noch eine allgemeinere Fassung des Begriffs der Konservativität eines Systems bzw. eines Verfahrens gegeben.

Den Systemen aus den Beispielen 1.2, 1.3 und 1.4 ist es gemeinsam, dass die Zusammensetzung der einzelnen Komponenten jedes Gleichungssystems nicht gleich ist, sondern dass die Zusammensetzungen der einzelnen Komponenten durch eine nicht triviale Stoichiometriematrix beschrieben werden.

Weiterhin sind viele Modelle nicht quellen- und senkenfrei definiert. Aber auch dies ist eine Eigenschaft, die zur Anwendung der modifizierten Patankar Typ Verfahren sichergestellt werden muss.

Zur Veranschaulichung wird sich eines Beispiels aus [BBKS07] bedient und die Begriffe aus Definition 2.1 werden an dem einfachen biochemischen System aus Beispiel 1.2 erklärt. Die rechte Seite aus Beispiel 1.2 besteht aus chemischer Sicht aus zwei Reaktionen r_1 und r_2. Diese lassen sich quantitativ veranschaulichen durch

$$\begin{array}{rcl} aC + bN & \overset{r_1(C,N,P)}{\longrightarrow} & P \text{ und} \\ P & \overset{r_2(P)}{\longrightarrow} & D. \end{array}$$

In Worten heißt das, aus Anzahl a Kohlenstoffeinheiten C und Anzahl b Nitrateinheiten N wird durch die Reaktion r_1 eine Phytoplanktoneinheit P. Und eine Phytoplanktoneinheit P wird via r_2 zu einer Einheit toter Materie D.

Die beiden Reaktionen r_1 und r_2 haben die konkrete Gestalt

$$\begin{aligned} r_1(C,N,P) &= r_{\max} \frac{C}{K_c + C} \frac{N}{K_n + N} P, \\ r_2(P) &= eP. \end{aligned} \tag{2.26}$$

Die rechte Seite von Beispiel 1.2 kann man nun mit Hilfe dieser Reaktionen als das Produkt einer Matrix-Vektor-Multiplikation schreiben.

Sei

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \quad , \quad S = \begin{pmatrix} -a & 0 \\ -b & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad \mathbf{c} = \begin{pmatrix} C \\ N \\ P \\ D \end{pmatrix}, \tag{2.27}$$

dann kann man das System aus Beispiel 1.2 schreiben als

$$\mathbf{c}' = S\, \mathbf{r}(\mathbf{c}).$$

Dieses System wäre beispielsweise für die Wahl $a = 0$ und $b = 1$ patankarkonservativ. Physikalisch gedeutet heißt das, dass man den Kohlenstoff, C, aus der Betrachtung nimmt und alle Größen nur noch über ihren Stickstoffgehalt miteinander in Verbindung setzt. Dieses System bestünde dann aus den (interessanten) Komponenten N, P und D. Daher kommt der in der Bemerkung 2.7 erwähnte Name, NPD konservativ.

Nun folgt die allgemeine Fassung des Begriffs der Konservativität.

Definition 2.36 *Ein System (2.2) heißt* konservativ, *falls es nicht triviale lineare Invarianten gibt, d.h. es existiert eine Matrix $E \in \mathbb{R}^{i\times N}, E \neq \mathbf{0}, i \in \mathbb{N}$, so dass für alle $h > 0$ und $t \geq 0$*

$$E(\mathbf{c}(t+h) - \mathbf{c}(t)) = \mathbf{0}$$

gilt. Diese Matrix E heißt Komponentenzusammensetzungsmatrix.

Die Matrix E ist zu einem gegebenen System nicht eindeutig bestimmt. Speziell ist die Anzahl der Zeilen i nicht eindeutig. Allerdings hängen mögliche Werte für i vom betrachteten System ab. Dies illustrieren die folgenden Beispiele.

Eine Matrix E für das Beispiel 1.2 lautet

$$E = \begin{pmatrix} 1 & 0 & a & a \\ 0 & 1 & b & b \end{pmatrix}. \tag{2.28}$$

Andere mögliche Wahlen für Beispiel 1.2 für die Komponentenzusammensetzungsmatrix E sind

$$E = \zeta_1 \begin{pmatrix} 1 & 0 & a & a \end{pmatrix}, E = \zeta_1 \begin{pmatrix} 0 & 1 & b & b \end{pmatrix} \text{ oder } E = \begin{pmatrix} \zeta_1 & 0 & \zeta_1 a & \zeta_1 a \\ 0 & \zeta_2 & \zeta_2 b & \zeta_2 b \end{pmatrix}$$

für $\zeta_1, \zeta_2 \in \mathbb{R} \setminus \{0\}$.

Das triviale System ist ebenfalls konservativ, da aus $\mathbf{c}' = \mathbf{0}$ unmittelbar $\mathbf{c}(t+h) = \mathbf{c}(t)$ für alle h folgt. Hier erfüllt offensichtlich jede Matrix $E \neq \mathbf{0} \in \mathbb{R}^{i\times N}$ die Bedingung aus Definition 2.36.

Die meisten konservativen Systeme werden auch durch den folgenden Zusammenhang beschrieben.

Satz 2.37 *Gegeben sei ein System (2.2) dessen Komponenten c_i stetig differenzierbar sind.*

Ist das System konservativ mit Komponentenzusammensetzungsmatrix E und seien die Komponenten des Reaktionsvektors $\mathbf{r}$ linear unabhängig, so liegt S im Kern von E.

Existiert andererseits eine Matrix $E \neq \mathbf{0}$ mit $ES = \mathbf{0}$, so ist das System konservativ mit Komponentenzusammensetzungsmatrix E.

Beweis: Mit Hilfe des Mittelwertsatzes ergibt sich für ein $\xi \in [t, t+h]$

$$E(\mathbf{c}(t+h) - \mathbf{c}(t)) = E\,(h\mathbf{c}'(\xi)) = E\,(hS\mathbf{r}(\xi)) = h\,E\,S\,\mathbf{r}(\xi). \tag{2.29}$$

Sei das System konservativ mit Komponentenzusammensetzungsmatrix $E \neq \mathbf{0}$. Man erhält dann aus (2.29) und der Setzung $\mathbf{r} = \mathbf{r}(\xi)$

$$\mathbf{0} = ES\mathbf{r}.$$

Man berechnet für jede Komponente i

$$0 \;=\; (ES\mathbf{r})_i = \sum_{k=1}^{K} (ES)_{(ik)} r_k.$$

Da die $r_1, \ldots, r_K$ nach Voraussetzung linear unabhängig sind, folgt $(ES)_{(ik)} = 0$ für alle i und k, also gilt $ES = \mathbf{0}$. Und S liegt wie gefordert im Kern von E.

Ist andererseits $ES = \mathbf{0}$, so liefert Einsetzen in (2.29) die gewünschte Aussage. Das System ist konservativ mit Komponentenzusammensetzungsmatrix E. Damit sind beide Richtungen gezeigt. □

Weiterhin erkennt man leicht den Zusammenhang zur Definition 2.12.

Bemerkung 2.38 *Ist ein System patankar-konservativ nach Definition 2.5, so ist es ein konservatives System mit*

$$E = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}. \tag{2.30}$$

Die Umkehrung gilt nicht, da für E wie in (2.30) nur

$$\sum_{i=1}^{N} S_{ik} = 0$$

für alle k folgt. Das ist aber offensichtlich schwächer als Definition 2.5. Vergleich dazu auch Bemerkung 2.9.

Verfahren, die die Konservativität eines Systems respektieren, lassen sich folgendermaßen charakterisieren.

Definition 2.39 *Ein Verfahren, das angewendet auf ein konservatives System die Eigenschaft*

$$E(\boldsymbol{c}^{n+1} - \boldsymbol{c}^n) = E\; h\; \phi\left(h, \mathbf{c}^n, \mathbf{c}^{n+1}\right) = \boldsymbol{0}$$

besitzt, heißt konservatives *Verfahren.*

Zusammen mit dem Satz 2.37 ergibt sich direkt die nächste Aussage.

Korollar 2.40 *Gegeben sei ein konservatives System mit K linear unabhängigen Reaktionen, Stoichiometriematrix S und einer stetig differenzierbaren Lösung $\mathbf{c}$. Ein numerisches Verfahren, angewendet auf dieses System, für das*

$$\phi = S\mathbf{v}$$

mit $\mathbf{v} \in \mathbb{R}^K$ gilt, ist konservativ.

Abschließend bleibt noch die folgende Bemerkung anzufügen.

Bemerkung 2.41 *Ein gestörtes Runge Kutta Verfahren nach Definition 2.15 ist, falls anwendbar, immer konservativ nach Korollar 2.40.*

2.4.1 Die Bruggeman Verfahren

In [BBKS07] werden Verfahren vorgestellt, welche in allgemeiner Weise konservativ nach Definition 2.36 und positiv sind. Allerdings sind sie nicht in der Lage, steife Probleme wie z.B. den Robertson Testfall oder das Orego Problem mit praktikabler Schrittweite $h > 0$ zu lösen (beide aus [HWN02]).

Die Grundidee der im Folgenden als Bruggeman Verfahren bezeichneten Löser ist eine Schrittweitensteuerung in Abhängigkeit der bedrohten (möglicherweise negativ werdenden) Größen.

Für das Euler Verfahren

$$\mathbf{c}^{n+1} = \mathbf{c}^n + h\mathbf{f}(\mathbf{c}^n) = \mathbf{c}^n + hS\mathbf{r}(\mathbf{c}^n)$$

verwendet Bruggeman eine modifizierte Schrittweite

$$\mathbf{c}^{n+1} = \mathbf{c}^n + h^\star\mathbf{f}(\mathbf{c}^n) = \mathbf{c}^n + h^\star S\mathbf{r}(\mathbf{c}^n) \tag{2.31}$$

mit

$$h^\star = h\ p_0.$$

Da die Schrittweitenanpassung hier global geschieht, es werden alle Reaktionen mit dem gleichen Zeitschritt $h^\star$ berechnet, wird zur Vereinfachung im Weiteren die konventionelle Notation verwendet.

Definition 2.42 *Das Verfahren (2.31) heißt Bruggeman Euler Verfahren und wird im Folgenden als BmE Verfahren bezeichnet.*

Man sieht direkt die folgende Eigenschaft.

Korollar 2.43 *Das BmE Verfahren ist nach Konstruktion konservativ nach Definition 2.36.*

Im Kontext ihrer Ausführungen schlagen Bruggeman et al. in [BBKS07] vor, trotz der „tatsächlich“ verwendeten Schrittweite $h^\star$, im weiteren Verlauf eine vollzogene Schrittweite von h anzunehmen.

p_0 wird so gewählt, dass die berechneten Näherungen positiv sind und Reaktionen die Richtung nicht ändern, d.h. für $f_i < 0$ sollte auch folgen, dass $c_i^{n+1} < c_i^n$ ist. Der folgende Satz klärt die Voraussetzungen an die Schrittweite zur Gewährleistung dieser beiden Bedingungen und präzisiert damit ein Ergebnis von Bruggeman et al.

Satz 2.44 *Wählt man* $p_0 \in I = \left(0, \min\left(1, \frac{1}{\max_{j\in J^n}|a_j|}\right)\right)$ *mit* $J^n = \{1 \leq i \leq N | f_i(\mathbf{c}^n) < 0\}$ *und* $a_j = \frac{h\, f_j(\mathbf{c}^n)}{{c_j}^n}$ *ist das BmE Verfahren (2.31) allgemein positiv und reaktionsrichtungserhaltend, d.h. Prozesse können nicht das Vorzeichen ändern.*

Beweis: Sei $i \in J^n$, d.h. speziell gilt $a_i < 0$. Dann erhält man aus (2.31)

$$ {c_i}^{n+1} = {c_i}^n + h\; p_0 f_i(\mathbf{c}^n) \Rightarrow \frac{{c_i}^{n+1}}{{c_i}^n} = 1 + p_0 \frac{h\; f_i(\mathbf{c}^n)}{{c_i}^n} = 1 + p_0\; a_i = 1 - p_0\; |a_i|. $$

Da $i \in J^n$ ist, sollte man sicherstellen, dass

$$ {c_i}^{n+1} < {c_i}^n $$

ist. Diese Forderung ist äquivalent zu

$$ \frac{{c_i}^{n+1}}{c_i^n} < 1. $$

Das ist eine direkte Folge aus

$$ 1 > 1 - \underbrace{p_0\; |a_i|}_{>0} = \frac{{c_i}^{n+1}}{c_i^n} $$

für $p_0 > 0$. Zusätzlich erwartet man auch

$$ {c_i}^{n+1} > 0. $$

Das lässt sich umformulieren zu

$$ \frac{{c_i}^{n+1}}{c_i^n} > 0. $$

Dafür betrachte man die beiden möglichen oberen Grenzen von I. Für $1 \leq \frac{1}{\max_{j\in J^n}|a_j|} \Leftrightarrow 1 \geq \max_{j\in J^n}|a_j|$ folgt $p_0 < 1$. Einsetzen liefert

$$1 - p_0\,|a_i| \geq 1 - \underbrace{p_0\,\max_{j\in J^n}|a_j|}_{<1} > 0.$$

Für $1 < \max_{j\in J^n}|a_j|$ folgt $p_0 < \frac{1}{\max_{j\in J^n}|a_j|}$. Einsetzen liefert

$$1 - p_0\,|a_i| > 1 - \underbrace{\frac{1}{\max_{j\in J^n}|a_j|}\,|a_i|}_{<1} \geq 0.$$

□

Das konkrete p_0 wird von Bruggeman et al. als eine Lösung der Gleichung

$$0 = \prod_{i\in J^n}(1 + a_i\,p_0) - p_0 \tag{2.32}$$

aus dem Intervall I bestimmt. Es gelten die Definitionen aus dem vorhergehenden Satz 2.44. In [BBKS07] wird ebenfalls gezeigt, dass das p_0 in dieser Weise bestimmt eindeutig ist.

Aus der Darstellung (2.31) ergibt sich sofort die folgende Aussage.

Korollar 2.45 *Das BmE Verfahren ist so wie das Euler Verfahren ein Verfahren erster Ordnung.*

Bruggeman formuliert auch eine Modifikation des Verfahrens von Heun

$$s_i^2 = c_i^n + h f_i(\mathbf{c}^n)$$

und

$$c_i^{n+1} = c_i^n + \frac{h}{2}\left(f_i(\mathbf{c}^n) + f_i(\mathbf{s}^2)\right)$$

durch

$$s_i^2 = c_i^n + p_0\,h\,f_i(\mathbf{c}^n)$$

und

$$c_i^{n+1} = c_i^n + q_0 q_1 \frac{h}{2}\left(f_i(\mathbf{c}^n) + f_i(\mathbf{s}^2)\right). \tag{2.33}$$

Hier ist $J^n = \{1 \leq j \leq N | a_j < 0\}$ und $a_j = h\frac{f_j(\mathbf{c}^n)+f_j(\mathbf{s}^2)}{2}$. Weiterhin ist $q_0 = \prod_{j\in J^n}\frac{c_j^n}{s_j^2}$ und $q_1 \in I$ eine Lösung der Gleichung $0 = \prod_{j\in J^n}(1 + a_j q_0 q_1) - q_1$.

Das so beschriebene Verfahren (2.33), im Folgenden mit BmH abgekürzt, ist zweiter Ordnung, allgemein positiv und konservativ. Die entsprechenden Beweise finden sich in [BBKS07, Zar05].

Welchen Vorteil die Wahl des Faktors p_0, wie in (2.32), gegenüber anderen auch konsistenten Wahlen $p_0 \in I$ hat, wird nicht begründet. Wobei man festhalten muss, dass die Lösung der Gleichung (2.32) einen ernst zu nehmenden Anteil am Gesamtrechenaufwand hat [BRBM07, BBKS07]. Demgegenüber wird hier kurz ein vereinfachtes Vorgehen beschrieben.

2.4.2 Minimalistische Schrittweitensteuerung

Zu Testzwecken sei an dieser Stelle eine alternative minimalistische Schrittweitensteuerung zu den Bruggeman Verfahren beschrieben.

Runge Kutta Verfahren entsprechen exakten Taylorentwicklungen bis zu einer gewissen Ordnung (siehe z.B. [HNW91]). Die modifizierten Patankar und Bruggeman Verfahren entsprechen nur im Grenzübergang $h \to 0$ den abgeschnittenen Taylorentwicklungen (vergleiche Unterkapitel 2.1 oder [Zar05]). Nach Definition sind alle Runge Kutta Verfahren konservativ (Bemerkung 2.41). Daher scheint es nahe liegend zu versuchen, die bekannten Runge Kutta Verfahren möglichst wenig und nur da, wo zur Wahrung der Positivität nötig, zu modifizieren.

Wählt man eine Sicherheitsschranke, die die maximale Reduzierung einer Größe in einem Zeitschritt auf z.B. $c_{\min} = 3\%$ ihres letzten Wertes gewährleistet, kann man aus der Form (2.31) und der oben beschriebenen Maxime eine minimalistische Schrittweitensteuerung konstruieren.

Es soll für alle i gelten, dass

$$c_i^n \; c_{\min} \leq c_i^{n+1} = c_i^n + h^\star f_i(\mathbf{c}^n).$$

Für eine Reduktion der Schrittweite sind alle $i \in J^n$ entsprechend der Definition des BmE Verfahrens (s. Satz 2.44) von Bedeutung.

Für alle $i \in J^n$ ergibt sich als maximale, den Anforderungen genügende, Schrittweite

$$h_i^\star = \frac{(1 - c_{\min})c_i^n}{|f_i(\mathbf{c}^n)|}. \tag{2.34}$$

Definition 2.46 *Das Euler Verfahren wie in (2.31) mit der modifizierten Schrittweite*

$$h^\star = \min_{i \in J^n} (1, h_i^\star)$$

mit $h_i^\star$ wie in (2.34) wird als Emini Verfahren *bezeichnet.*

Korollar 2.47 *Das Emini Verfahren ist allgemein positiv, konservativ und erster Ordnung.*

Dieses Verfahren unterliegt ähnlichen Einschränkungen der Schrittweite wie die zuvor geschilderten Bruggeman Verfahren. Auch dieses Verfahren ist nicht in der Lage steife Probleme mit praktikabler Schrittweite zu lösen. Numerische Experimente des Autors zu dem Robertson Testfall aus [Tes09] und dem Orego Problem (Beispiel 1.3) aus derselben Quelle lieferten instabiles Verhalten für Schrittweiten bis zu einer Größenordnung von $h(n) = 1.0002^n \cdot 10^{-15}$ und $h = 10^{-10}$ respektive.

2.4.3 Das verallgemeinerte modifizierte Patankar

Ziel des folgenden Abschnitts ist die Konstruktion und der Beweis eines Verfahrens, welches sowohl allgemeine Konservativität nach Definition 2.36 erhält, so wie es die in den vorhergehenden Abschnitten beschriebenen Bruggeman und Emini Verfahren leisten, aber im Gegensatz zu diesen zusätzlich auch in der Lage ist, wie die modifizierten Patankar Verfahren aus Unterkapitel 2.2 und 2.3 steife Probleme zu lösen. Dafür werden die Ansätze all dieser Verfahren mit ihren jeweiligen Schwächen noch einmal kurz zusammengefasst.

Sowohl die Bruggeman Verfahren als auch das Emini Verfahren verwenden einen einzigen Dämpfungsfaktor (p_0 und q_0 bzw. $h^\star$) für alle auftretenden Reaktionen und Prozesse. Im Gegensatz dazu verwendet das mPE Verfahren unterschiedliche Dämpfungsfaktoren $\left(\frac{\tilde{s}_j^v}{\tilde{s}_j^{v-1}} \text{ bzw. } \frac{c_j^{n+1}}{\tilde{s}_j^r}\right)$ für jede Reaktion r_k, welche nach Definition 2.12 jeweils eine Transformation von einer Größe c_i zu einer Größe c_j ist. Nur einen Dämpfungsfaktor zu verwenden, ermöglicht es den Bruggeman und dem Emini Verfahren unter anderem allgemeine Positivität zu erreichen; allerdings werden auch uneingeschränkte Größen nicht entsprechend der exakten Taylorentwicklung angenähert (nach Maßgabe des zugrunde liegenden Runge Kutta Verfahrens), sondern genauso in der Entwicklung gedämpft wie alle positiven Größen. Das ist nicht wünschenswert. Auch sind diese Verfahren nicht in der Lage, steife Probleme wie z.B. das Orego Problem (Beispiel 1.3) oder den Robertson Testfall aus [Tes09, HWN02] zu lösen. Letzteres ist ein Problem, welches den formalen Anforderungen des mPE Verfahrens genügt und von diesem auch bewältigt wird [BDM03, Zar05].

Nun folgt eine Betrachtung zu den Unterschieden der beiden Konservativitätsbegriffe (Definition 2.12 und 2.36). Grundsätzlich ist ein System patankarkonservativ, wenn jede Reaktion zwei Bedingungen erfüllt. Sie transformiert genau eine Größe in genau eine andere und die Masse, welche bei der einen abgebaut wird, wird der anderen ohne Modifikation zugeschlagen. Diese beiden Bedingungen werden vom mPE Verfahren erhalten und die berechneten Näherungen sind in diesem Sinne konservativ. Wenn das System diesen beiden Bedingungen aber nicht genügt, sondern Reaktionen aufweist, welche z.B. mehr als zwei Komponenten miteinander in Verbindung setzen, ist das mPE Verfahren nicht mehr anwendbar. Den Konflikt illustriert das folgende Beispiel.

Beispiel 2.48 *Es wird nochmal die Reaktion r_1 aus dem Beispiel 1.2 genauer betrachtet. Vergleiche dazu auch (2.26) und (2.27).*

Es gilt eingeschränkt auf r_1

$$\begin{pmatrix} C \\ N \\ P \end{pmatrix}' = \begin{pmatrix} -ar_1 \\ -br_1 \\ r_1 \end{pmatrix} = \underbrace{\begin{pmatrix} -a \\ -b \\ 1 \end{pmatrix}}_{S} \underbrace{r_1}_{r}. \tag{2.35}$$

Gälte $a, b > 0$ und $a+b = 1$, also $\sum_{i=1}^{3} S_i = 0$, so könnte man, ohne formal etwas zu ändern, die Reaktion r_1 additiv in r_{1a} und r_{1b} zerlegen. Vergleiche hierzu auch Bemerkung 2.9. Eine Möglichkeit wäre

$$r_{1a} = ar_1 \text{ und } r_{1b} = br_1 \text{ mit } r_1 = r_{1a} + r_{1b}.$$

Dann ließe sich (2.35) schreiben als

$$\begin{pmatrix} C \\ N \\ P \end{pmatrix}' = \begin{pmatrix} -r_{1a} \\ -r_{1b} \\ r_{1a} + r_{1b} \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix}}_{S} \underbrace{\begin{pmatrix} r_{1a} \\ r_{1b} \end{pmatrix}}_{\mathbf{r}}.$$

Das mPE Verfahren würde diese zwei Reaktionen jetzt entkoppelt betrachten und sie mit unterschiedlichen Faktoren, r_{1a} mit $\frac{C^{n+1}}{C^n}$ und entsprechend r_{1b} mit $\frac{N^{n+1}}{N^n}$, dämpfen. Das zerstörte aber die eigentlich modellierten stoichiometrischen Verhältnisse. Daher ist das mPE Verfahren selbst in dem seltenen Fall, dass $a+b = 1$ ist, nicht Erfolg versprechend anwendbar.

Ebenso ist das mPE Verfahren nicht auf Modelle anwendbar, in denen Quellen und Senken ohne jeweilige Entsprechungen (zu jeder Quelle eine Senke und umgekehrt) formuliert sind. Diese Einschränkungen sind (mindestens für konkrete Probleme) aufhebbar.

Die Idee ist also, im Gegensatz zu den Bruggeman Verfahren, für jede Reaktion r_k einen eigenen Modifikator β_k zu verwenden und so eine größere Stabilität zu erlangen, wie sie auch von den modifizierten Patankar Verfahren erhalten wird. Dies führt zu Verfahren des gestörten Runge Kutta Typs und ist entsprechend Bemerkung 2.41 konservativ.

Der Modifikator soll in Anlehnung an das mPE Verfahren definiert sein. Zwei Konventionen, die das Lesen des folgenden Formalismus erleichtern sollen, sind die Festlegung der Indizes. i steht immer für eine Größe, k steht immer für eine Reaktion. Auch sei noch einmal daran erinnert, dass immer N Größen und K Reaktionen betrachtet werden.

Da im allgemeinen Fall (2.2) nicht mehr klar ist, wie viele positive Größen durch eine Reaktion bedroht sind, sei dazu für jede Reaktion r_k die Indexmenge der möglicherweise bedrohten Größen

$$G_k = \{i \in \{1, \ldots, N\} \,|S_{ik} < 0 \textit{ und } c_i \textit{ ist positive Größe}\} \tag{2.36}$$

definiert. Zur allgemeinen Darstellung sind ebenso folgende Mengen hilfreich. Zu einem System (2.2) sei

$$P_i = \{k \in \{1, \ldots, K\} \,|S_{ik} > 0 \text{ und } \; G_k \neq \emptyset\}$$

bzw.

$$D_i = \{k \in \{1, \ldots, K\} \,|S_{ik} < 0 \text{ und } \; G_k \neq \emptyset\}$$

die Menge aller Indizes von Reaktionen, die für Größe c_i eine Quelle bzw. Senke und die für wenigstens eine positive Größe eine Senke sind. Die noch nicht berücksichtigten Beiträge $\neq 0$ zur i-ten Größe werden in

$$L_i = \{k \in \{1, \ldots, K\} \,|G_k = \emptyset \text{ und } \; S_{ik} \neq 0\}$$

zusammengefasst. Die Reaktionen dieser Mengen bedrohen keine positiven Größen und müssen daher nicht gedämpft werden. Daher definiert man noch

$$\bar{c_i}^n \;\; = \;\; c_i^n + h \sum_{k \in L_i} S_{ik} r_k \left(\mathbf{c}^n\right). \tag{2.37}$$

Es wird auch eine Einteilung der Indizes der N Größen benötigt. Man unterscheidet alle möglicherweise bedrohten positiven Größen,

$$N_1 = \{i \in \{1, \ldots, N\} | c_i \text{ ist positive Größe und } \exists k : S_{ik} < 0\}$$

und alle sicher unbedrohten Größen (alle anderen), $N_2 = \{1, \ldots, N\} \setminus N_1$. Äquivalente Definitionen der Mengen lauten

$$N_1 = \{i \in \{1, \ldots, N\} | \exists k : i \in G_k\}$$

und

$$N_2 = \{i \in \{1, \ldots, N\} | \nexists k : i \in G_k\}.$$

Nun wählt man

$$\beta_k = \begin{cases} 1 & G_k = \emptyset \\ 0 \text{ oder } \frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} & G_k \neq \emptyset \end{cases}, \tag{2.38}$$

wobei noch zu klären ist, welcher Index $\iota_k \in G_k$ zu wählen ist. Hierfür wird eine Heuristik herangezogen. Zur Berechnung dient

$$\begin{aligned} \iota_k &= \arg\min_{i \in G_k} \left\{ \frac{\bar{c}_i^n + h S_{ik} r_k(\mathbf{c}^n)}{\bar{c}_i^n} \right\} \\ &= \arg\min_{i \in G_k} \left\{ 1 + \frac{h S_{ik} r_k(\mathbf{c}^n)}{\bar{c}_i^n} \right\} = \arg\min_{i \in G_k} \left\{ \frac{S_{ik}}{\bar{c}_i^n} \right\}. \end{aligned} \tag{2.39}$$

Hierbei meint $\arg\min_{i \in G_k}$ das Element i aus G_k, für das der Ausdruck $\frac{S_{ik}}{\bar{c}_i^n}$ minimal ist. Es werden also Euler Näherungen mit der Reaktion r_k für alle $i \in G_k$ verwendet.

Des Weiteren weiß man noch nicht, wann $\beta_k = 0$ zu wählen ist; unzweifelhaft ist dieser Fall idealerweise zu vermeiden! Details dazu ergeben sich im weiteren Verlauf und werden entsprechend ausdrücklich erwähnt.

Definition 2.49 *Das* verallgemeinerte modifizierte Patankar Euler Verfahren *(vmPE) ist definiert durch*

$$c_i^{n+1} = c_i^n + h \sum_{k=1}^{K} S_{ik} r_k(\mathbf{c}^n) \beta_k,$$

mit dem Modifikator β_k wie (2.38) und ι_k wie in (2.39).

Die Fälle, in denen $\beta_k = 0$ gesetzt werden muss, sind stark problemabhängig. Sie werden hier nicht weiter allgemein erörtert, sondern im weiteren Verlauf der Untersuchungen genauer spezifiziert.

Das erste Ergebnis sieht man leicht.

Korollar 2.50 *Das vmPE ist nach Bemerkung 2.41 konservativ.*

Im Gegensatz zu den Bruggeman und dem Emini Verfahren steuert die Auswahl der Menge G_k im vmPE Verfahren die zu dämpfenden Reaktionen. Dies führt zu dem Ergebnis, dass für uneingeschränkte Größen, soweit wie möglich ohne die zu bewahrende Positivität der positiven Größen zu riskieren, die exakten Taylorentwicklungen als Näherungen verwendet werden.

Die konkrete Berechnung der Näherungen $\mathbf{c}^{n+1}$ findet durch die Lösung eines linearen Gleichungssystems

$$M\mathbf{c}^{n+1} = \bar{\mathbf{c}}^n, M \in \mathbb{R}^{N\times N} \tag{2.40}$$

statt.

Für das nun Folgende wird die Möglichkeit $\beta_k = 0$ vernachlässigt. Dies ist zulässig, da sich dadurch die Zuordnung der Terme $S_{ik}r_k\left(\mathbf{c}^n\right)\beta_k$ zu den Mengen L_i, P_i, D_i nicht ändert. Es erspart die ansonsten formal notwendige Unterscheidung zwischen $\beta_k = \frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n}$ und $\beta_k = 0$. Letzteres macht die Ausdrücke aber mathematisch nicht komplexer, es würde nur die auszuschließenden Ausnahmen in jeder Zeile erhöhen.

Die Gestalt von M lässt sich in der folgenden Art und Weise herleiten,

$$\begin{aligned} c_i^{n+1} &= c_i^n + h\sum_{k=1}^{K} S_{ik}r_k\left(\mathbf{c}^n\right)\beta_k \\ &= c_i^n + h\left[\sum_{k\in L_i} S_{ik}r_k\left(\mathbf{c}^n\right) + \sum_{k\in P_i} S_{ik}r_k\left(\mathbf{c}^n\right)\frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} - \sum_{k\in D_i} |S_{ik}|r_k\left(\mathbf{c}^n\right)\frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n}\right]. \end{aligned}$$

Stellt man dies nach den neuen Zuständen $\mathbf{c}^{n+1}$ um, resultiert unter Verwendung der Identität (2.37)

$$\bar{c_i}^n = c_i^{n+1} + h\left[\sum_{k\in D_i} |S_{ik}|r_k\left(\mathbf{c}^n\right)\frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} - \sum_{k\in P_i} S_{ik}r_k\left(\mathbf{c}^n\right)\frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n}\right].$$

Hieraus ergeben sich die Elemente der Matrix M durch

$$M_{ij} = \delta_{ij} + \frac{h}{c_j^n}\left(\sum_{k\in D_i,\iota_k=j} |S_{ik}|r_k\left(\mathbf{c}^n\right) - \sum_{k\in P_i,\iota_k=j} S_{ik}r_k\left(\mathbf{c}^n\right)\right). \tag{2.41}$$

Dies lässt sich weiter präzisieren. Die Spalte eines Eintrags gibt an, welches die vermeintlich am meisten bedrohte Größe ist. Die Zeile erklärt, welche Größe modifiziert wird. Für $i = j$ gilt

$$\sum_{k \in P_i, \iota_k = i} S_{ik} r_k(\mathbf{c}^n) = 0,$$

da aus $k \in P_i \overset{(2.36)}{\Rightarrow} i \notin G_k \overset{(2.39)}{\Rightarrow} i \neq \iota_k \Rightarrow \{k \in P_i | \iota_k = i\} = \emptyset$ folgt. Somit erhält man

$$M_{ii} \geq 1.$$

Sei $j \in N_2$. Dann zeigt ein ähnliches Argument für $j \in N_2 \overset{(2.4.3)}{\Rightarrow} \nexists k : j \in G_k \overset{(2.39)}{\Rightarrow} \nexists k : \iota_k = j$. Hieraus erhält man wie zuvor für $i \in N_1 \cup N_2$

$$\{k \in P_i | \iota_k = j\} = \{k \in D_i | \iota_k = j\} = \emptyset.$$

Man erhält hieraus

$$\sum_{k \in D_i, \iota_k = j} |S_{ik}| r_k(\mathbf{c}^n) = 0$$

und

$$\sum_{k \in P_i, \iota_k = j} S_{ik} r_k(\mathbf{c}^n) = 0.$$

Mit (2.41) ergibt sich zusammenfassend für $i \in N_1, j \in N_2$ und $i, j \in N_2, i \neq j$

$$M_{ij} = 0,$$

und für $i \in N_2$ folgt $M_{ii} = 1$.

M hat also eine Blockstruktur mit zwei Matrizen $X \in \mathbb{R}^{\#N_1 \times \#N_1}$ und $Y \in \mathbb{R}^{\#N_2 \times \#N_1}$ der Form

$$M = \left(\begin{array}{c|c} X & \mathbf{0} \\ \hline Y & \mathbf{1} \end{array}\right),$$

wobei $\#N_1$ und $\#N_2$ die Anzahl der Elemente der jeweiligen Menge bezeichnet. Man bestimmt $\mathbf{c}^{n+1}$ nun durch die formale Invertierung von M

$$\mathbf{c}^{n+1} = M^{-1} \bar{\mathbf{c}}^n.$$

Offensichtlich ist

$$M^{-1} = \left(\begin{array}{c|c} X^{-1} & \mathbf{0} \\ \hline -YX^{-1} & \mathbf{1} \end{array}\right).$$

M ist also invertierbar, wenn X invertierbar ist. Hierfür aber allgemeine Bedingungen aufzuschreiben ist sehr komplex und nur von bedingtem praktischen Nutzen. Daher wird nur für den konkreten Einzelfall bestimmt, ob eine Invertierung möglich ist. Ähnlich schwierig ist die allgemeine Abschätzung der Matrix M^{-1} (und damit der Nachweis der Ordnung). Die Ordnung wird sich in den Anwendungsbeispielen aber zeigen lassen.

Satz 2.51 *Das vmPE Verfahren ist von genau erster Ordnung, wenn die Matrix M^{-1} existiert und $M^{-1} = \mathbb{O}(1)$ ist.*

Beweis: Der Beweis verläuft entlang den Gedanken des entsprechenden Beweises zum mPE Verfahren in [BDM03, Zar05].

Nach Identität (2.40) gilt $\mathbf{c}^{n+1} = M^{-1}\bar{\mathbf{c}}^n$. Da $\bar{\mathbf{c}}^n \overset{(2.37)}{=} \mathbf{c}^n + \mathbb{O}(h)$ folgt $\mathbf{c}^{n+1} - \bar{\mathbf{c}}^n = (M^{-1} - I)\,\bar{\mathbf{c}}^n = \mathbb{O}(1)$.

Daraus erhält man $\mathbb{O}(1) = \frac{c_i^{n+1}}{c_i^n} - \frac{\bar{c}_i^n}{c_i^n} = \frac{c_i^{n+1}}{c_i^n} - \frac{c_i^n + \mathbb{O}(h)}{c_i^n} = \frac{c_i^{n+1}}{c_i^n} - 1$. Umstellen ergibt

$$\frac{c_i^{n+1}}{c_i^n} = \mathbb{O}(1).$$

Setzt man dies in die allgemeine Form des vmPE Verfahrens ein, erhält man für alle i

$$\begin{aligned} c_i^{n+1} &= c_i^n + h \underbrace{\left[\sum_{k \in L_i} S_{ik} r_k(\mathbf{c}^n) + \sum_{k \in P_i} S_{ik} r_k(\mathbf{c}^n) \frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} - \sum_{k \in D_i} |S_{ik}| r_k(\mathbf{c}^n) \frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} \right]}_{=\mathbb{O}(1)} \\ &= c_i^n + \mathbb{O}(h). \end{aligned}$$

Speziell gilt dies auch für $i = \iota_k$. Wiederum eine Umstellung liefert

$$\beta_k = \frac{c_{\iota_k}^{n+1}}{c_{\iota_k}^n} = 1 + \mathbb{O}(h).$$

Da das zugrunde liegende Euler Verfahren erster Ordnung ist, liefert Korollar 2.17 die Aussage. □

Es folgt eine weitere schlechte Nachricht. Für die allgemeine Form, in der das Verfahren definiert ist, lässt sich die Positivität nicht zeigen. Wenn man die Einschränkungen für das mPE Verfahren (siehe Unterkapitel 2.2) annimmt, erhalten die Matrizen M eine so stark festgelegte Gestalt, dass Positivität allgemein

gezeigt werden kann (siehe [BDM03, Zar05]). Lässt man einen Großteil der Forderungen fallen, wie zur Definition des vmPE geschehen, verliert man auch die mathematische Sicherheit, dass man im Allgemeinen die gewünschte Positivität für beliebige Schrittweiten h erhält. Für konkrete Testfälle (Beispiel 1.2 und 1.3) lassen sich aber, wie noch gezeigt werden wird, die gewünschte Positivität und Konservativität erzielen.

Satz 2.52 *Das vmPE liefert im Allgemeinen ohne die Berücksichtigung der Möglichkeit $\beta_k = 0$ uneingeschränkte Näherungen (vergleiche dazu Definition 2.4).*

Beweis: Sei für zwei positive Größen, c_1 und c_2, folgendes System betrachtet

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}' = S\mathbf{r} = \begin{pmatrix} m_{11} & -m_{12} \\ -m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} m_{11}r_1 - m_{12}r_2 \\ -m_{21}r_1 + m_{22}r_2 \end{pmatrix} \tag{2.42}$$

mit $m_{11}, m_{12}, m_{21}, m_{22} > 0$. Aus $G_1 = \{2\}$ folgt $\iota_1 = 2$ und entsprechend aus $G_2 = \{1\}$, dass $\iota_2 = 1$ ist. Weiterhin sieht man $P_1 = \{1\}$, $D_1 = \{2\}$ und $P_2 = \{2\}$, $D_2 = \{1\}$ sowie $L_1 = L_2 = \emptyset$. Abschließend ist festzuhalten, dass $N_1 = \{1, 2\}$ und $N_2 = \emptyset$ gilt.

Da beide Mengen G_1 und G_2 einelementig sind, gibt es nur eine mögliche Matrix M. Sie hat die Gestalt

$$M = \begin{pmatrix} 1 + \frac{h}{c_1} m_{12} r_2 & -\frac{h}{c_2} m_{11} r_1 \\ -\frac{h}{c_1} m_{22} r_2 & 1 + \frac{h}{c_2} m_{21} r_1 \end{pmatrix}$$

mit der Determinante

$$\begin{aligned} d &= \left(1 + \frac{h}{c_1} m_{12} r_2\right)\left(1 + \frac{h}{c_2} m_{21} r_1\right) - \frac{h}{c_2} m_{11} r_1 \frac{h}{c_1} m_{22} r_2 \\ &= 1 + \frac{h}{c_1} m_{12} r_2 + \frac{h}{c_2} m_{21} r_1 + \frac{h}{c_1} m_{12} r_2 \frac{h}{c_2} m_{21} r_1 - \frac{h}{c_2} m_{11} r_1 \frac{h}{c_1} m_{22} r_2 \\ &= 1 + \underbrace{\frac{h}{c_1} m_{12} r_2}_{>0} + \underbrace{\frac{h}{c_2} m_{21} r_1}_{>0} + \underbrace{\frac{h^2}{c_1 c_2} r_1 r_2}_{>0} \underbrace{\left(\underbrace{m_{12} m_{21}}_{>0} - \underbrace{m_{22} m_{11}}_{>0}\right)}_{\text{uneingeschränkt}}. \end{aligned} \tag{2.43}$$

Für $d \neq 0$ erhält man

$$M^{-1} = \frac{1}{d} \underbrace{\begin{pmatrix} 1 + \frac{h}{c_2} m_{21} r_1 & \frac{h}{c_2} m_{11} r_1 \\ \frac{h}{c_1} m_{22} r_2 & 1 + \frac{h}{c_1} m_{12} r_2 \end{pmatrix}}_{>\mathbf{0}}.$$

Da d für $d \neq 0$ sowohl positiv als auch negativ sein kann, ist auch M^{-1} entweder vollständig negativ oder positiv. Um Positivität zu garantieren müsste man $\beta_1 = \beta_2 = 0$ setzen. Dann hat man aber ein Verfahren für das $c_i^{n+1} = c_i^n$ gilt und das ist nur für ausgewählte Systeme richtig. □

Andererseits gibt das System (2.42) aber auch Grund zur Hoffnung.

Satz 2.53 *Ist das System (2.42) konservativ nach Definition 2.36 mit einer Komponentenzusammensetzungsmatrix E, sodass $E\,S = \mathbf{0}$ ist, dann ist das vmPE für dieses Problem uneingeschränkt positiv.*

Beweis: Nach Voraussetzung existiert ein Vektor $E = (e_1, e_2) \neq \mathbf{0}$ mit

$$ES = \mathbf{0}.$$

Hieraus ergibt sich aber, dass die Zeilen der Matrix S linear abhängig sind und somit

$$det(S) = 0$$

gilt, was äquivalent ist mit $m_{12}m_{21} - m_{22}m_{11} = 0$. Das bedeutet aber in (2.43) eingesetzt, dass M^{-1} existiert und $M^{-1} > 0$ ist. Damit ist alles gezeigt. □

Exemplarisch werden nun die Matrizen M bzw. M^{-1} für die Probleme aus den Beispielen 1.2 und 1.3 aufgestellt und abgeschätzt.

Sei dazu das erstgenannte Beispiel genauer untersucht.

Satz 2.54 *Das vmPE Verfahren ist angewendet auf das Modellproblem aus Beispiel 1.2 erster Ordnung.*

Beweis: Alle Größen des Problems sind positiv. Die Definition von $\mathbf{r}$ und S befindet sich in (2.27).

Die erste Reaktion r_1 hat zwei positive Größen, C und N, für die sie als Senke fungiert, d.h. $G_1 = \{1, 2\}$. Vergleiche dies mit den Einträgen $S_{11} = -a$ und $S_{21} = -b$ mit $a, b > 0$.

Die Reaktion r_2 besitzt nur eine positive Größe, für die sie eine Senke ist. Auch dies ergibt die Matrix S mit dem Eintrag $S_{32} = -1$. Man erhält $G_2 = \{3\}$.

Insgesamt ergibt sich $\iota_1 \in G_1 = \{1, 2\}$ und $\iota_2 \in G_2 = \{3\}$, was aber bedeutet, dass $\iota_2 = 3$ ist. Es gibt also zwei mögliche Kombinationen für ι_1 und ι_2. Diese Kombination wird durch die Wahl von ι_1 bestimmt, da ι_2 immer der gleiche

Ausdruck ist. Das bedeutet entsprechend gibt es zwei mögliche Formen der Matrix M, im Weiteren als M_1 und M_2 bezeichnet.

Seien ergänzend noch die weiteren zuvor definierten Mengen benannt. Es gilt für die Produktionsmengen $P_1 = \emptyset$, $P_2 = \emptyset$, $P_3 = \{1\}$ und $P_4 = \{2\}$. Die Destruktionsmengen sind gegeben durch $D_1 = \{1\}$, $D_2 = \{1\}$, $D_3 = \{2\}$ und $D_4 = \emptyset$. Für die uneingeschränkten Mengen gilt $L_1 = L_2 = L_3 = L_4 = \emptyset$. Es gilt also $\bar{\mathbf{c}}^n = \mathbf{c}^n$.

Betrachtet man nun als ersten Fall $\iota_1 = 1$, so gilt $\beta_1 = \frac{C^{n+1}}{C^n}$. Das bedeutet, dass die Verfahrensmatrix die folgende Gestalt hat.

$$M_1 = \begin{pmatrix} 1 + \frac{h}{C} a r_1 & 0 & 0 & 0 \\ \frac{h}{C} b r_1 & 1 & 0 & 0 \\ -\frac{h}{C} r_1 & 0 & 1 + \frac{h}{P} r_2 & 0 \\ 0 & 0 & -\frac{h}{P} r_2 & 1 \end{pmatrix}. \tag{2.44}$$

Im anderen Fall, es gilt also $\iota_1 = 2$, was $\beta_1 = \frac{N^{n+1}}{N^n}$ bedeutet, ist

$$M_2 = \begin{pmatrix} 1 & \frac{h}{N} a r_1 & 0 & 0 \\ 0 & 1 + \frac{h}{N} b r_1 & 0 & 0 \\ 0 & -\frac{h}{N} r_1 & 1 + \frac{h}{P} r_2 & 0 \\ 0 & 0 & -\frac{h}{P} r_2 & 1 \end{pmatrix}. \tag{2.45}$$

Diese beiden Matrizen lassen sich problemlos invertieren. Für

$$M_1^{-1} = \begin{pmatrix} \frac{C}{a r_1 h + C} & 0 & 0 & 0 \\ \frac{-b r_1 h}{a r_1 h + C} & 1 & 0 & 0 \\ \frac{r_1 h P}{(a r_1 h + C)(r_2 h + P)} & 0 & \frac{P}{r_2 h + P} & 0 \\ \frac{r_2 h^2 r_1}{(r_2 h + P)(a r_1 h + C)} & 0 & \frac{r_2 h}{r_2 h + P} & 1 \end{pmatrix} \tag{2.46}$$

sieht man sofort, dass dies wohldefiniert ist, falls alle Größen positiv sind, was nach Voraussetzung erfüllt ist. Weiter kann man jedes Matrixelement abschätzen. So ist

$$\left| \left(M_1^{-1}\right)_{31} \right|, \left| \left(M_1^{-1}\right)_{41} \right| \leq \frac{1}{a}$$

und

$$\left| \left(M_1^{-1}\right)_{21} \right| \leq \frac{b}{a}.$$

Für die anderen Elemente gilt $\left|\left(M_1^{-1}\right)_{ij}\right| \leq 1$. Ebenso lässt sich dies für die Matrix M_2^{-1} zeigen. Es folgen analoge Abschätzungen mit vertauschten Rollen von a und b. Auf die Darstellung der Rechnung wird hier aber verzichtet. Womit man insgesamt gezeigt hat, dass M^{-1} existiert und $M^{-1} = \mathbb{O}(1)$ ist, was den Beweis unter Verwendung von Satz 2.51 abschließt. □

Es lässt sich darüber hinaus zeigen, dass das vmPE, ohne Berücksichtigung eines möglichen Setzens von $\beta_k = 0$, für dieses Problem uneingeschränkt positiv ist.

Satz 2.55 *Das vmPE Verfahren liefert unter Vernachlässigung der Möglichkeit $\beta_k = 0$ mit der Definition 2.49 und der strikten Wahl*

$$\beta_1 = \frac{c_{\iota_1}^{n+1}}{c_{\iota_1}^{n}} \text{ und } \beta_2 = \frac{c_3^{n+1}}{c_3^{n}},$$

und ι_1 wie in (2.39) für das Problem (1.3) uneingeschränkt positive Näherungen.

Beweis: In Analogie zum Beweis des vorhergehenden Satzes 2.54 sind wiederum zwei Fälle zu betrachten. Alle dort verwendeten Definitionen und Schreibweisen finden wieder Anwendung. Die beiden Fälle gliedern sich wie dort nach der Wahl von ι_1.

Für den ersten Fall betrachte $\iota_1 = 1$. Das bedeutet mit (2.39), dass

$$\frac{-a}{C^n} \leq \frac{-b}{N^n}$$

ist, was wiederum folgende Äquivalenzen hat

$$\frac{a}{C^n} \geq \frac{b}{N^n} \Leftrightarrow aN^n \geq bC^n \Leftrightarrow aN^n - bC^n \geq 0.$$

Der einzige negative Ausdruck in der Matrix M_1^{-1}, siehe (2.46), ist das Indexpaar $(2,1)$. Insofern könnte höchstens die Größe N^{n+1} negativ werden. Dies ist konsistent mit der Tatsache, dass die Reaktion r_1 die Größe N bedroht, von dieser

aber nicht direkt gedämpft wird. Schaut man sich N^{n+1} genauer an, sieht man

$$\begin{aligned} N^{n+1} &= \left(\frac{-br_1h}{ar_1h+C^n}, 1, 0, 0\right) \begin{pmatrix} C^n \\ N^n \\ P^n \\ D^n \end{pmatrix} \\ &= N^n - C^n \frac{br_1h}{ar_1h+C^n} \\ &= \frac{N^n(ar_1h+C^n) - br_1hC^n}{ar_1h+C^n} \\ &= \frac{\overbrace{N^nC^n}^{>0} + \overbrace{r_1h}^{>0}\overbrace{(aN^n - bC^n)}^{\geq 0}}{\underbrace{ar_1h+C^n}_{>0}} \\ &> 0. \end{aligned}$$

Die Behandlung des Falls $M = M_2$ erfolgt analog. Somit ist alles gezeigt. □

Führt man eine ähnliche Analyse für das Orego Problem (Beispiel 1.3) durch, so stößt man leider auf einige Schwierigkeiten. Die Ordnung kann man noch uneingeschränkt zeigen. Bei der Positivität erweist es sich, dass man die Möglichkeit $\beta_k = 0$ (wenigstens theoretisch) nicht vollständig vernachlässigen kann.

Zur Vereinfachung und größeren Übersicht wird vorab der eingeführte Formalismus auf das Orego Problem angewendet. Die Definitionen der Stoichiometriematrix S und des Reaktionsvektors $\mathbf{r}$ finden sich in (1.5). Die zugehörigen Mengen zu den Reaktionen sind

$$G_1 = \{1\}, G_2 = \{2\}, G_3 = \{3\}, G_4 = \{1,2\} \text{ und } G_5 = \emptyset. \tag{2.47}$$

Die Mengen zu den Größen sind gegeben durch $P_1 = \{2\}$, $P_2 = \{3\}$, $P_3 = \emptyset$, sowie $D_1 = \{1,4\}$, $D_2 = \{2,4\}$, $D_3 = \{3\}$ und abschließend $L_1 = L_3 = \{5\}$ und $L_2 = \emptyset$. Aus der Gestalt der Mengen L_i erkennt man, dass $\bar{\mathbf{c}}^n \neq \mathbf{c}^n$ ist, vergleiche dazu (2.37).

Satz 2.56 *Das vmPE Verfahren unter Verwendung der Heuristik (2.39) angewendet auf das Orego Problem (1.4) ist exakt erster Ordnung.*

Beweis: Wieder ist es laut Satz 2.51 ausreichend, die Wohldefiniertheit von M^{-1} und $M^{-1} = \mathbb{O}(1)$ zu zeigen. Dafür werden alle möglichen Matrizen M

untersucht. Zur Zeichenersparnis gilt $c_1 = c_1^n, c_2 = c_2^n, c_3 = c_3^n$. Weiterhin werden zur Abkürzung die Reaktionen nummeriert, $r_1 = qc_1^2, r_2 = c_2, r_3 = c_3$ und $r_4 = c_1c_2$. Beachte hier, in der Reaktion r_1 ist der Exponent zu c_1 tatsächlich eine Potenz, ebenso der Exponent zur Konstanten s in den weiteren Rechnungen.

Eine Betrachtung der Mengen G_k aus (2.47) ergibt, dass es für die Wahl der ι_k nur eine echte Wahl gibt. Dies ist die Wahl des ι_4; alle anderen ι_k sind als Elemente der zugehörigen Menge G_k festgelegt, da diese bis auf G_4 alle einelementig oder leer sind.

Man hat es hier also auch, wie beim vorher untersuchten Problem, Beispiel 1.2, mit zwei möglichen Matrizen M_1, falls c_1 das begrenzende Element ist (also $\iota_4 = 1$) und M_2, falls c_2 begrenzend wirkt (entsprechend $\iota_4 = 2$), zu tun. Man findet

$$M_1 = \begin{pmatrix} \frac{s(r_1+r_4)h}{c_1} + 1 & -\frac{sr_2h}{c_2} & 0 \\ \frac{r_4h}{sc_1} & \frac{r_2h}{sc_2} + 1 & -\frac{r_3h}{sc_3} \\ 0 & 0 & \frac{wr_3h}{c_3} + 1 \end{pmatrix} \tag{2.48}$$

und

$$M_2 = \begin{pmatrix} \frac{sr_1h}{c_1} + 1 & sh\frac{r_4-r_2}{c_2} & 0 \\ 0 & h\frac{r_4+r_2}{sc_2} + 1 & -\frac{r_3h}{sc_3} \\ 0 & 0 & \frac{wr_3h}{c_3} + 1 \end{pmatrix}. \tag{2.49}$$

Formale Invertierung durch Maple 11.0 liefert

$$\begin{aligned} M_1^{-1} &= \begin{pmatrix} \frac{c_1(r_2h+sc_2)}{d} & \frac{c_1hr_2s^2}{d} & \frac{r_2h^2r_3c_1s}{(wr_3h+c_3)d} \\ -\frac{hr_4c_2}{d} & \frac{c_2s(sr_1h+shr_4+c_1)}{d} & \frac{(sr_1h+shr_4+c_1)r_3hc_2}{(wr_3h+c_3)d} \\ 0 & 0 & \frac{c_3}{wr_3h+c_3} \end{pmatrix} \\ &= \begin{pmatrix} \frac{d_5+d_6}{d} & \frac{d_5s^2}{d} & \frac{hr_3s\,d_5}{(wr_3h+c_3)d} \\ -\frac{d_4}{s^2\,d} & \frac{d_3+d_4+d_6}{d} & \frac{(d_3+d_4+d_6)r_3h}{s(wr_3h+c_3)d} \\ 0 & 0 & \frac{c_3}{wr_3h+c_3} \end{pmatrix} \end{aligned} \tag{2.50}$$

mit

$$d = \underbrace{2r_4h^2r_2s}_{=:d_1} + \underbrace{sr_1h^2r_2}_{=:d_2} + \underbrace{s^2r_1hc_2}_{=:d_3} + \underbrace{s^2hr_4c_2}_{=:d_4} + \underbrace{c_1r_2h}_{=:d_5} + \underbrace{c_1sc_2}_{=:d_6}$$

und

$$M_2^{-1} = \begin{pmatrix} \frac{c_1}{sr_1h+c_1} & \frac{c_1h(r_2-r_4)s^2}{(r_2h+r_4h+sc_2)(sr_1h+c_1)} & \frac{(r_2-r_4)h^2r_3c_1s}{(sr_1h+c_1)(r_2h+r_4h+sc_2)(wr_3h+c_3)} \\ 0 & \frac{sc_2}{r_2h+r_4h+sc_2} & \frac{c_2r_3h}{(wr_3h+c_3)(r_2h+r_4h+sc_2)} \\ 0 & 0 & \frac{c_3}{wr_3h+c_3} \end{pmatrix}. \quad (2.51)$$

Da alle Größen positiv sind, sind alle Nenner ungleich Null und damit die Inversen wohldefiniert. Weiter lassen sich diese Matrizen nun elementweise abschätzen. Dafür geht man wie folgt vor. Die Konstanten s und w sind problemabhängig. Über alle anderen Größen weiß man nur, dass sie positiv sind. Nun ist es das Ziel, die Brüche in die Form $const\frac{1}{1+\epsilon}, \epsilon > 0$ zu bringen. Dafür vergleicht man die Zählerterme mit den Nennertermen. Für das Indexpaar (1,2) ergibt sich die Abschätzung durch d_5. Man erhält

$$\left(M_1^{-1}\right)_{12} = s^2\frac{d_5}{d} = s^2\frac{\overbrace{1}^{<1}}{1+\underbrace{\left(\frac{d-d_5}{d_5}\right)}_{>0}} < s^2,$$

da $d > d_5 > 0$ ist. In gleicher Weise gewinnt man auch die Abschätzungen der anderen Matrixelemente

$$0 \leq \left(M_1^{-1}\right)_{13} \leq \frac{s}{w},$$

$$0 > \left(M_1^{-1}\right)_{21} \geq \frac{-1}{s^2},\ 0 \leq \left(M_1^{-1}\right)_{23} \leq \frac{1}{s},$$

$$0 \leq \left|\left(M_2^{-1}\right)_{12}\right| \leq s^2,\ 0 \leq \left|\left(M_2^{-1}\right)_{13}\right| \leq \frac{s}{w} \text{ und}$$

$$0 \leq \left(M_2^{-1}\right)_{23} \leq \frac{1}{sw}.$$

Für alle weiteren Elemente gilt $0 \leq \left(M_1^{-1}\right)_{ij}, \left(M_2^{-1}\right)_{ij} \leq 1$.

Auch hier existieren die Inversen und es ergibt sich ebenso $M_1^{-1} = \mathbb{O}(1)$ und $M_2^{-1} = \mathbb{O}(1)$, was den Beweis abschließt. □

Um die Positivität zu garantieren, muss man die Inversen aus dem letzten Beweis weiter untersuchen.

Satz 2.57 *Das vmPE Verfahren unter Verwendung der Heuristik (2.39) angewendet auf das Orego Problem (1.3) ist uneingeschränkt positiv, falls $\iota_4 = 1$ oder $\beta_4 = 0$.*

Beweis: Hierfür müssen wiederum zwei Fälle unterschieden werden. Auch in diesem Beweis wird wieder $c_1 = c_1^n, c_2 = c_2^n, c_3 = c_3^n$ zur Zeichenersparnis gesetzt.

Betrachtet man den ersten Fall, $\iota_4 = 1$, und wählt in Übereinstimmung mit der Definition (2.38) $\beta_4 = \frac{c_1^{n+1}}{c_1^n}$, so ergibt sich aus der Heuristik (2.39), dass

$$\frac{-s}{\bar{c}_1} \leq \frac{-1}{s\bar{c}_2}$$

gilt, was wegen $\bar{c}_i > 0$ die folgenden Äquivalenzen besitzt

$$\frac{s}{\bar{c}_1} \geq \frac{1}{s\bar{c}_2} \Leftrightarrow s^2\bar{c}_2 \geq \bar{c}_1 \Leftrightarrow s^2\bar{c}_2 - \bar{c}_1 \geq 0. \tag{2.52}$$

Erinnere, M_1^{-1} steht in (2.50). Die einzige Komponente, die bei der Operation $M_1^{-1}\bar{\mathbf{c}}^n$ negativ werden könnte, ist c_2^{n+1}, da nur das Element $\left(M_1^{-1}\right)_{21}$ negativ ist. Betrachte dazu unter Verwendung der Abschätzung (2.52) und der Erweiterung mit dem kleinsten gemeinsamen Vielfachen der Nenner der zweiten Zeile

$$\begin{aligned}
& c_2^{n+1}(wr_3h + c_3)\, d\, s^2 \\
&= (wr_3h + c_3)\, d\, s^2 \sum_{i=1,2,3} \left(M_1^{-1}\right)_{2i} \bar{c}_i \\
&= -\bar{c}_1 d_4(wr_3h + c_3) + \bar{c}_2 s^2(d_3 + d_4 + d_6)(wr_3h + c_3) + \bar{c}_3\,(d_3 + d_4 + d_6)r_3hs \\
&= d_4\left(-\bar{c}_1(wr_3h + c_3) + s^2\bar{c}_2(wr_3h + c_3)\right) + \underbrace{d_4 s\bar{c}_3 r_3 h}_{>0} + \epsilon \\
&> d_4\left(\underbrace{(-\bar{c}_1 + s^2\bar{c}_2)}_{\geq 0}(wr_3h + c_3)\right) \\
&\geq 0,
\end{aligned}$$

wobei $\epsilon \in \mathbb{R}^+$ alle nicht weiter explizit aufgeführten Terme mit d_3 oder d_6 als Faktor beinhaltet. Damit ist der erste Fall abgehandelt und die uneingeschränkte Positivität der berechneten Näherungen für diesen Fall gezeigt.

Sei nun andererseits $\iota_4 = 2$, so erhält man wiederum aus (2.39) die äquivalenten Abschätzungen

$$\frac{-s}{\bar{c}_1} > \frac{-1}{s\bar{c}_2} \Leftrightarrow \frac{s}{\bar{c}_1} < \frac{1}{s\bar{c}_2} \Leftrightarrow s^2\bar{c}_2 < \bar{c}_1 \Leftrightarrow \bar{c}_1 - s^2\bar{c}_2 > 0.$$

Verwendet man aber $\beta_4 = \frac{c_2^{n+1}}{c_2^n}$, im Gegensatz zur Forderung $\beta_4 = 0$, so ergibt sich die vollständige Matrix M_2^{-1}, wie sie sich in (2.51) findet. Man sieht, dass die beiden möglicherweise negativen Elemente von M_2^{-1} zu den Indexpaaren $(1, 2)$ und

$(1,3)$ gehören. Daher untersucht man $c_1^{n+1} = \left(M_2^{-1}\bar{\mathbf{c}}^n\right)_1$ erweitert um den maximalen Nenner der Brüche aus der ersten Zeile der Matrix M_2^{-1}. Dies ausführend erhält man

$$\begin{aligned}
&c_1^{n+1}(sr_1h + c_1)(r_2h + r_4h + sc_2)(wr_3h + c_3)\\
&= (sr_1h + c_1)(r_2h + r_4h + sc_2)(wr_3h + c_3)\sum_{i\in 1,2,3}\left(M_2^{-1}\right)_{1i}\bar{c}_i\\
&= \bar{c}_1c_1(\underline{r_2h} + r_4h + sc_2)(wr_3h + c_3) + \bar{c}_2c_1h(\underline{r_2} - r_4)s^2(wr_3h + c_3)\\
&\quad +\bar{c}_3(\underline{r_2} - r_4)h^2r_3c_1s\\
&> \bar{c}_1c_1r_4h(wr_3h + c_3) + \bar{c}_1c_1sc_2(\underline{wr_3h} + c_3) - \bar{c}_2c_1hr_4s^2(wr_3h + c_3)\\
&\quad -\bar{c}_3r_4h^2r_3c_1s\\
&> \underline{c_1r_4h(wr_3h + c_3)\overbrace{\left(\bar{c}_1 - \bar{c}_2s^2\right)}^{>0}} + \bar{c}_1\underbrace{c_1c_2}_{=r_4}sc_3 - \bar{c}_3r_4h^2r_3c_1s\\
&= r_4s\left(\bar{c}_1c_3 - \bar{c}_3c_1h^2r_3\right),
\end{aligned}$$

wobei die jeweils unterstrichenen (positiven) Terme zur nächsten Zeile entfallen.

Man erkennt zwei Dinge. Zuallererst ist das Verfahren in dieser Form nicht uneingeschränkt positiv. Das kann man aber erzwingen, indem man $r_4 = 0$ wählt, da sich in allen negativen Termen r_4 als Faktor findet. Dies geschieht explizit durch die Wahl $\beta_4 = 0$. Beachte weiterhin, für $r_4 = 0$ gilt $M_1 = M_2$ mit der entsprechenden Gleichheit für die Inversen.

Als zweites sieht man aber auch, dass die „negativen Anteile“ für $h \to 0$ ein $\mathbb{O}(h^2)$ sind. Man muss also für die Asymptotik zum Beweis der Ordnung keine weitere Matrix mit $r_4 = 0$ betrachten und die negative Störung ist über die Schrittweite kontrollierbar. □

Für die Programmierung ergibt sich also die Notwendigkeit eine Ausnahme für den Fall $M = M_2$ und $c_1^{n+1} \leq 0$ vorzusehen. Diese lässt sich aber problemlos über die Wahl $\beta_4 = 0$, d.h. $\beta_4 r_4 = 0$ abhandeln. Für die Ordnung des Verfahrens ergibt sich durch dieses „Abschneiden“ kein weiteres Problem, da c_1^{n+1} für $h \to 0$ garantiert positiv wird.

An dieser Stelle sei aber explizit darauf hingewiesen, dass keine der im Weiteren gezeigten Rechnungen bei voll besetzter Matrix, d.h. $\beta_k \neq 0$ gewählt, ent-

sprechend der Heuristik (2.39)

$$\beta_4 = \begin{cases} \frac{c_1^{n+1}}{c_1^n} & , \quad \text{für } \frac{-s}{\bar{c}_1^n} \leq \frac{-1}{s\bar{c}_2^n} \\ \frac{c_2^{n+1}}{c_2^n} & , \quad \text{sonst} \end{cases} ,$$

negative Werte lieferte.

Alle Testrechnungen der in diesem Kapitel vorgestellten Verfahren zu den Modellproblemen aus Kapitel 1 befinden sich in Kapitel 5.

Teil II

Zur erweiterten 2 D Flachwassergleichung mit Phosphorzyklus

Kapitel 3

Modellbildung zur erweiterten zweidimensionalen Flachwassergleichung

Um die modifizierten Patankar Verfahren im Kontext der Seenmodellierung untersuchen zu können, bedarf es eines Testfalls. Das vorliegende Kapitel widmet sich der Beschreibung des in dieser Arbeit verwendeten Problems. Es besteht einerseits aus der zweidimensionalen Flachwassergleichung und andererseits aus einem Phosphorzyklus.

Das erste Unterkapitel widmet sich der Darstellung der zweidimensionalen Flachwassergleichung und eines Dammbruchproblems aus [Tor01b] für die reine Flachwassergleichung als Testfall des reinen Strömungslösers.

Das Folgende umfasst eine kurze Beschreibung verschiedener Seenmodelle und legt den Schwerpunkt auf die, von ihnen beschriebenen, ökologischen Prozesse und vernachlässigt in den Modellen möglicherweise vorhandene Berücksichtigungen der Hydrodynamik. Ebenso findet sich dort eine Übersicht zum verwendeten Phosphorzyklus (die Details finden sich in Anhang A). Es schließen sich einige Bemerkungen zu typischen Modellierungstermen ökologischer Phänomene an.

Abschließend findet sich die Zusammenführung des Gesamtsystems sowie der verwendeten Rand- und Anfangswerte.

3.1 Die Flachwassergleichung

Die zweidimensionale Flachwassergleichung (in der englischen Literatur 2D shallow water equation, daher kurz *swe*) ist ein nicht lineares System von hyperbolischen Erhaltungsgleichungen. Sie beschreibt die Änderungen der Strömungsgeschwindigkeiten und Wasserhöhe in Abhängigkeit der Fließgeschwindigkeiten, der Wasserhöhe, den Volumenkräften (zu meist nur der Gravitation) sowie Quelltermen. Die Quellterme können je nach Anwendung Ausdrücke für Niederschlag und Verdunstung, Bodenprofil, Windschub, Bodenreibung, Corioliskräfte oder auch Turbulenzmodelle enthalten. Beispiele zu letzterem finden sich unter anderem in [VC07].

Die Gleichungen werden üblicherweise aus den Eulergleichungen der Gasdynamik hergeleitet. Die vereinfachenden Annahmen sind,

- die Dichte ist konstant und
- die relevante Skala für die vertikalen Prozesse (z) ist deutlich kleiner als die entsprechende Skala für die horizontalen Abläufe (x).

Detaillerte Herleitungen finden sich zum Beispiel in [Lev02, Tor01b]. Eine ausführliche Herleitung durch eine asymptotische Entwicklung nach $\sigma = \frac{z^2}{x^2}$ findet sich in [Sto57], welches sich aber auch noch anderen Arten der Herleitung umfassend widmet. Die mathematische Struktur der swe ist äquivalent zu den isentropen Eulergleichungen der Gasdynamik mit dem adiabatischen Faktor zwei (siehe z.B. [Lev06]).

3.1.1 Die mathematische Form der zweidimensionalen Flachwassergleichung

In dieser Arbeit wird die swe in der folgenden Form für einen Vektor der konservativen Größen $\mathbf{u}$ verwendet,

$$\partial_t \mathbf{u}_{swe} + \partial_{x_1} \mathbf{f}_1(\mathbf{u}_{swe}) + \partial_{x_2} \mathbf{f}_2(\mathbf{u}_{swe}) = \mathbf{r}_{swe}(\mathbf{u}_{swe}), \tag{3.1}$$

wobei ∂_t und ∂_{x_i} die partiellen Ableitungen nach der Zeit t respektive den beiden horizontalen Raumkoordinaten x_1, x_2 bezeichnen. Es gilt

$$\mathbf{u}_{swe} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} \Phi \\ \Phi\ v_1 \\ \Phi\ v_2 \end{pmatrix}, \mathbf{f}_i(\mathbf{u}_{swe}) = \begin{pmatrix} u_{i+1} \\ \frac{u_2\ u_{i+1}}{u_1} + \delta_{1i}\frac{u_1^2}{2} \\ \frac{u_3\ u_{i+1}}{u_1} + \delta_{2i}\frac{u_1^2}{2} \end{pmatrix}.$$

Weiterhin ist $\Phi = g\ H$ mit der Gravitationsbeschleunigung $g = 7.32312576 \cdot 10^{10}\ \frac{m}{d^2}$ bzw. $g = 9.81\ \frac{m}{s^2}$ (je nach Anwendungsfall, ökologisches Problem oder Dammbruchproblem) und H der tatsächlichen Wasserhöhe, δ_{ij} das Kroneckerdelta und v_i die Fließgeschwindigkeit in x_i Richtung. Häufig wird Φ als Geopotential bezeichnet. Die Einheit für $[g] = \frac{m}{d^2}$ bzw. $[g] = \frac{m}{s^2}$ wird jeweils gewählt, um der Zeiteinheit für die Ausgabe der Ergebnisse des noch zu beschreibenden ökologischen Modells (Unterkapitel 3.2) bzw. Dammbruchproblems (Unterkapitel 3.1.2) Rechnung zu tragen. Damit ergeben sich die Einheiten für $[v_i] = \frac{m}{d}$ bzw. $[v_i] = \frac{m}{s}$ und $[\Phi] = \frac{m^2}{d^2}$ bzw. $[\Phi] = \frac{m^2}{s^2}$.

Die im Rahmen des ökologischen Modells (Unterkapitel 3.2) berücksichtigten Quellterme lauten

$$\mathbf{r}_{swe}(\mathbf{u}_{swe}) = \begin{pmatrix} g\ q \\ g\ q\ v_1 \\ g\ q\ v_2 \end{pmatrix}. \tag{3.2}$$

Hierbei bezeichnet q den Niederschlag bzw. die Verdunstung mit der Einheit $[q] = \frac{m}{d}$. Das bedeutet, dass die Änderungen der Wasserhöhe und der Geschwindigkeiten zusätzlich zu den Transporttermen nur von Regen und Verdunstung abhängig angenommen werden. Auf die Wasserhöhe H, bzw. das Geopotential $\Phi = g\,H$, wirkt sich q direkt aus. Für die Änderungen der Geschwindigkeiten gilt die Annahme, dass sich der auf das Wasser auftreffende Regen direkt mit den Fließgeschwindigkeiten der Strömung bewegt, da das Wasser als vollständig vertikal durchmischt angenommen wird.

Mögliche Einflüsse durch Bodenprofil, Turbulenz, Corioliskraft, Windschub, Bodenreibung o.a., werden vernachlässigt. Dies rechtfertigt sich für das ökologische Modell (Unterkapitel 3.2) über sehr geringe Fließgeschwindigkeiten im betrachteten Problem einerseits und andererseits darüber, dass der betrachtete See von so geringer Fläche ist und somit die Erdrotation keinen großen Einfluss hat. Des Weiteren sei eine ruhige Wetterlage ohne signifikante Winde angenommen.

Für das Dammbruchproblem (Unterkapitel 3.1.2) ergibt sich die Vernachlässigung aller Quellterme aus der Referenzimplementierung. Da die Ergebnisse, des noch vorzustellenden Strömungslösers (Unterkapitel 4.5), mit Näherungen, berechnet mit Hilfe des Sourcecodes von Toro nach [Tor01b], verglichen werden sollen und in dieser Referenzimplementierung keine Quellen verwendet werden, geschieht dies hier auch nicht.

Damit ist die rechte Seite $\mathbf{r}_{swe}$ der Gleichung vollständig beschrieben.

3.1.2 Ein kreisförmiges Dammbruchproblem

Um die Fähigkeiten des noch vorzustellenden Strömungslösers aus Kapitel 4 unabhängig von anderen Einflüssen beobachten zu können, wird ein Testfall für die reine swe (3.1) aus [Tor01b] verwendet.

Die Fragestellung ist so gewählt, dass die Ergebnisse des Strömungslösers (Kapitel 4) mit den Ergebnissen des Strömungslösers von E. F. Toro [Tor01b] verglichen werden können. Die Ergebnisse des Strömungslösers von Toro stehen nach freundlicher Freigabe des Sourcecodes durch den Autor zur Verfügung.

Ein weiterer Vorteil dieses Problems ist, dass es zweidimensional gestellt ist, die Lösung aber in einer Dimension vollständig beschrieben werden kann. Dies vereinfacht die Darstellung und den Vergleich der Näherungen.

In einem See mit quadratischem Grundriss, gleichmäßiger Wasserhöhe von 0.5 Metern und ohne Bodenprofil befindet sich in der Mitte des Sees eine Wassersäule mit 2.5 Metern Durchmesser und einer Wasserhöhe von ebenfalls 2.5 Metern.

Das Rechengebiet ist quadratisch mit einer Kantenlänge von 40 Metern. Es wird in der eigenen Implementierung mit einer Sekundärnetzmethode über einem unstrukturierten Gitter mit 46626 Zellen diskretisiert.

Die Anfangswerte sind

$$v_1(x, y, 0) = v_2(x, y, 0) = 0$$

und

$$H(x, y, 0) = \begin{cases} 0.5 & , \quad (x - x_c)^2 + (y - y_c)^2 > 2.5^2 \\ 2.5 & , \quad (x - x_c)^2 + (y - y_c)^2 \leq 2.5^2 \end{cases}$$

mit $x_c = y_c = 20$.

An den Rändern werden Neumann Randbedingungen durch

$$\frac{\partial \mathbf{u}_{swe}}{\partial \mathbf{n}} = \mathbf{0} \tag{3.3}$$

festgelegt, wobei die Rechnung beendet wird, bevor die erste Welle mit dem Rand interagiert.

Als Referenz dient die Näherung nach Toro [Tor01b]. Dieser Code verwendet strukturierte Gitter. Er rechnet mit 40000 Zellen, da diese in der Darstellung des Autors identische Ergebnisse wie eine Rechnung mit 10^6 Zellen liefert.

3.2 Die Phosphor- und Biomassedynamik

Um die Anwendbarkeit der vorgestellten modifizierten Patankar Euler und Heun Verfahren (Unterkapitel 2.2) und deren Erweiterungen im Bereich der ökologischen Seenmodellierung zeigen zu können, wird die swe mit einem Phosphorzyklus gekoppelt.

Dieses Unterkapitel widmet sich der Vorstellung einiger verbreiteter Seenmodelle und einer Untersuchung ihrer Eignung als Testfälle für die modifizierten Patankar Verfahren. Anschließend wird das Modell von Hongping und Jianyi aus [HJ02] kurz beschrieben. Dieses Modell dient als Basis für das in dieser Arbeit verwendete System. Die Details des verwendeten Modells sind im Anhang A zusammengefasst.

3.2.1 Die Modellauswahl

Der Vorauswahl lagen die Übersichtsartikel [BH07, AO03, SMB+04] zu Grunde. Die konkrete Auswahl des Modells erfolgte aus mathematisch praktischen Überlegungen sowie Gedanken zur biologischen und ökologischen Plausibilität.

Da dies eine Grundlagenarbeit ist und kein konkretes ökologisches Szenario vorlag, wurde das Modell als besser eingestuft, welches komplexere Phänomene modelliert. Die Algorithmen sollten auf ihre Eignung für realistische Fragestellungen hin untersucht werden. Somit sollte ein Modell aus der ökologischen Praxis Anwendung finden.

Die technische Seite stellte zwei Bedingungen an das verwendete Modell. Die Klasse der modifizierten Patankar Verfahren und deren Ordnungserweiterung sollten auf das Modell anwendbar sein. Damit war es nötig, ein Modell auszuwählen, das einen vergleichsweise einschränkenden Begriff der Konservativität (vergl. Definition 2.5) erfüllte oder aber sich mit wenig Aufwand in die entsprechende Form umformulieren ließ, ohne dabei die grundlegenden Modellideen zu verlieren.

Als zweite, weniger rigorose Bedingung erwies sich die Frage der Überschaubarkeit. Das Modell sollte einerseits komplex genug sein, um realistische Testbedingungen zu garantieren. Andererseits sollten aber keine unnötigen zusätzlichen Details Bestandteil des zu schnürenden Programmpakets werden, die in der beabsichtigten Grundlagenforschung den Blick auf das Wesentliche verstellen würden.

Im Folgenden werden exemplarisch einige neuere Phosphor und Algendynamikmodelle (ein Modell von Omlin et al. zum Züricher See, Biola, LEEDS, PCLa-

ke, EcoLE und PROTECH) kurz charakterisiert. Dies geschieht mit Hauptaugenmerk auf ihre Eignung für die vorliegende Fragestellung.

Modell von Omlin

Das in [ORF01] von Omlin et al. vorgestellte Modell zur Modellierung der Algendynamik im Züricher See ist aus technischen Gründen für die gegebene Fragestellung nicht geeignet. Sie verwenden in ihrer Modellierung des Phosphors weder einen geschlossenen Massezyklus, noch setzen sie in ihren Masseflüssen gleichen Abbau wie Produktion durch eine gegebene Transformationsfunktion fest. Des Weiteren hätte es den zusätzlichen programmiertechnischen Aufwand bedeutet, Wasserschichtung zu berücksichtigen, welche zur besseren Überschaubarkeit des Codes nicht erwünscht war.

Aus ökologischer Sicht wäre aber eine Verwendung durchaus interessant gewesen, da an einigen Stellen eine realistischere Modellierung bekannter ökologischer Phänomene erfolgt. Das sei präzisiert. Speziell werden Prozesse hervorgehoben, die im verwendeten Modell nach Hongping und Jianyi in dieser Art nicht berücksichtigt werden.

Das Sediment, welches gerade für flache Seen eine außerordentliche Bedeutung inne hat, wird in zwei Schichten aufgeteilt und ermöglicht es somit, ein langfristiges Binden des Phosphors, und damit ein dauerhaftes Entziehen aus dem Nährstoffkreislauf des Sees, abzubilden.

Ebenso wird die Bindung des gelösten Phosphors an im Wasser vorhandene Schwebepartikel modelliert. Als Nährstoffe werden nicht nur der Phosphor betrachtet, sondern auch der für das Phänomen des „Umkippens eines Sees“ wichtige Sauerstoff sowie Ammonium und Stickstoff. Im Modell werden veränderliche Phosphoranteile in der Gesamtmasse der Algen, aber nicht des Zooplanktons, zugelassen. Weiterhin berücksichtigt das Modell für den Züricher See keine höheren trophischen Ebenen.

Biola

Das Modell Biola von C. Pers (siehe zur Modellbeschreibung und Anwendung in [Per02, Per05, Per06] sowie für einen Vergleich mehrerer Modelle unter anderem Biola und LEEDS in [DP04]) ist auf Grund der restriktiven Anforderungen an die zur Formulierung des Modells gewählte Konservativität (Definition 2.5) nicht geeignet. Dieses Modell beschreibt ebenso wie das Modell von Omlin Algendyna-

miken mit Hilfe der Nährstoffe Phosphor, Sauerstoff, Ammonium und Stickstoff. Es nimmt weiterhin höhere trophische Ebenen in die Betrachtung, pflanzen- und fischfressende Fische sowie Macrophyten. Allerdings werden in diesem Modell, wie in vielen anderen Modellen auch (siehe dazu die Absätze zu LEEDS, PROTECH oder EcoLE), konstante Phosphoranteile für alle Organismen angenommen. Insofern erübrigt sich eine explizite Formulierung der Nährstoffe in den Organismen und verhindert dadurch einen geschlossen formulierten Materiekreislauf für Nährstoffe im Sinne der Definition für patankar-konservative Systeme.

LEEDS

Das in [MH04, MBMP06] beschriebene Modell LEEDS modelliert Phosphorkonzentrationen zur Beurteilung der Eutrophierung eines Sees. Ein Vergleich für den schwedischen See Vänern für insgesamt vier unterschiedliche Modelle (von den hier aufgeführten LEEDS und Biola) findet sich in [DP04]. Allerdings werden Organismen nicht explizit in Arten aufgeteilt dargestellt. Das mag sich durch ausgewählte Ergebnisse oder Fragestellungen rechtfertigen lassen, wobei die Autoren selbst einräumen: „To account for just two biological compartments is evidently an over-simplification from biological point of view,...“. Darüber hinaus gibt es keine geschlossene Angabe der verwendeten Differentialgleichungen. Damit sind die Gründe benannt, die zum Ausschluss dieses Modells geführt haben.

PCLake

Ein anderes weit verbreitetes Modell ist PCLake bzw. sein Vorgänger PCLoos (beschrieben und angewendet unter anderem in [JA90, Jan05] sowie berücksichtigt in einem sehr umfassenden Vergleich von Modellen stark unterschiedlicher Komplexität in [BH07]). Es berücksichtigt weit über 50 Differentialgleichungen. Nährstoffkonzentrationen werden nicht in festen Verhältnissen zueinander modelliert. Dieses Modell ist auf Grund seiner Komplexität nicht als Testfall geeignet, da sich hier der Blick auf die wesentlichen Entwicklungsziele erschweren würde. Grundsätzlich ist eine Anwendung der in dieser Arbeit entwickelten und vorgestellten Verfahren aber möglich; dies würde lediglich einen enormen programmiertechnischen Aufwand bedeuten, der für diese Dissertation außerhalb der Zielsetzung liegt.

EcoLE

Das Modell EcoLE, beschrieben in [ZCB08], ist eine Erweiterung eines Modells beschrieben in [Boe99] für den See Erie. Auch EcoLE berücksichtigt eine Vielzahl von Nährstoffen als Zustandsgrößen: Stickstoff, Phosphor und Kohlenstoff. Allerdings verwendet auch dieses Modell konstante Phosphorkonzentrationen in den Biomassen. Somit ergeben sich ähnliche Gegenargumente wie zum Biola Modell.

PROTECH

Als letztes Modell wurde noch PROTECH (zu finden in [RIE01]) in Betracht gezogen. Hier wird ebenso mit konstanten Phosphorkonzentrationen in den Biomassen gearbeitet. Zusätzlich ist das Modell nicht in der Form eines Differentialgleichungssystems beschrieben. Stattdessen wird die Lösung einer linearen Differentialgleichung mit konstanten Koeffizienten als Lösungsalgorithmus verwendet, wobei die Koeffizienten aber nicht konstant gewählt sind. Durch diese Art der Formulierung ist eine Behandlung im Rahmen der modifizierten Patankar Verfahren nicht umsetzbar.

3.2.2 Das verwendete Modell - eine Erweiterung zum West Lake Modell von Hongping und Jianyi

Hongping und Jianyi modellieren in [HJ02] die Algendynamik des West Lakes bei Hangzhou nahe Shanghai mit dem limitierenden Nährstoff Phosphor. Sie verwenden das Modell zum Vergleich unterschiedlicher restorativer Maßnahmen zur Reduktion der zu erwartenden Algenblüten.

Zentral für die Auswahl dieses Modells waren drei Kriterien. Es werden keine konstanten Nährstoffverhältnisse in den Biomassen angenommen. Dies führt im ursprünglichen Modell zu einer geschlossenen Nährstoffkreislaufmodellierung und somit einer Gleichungsform, die der Konservativität nach Definition 2.12 genügt.

Das Modell ist von moderater Komplexität. Es wird gebildet (in seiner ursprünglichen Formulierung) über 13 gewöhnliche Differentialgleichungen. Damit liegt es weit über einfachen akademischen Problemen und ist trotzdem mit für diese Arbeit angemessenem programmiertechnischem Aufwand zu bewältigen.

Da kein ökologisches Problem den Untersuchungen zu Grunde lag, gab es auch keine Messdaten. Um eine Bewertung der berechneten Ergebnisse anstellen zu können, war daher eine Referenzimplementierung wünschenswert. Diese existiert

für Ecobas. Zusätzlich liefert Ecobas seine Ergebnisdaten in einer Form, die sich für einen Ergebnisvergleich anbietet.

Nimmt man ungestörtes Algenwachstum ($growth_i$) unter Idealbedingungen an, so haben die Algengruppen Verdopplungszeiten zwischen fünf und sieben Stunden. Dem gegenüber stehen das Wachstum ($assim$) des Zooplanktons mit einer Verdopplungszeit von knapp drei Tagen und 15 Stunden und Mineralisierung, deren langsamster ($minpd$) eine Halbierungszeit von knapp 70 Tagen aufweist. Die Details zu den Termen und den Konstanten befinden sich im Anhang A. Damit ergeben sich Zeitskalendifferenzen in der Größenordnung von 10^2. Das System ist somit nicht besonders steif. Dies widerspricht aber einer Verwendung der Erweiterung des West Lake Modells aus [HJ02] als Testproblem nicht, da das Ziel dieser Arbeit das erstmalige numerische lösen ökologischer Flachseenmodelle im Rahmen einer Adaption des TAU 2D Codes war und die Eignung der Erweiterungen der modifizierten Patankar Verfahren für steife Probleme anhand gewöhnlicher Differentialgleichungssysteme (Beispiel 1.3) gezeigt wird.

Eine wesentliche Eigenschaft des Sees ist die durchschnittliche Tiefe von nur 1.5m, was es ermöglicht, den See als vertikal durchmischt anzunehmen. Dies erlaubt eine sinnvolle Anwendung der swe (vergleiche dazu die vereinfachenden Annahmen aus Unterkapitel 3.1). Auch sind Hongping und Jianyi in der Lage, die relevanten Algengruppenarten auf vier einzuschränken und diese in ihrem Modell gesondert zu betrachten. Weiterhin verwenden sie keine konstanten Phosphoranteile in den Organismen und berücksichtigen in einfacher Form das Sediment. Das Modell ist als gewöhnliche Differentialgleichung formuliert.

Das Modell benötigt als Vorgaben die Temperatur des Wassers ($[T] = °\mathrm{C}$), des Sediments ($[TE] = °\mathrm{C}$) sowie die Lichtintensität an der Wasseroberfläche ($[I] =$Lux).

Das Modell beschreibt die Dynamiken der vier Algengruppenarten (Cyanophyta, Chlorophyta, Cryptophyta und Bacillariophyta) mit ihren Biomassen BA_i, $i = 1, \ldots, 4$ und deren entsprechenden Phosphormassen $PA_i, i = 1, \ldots, 4$. Darüber hinaus wird Zooplankton mit Biomasse BZ und der korrespondierenden Phosphormasse PZ berücksichtigt. Als Nährstoff wird Phosphor modelliert. In der verwendeten Erweiterung wird der Phosphor im Wasserkörper und im Sediment jeweils sowohl in gelöster, PS respektive PE_I, als auch in organisch gebundener, PD bzw. PE_O, Form betrachtet.

Die Differenzierung des Phosphors nach organischer und anorganischer Fraktion im Sediment ist eine Neuerung und im ursprünglichen Modell nicht enthalten.

Im ursprünglichen Modell wurden an dieser Stelle konstante Verteilungen von partikulärem und gelöstem Phosphor in den Größen PD und der einen Sedimentgröße „$PE_{\text{ursprüngliches Modell}}$“ angenommen.

Diese Aufsplittung in partikulären und gelösten Phosphor ermöglicht es, temperaturverzögerte Phosphorrückführung abzubilden. Der Phosphor im Sediment steht also nicht sofort wieder im diffusiven Ausgleich mit dem gelösten Phosphor des Wasserkörpers. Allerdings ist es durch diese Art der Modellierung nicht möglich, ein langfristiges Binden (über Jahre hinweg) des Phosphors an die Bodenmatrix zu reproduzieren, ein Phänomen, welchem im Bereich der flachen Seen eine wesentliche Rolle für die Nährstoffverteilung zukommt. Dies leisten beispielsweise die Modelle PCLake oder das Modell von Omlin et al.

Weitere vorgenommene Veränderungen betreffen die konkreten Formulierungen der Terme in (3.4) und werden im Anhang A beschrieben. Die Erweiterung des Modells um weitere Nährstoffe oder Prozesse, wie z.B. im Modell von Omlin oder PCLake, ist möglich, sollte ein gegebenes ökologisches Problem dies nahe legen.

Insgesamt ergibt sich das Gleichungssystem für $\mathbf{u}_p = (BA_A, BA_B, BA_C, BA_D, PA_A, PA_B, PA_C, PA_D, BZ, PZ, PS, PD, PE_I, PE_O)^T$

$$\mathbf{u}_p' = \begin{pmatrix} \partial_t BA_i \\ \partial_t PA_i \\ \partial_t BZ \\ \partial_t PZ \\ \partial_t PS \\ \partial_t PD \\ \partial_t PE_I \\ \partial_t PE_O \end{pmatrix} = \begin{pmatrix} growth_i - res_i - sink_i - graz_i \\ uptba_i - resp_i - setpa_i - gsinkp_i - assimp_i \\ \sum_{j=1}^4 assim_j - zres - zmor \\ \sum_{j=1}^4 assimp_j - zresp - zmorp \\ exchp_{ps} - exchp_{pe_i} - \sum_{j=1}^4 uptba_j + minpd + zresp + \sum_{j=1}^4 resp_j \\ -minpd - setpd + zmorp \\ -exchp_{ps} + exchp_{pe_i} + minpe \\ -minpe + setpd + \sum_{j=1}^4 (setpa_j + gsinkp_j). \end{pmatrix} =: \mathbf{r}_p. \tag{3.4}$$

Die zeitlichen Änderungen der Biomassen $(BA_i, BZ, i = 1, \ldots, 4)$ werden bestimmt durch Wachstum bzw. Aufnahme ($growth_i$ respektive $assim_j$), Stoffwechselprozesse zum Erhalt der Organismen ($res_i, zres$), Sedimentation ($sink_i, zmor$) und im Fall der Algen auch Grazing durch das Zooplankton ($graz_i$).

Die Prozesse für die assoziierten Phosphormassen erklären sich analog, wobei die Terme $gsinkp_i$ und $assimp_i$ für die Phosphormassen der Algen PA_i in $graz_i$ für BA_i zusammenfallen.

Die Dynamiken des nicht Organismen zugeordneten Phosphors im Wasser und Sediment (PS, PD, PE_I, PE_O) werden bestimmt über jeweils einen Prozess der Mineralisierung ($minpd, minpe$) und den diffusiven Austausch zwischen den gelösten Phosphorfraktionen ($exchp_\gamma, \gamma \in \{pe_i, ps\}$). Des Weiteren werden die entsprechenden Terme der Organismendynamiken ($uptba_i$, $resp_i$, $setpa_i$, $gsinkp_i$, $zresp$, $zmorp$) berücksichtigt.

Eine explizite Aufschlüsselung aller Terme findet sich im Anhang A. Die Zuordnung der verschiedenen Funktionen der rechten Seite werden im folgenden Schema (Abbildung 3.1) noch verdeutlicht. Zur Vereinfachung wurden in der Graphik nur eine Algengruppenart mit der entsprechenden Bio- und Phosphormasse berücksichtigt.

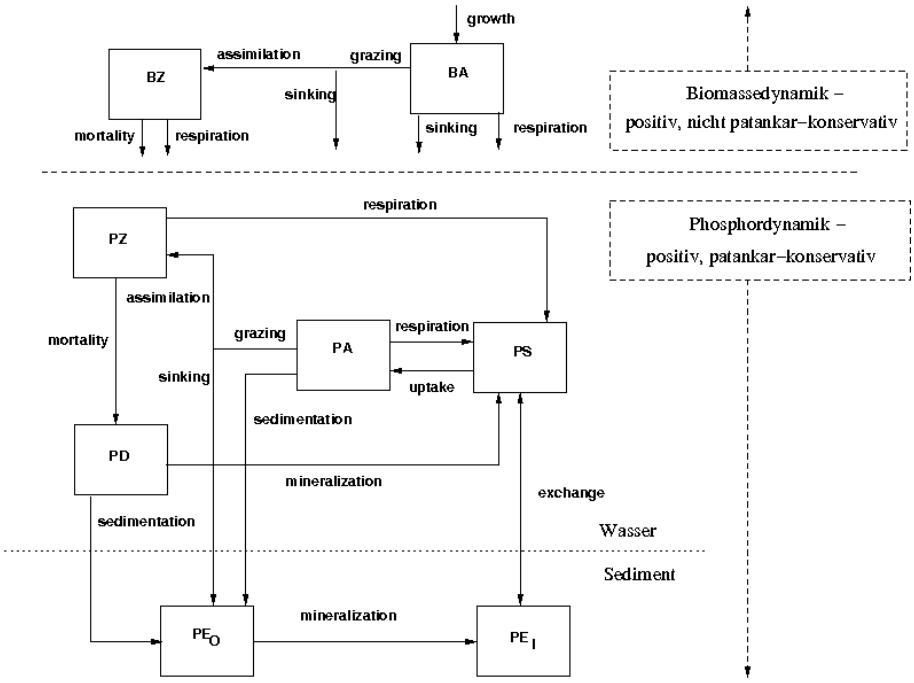

Abbildung 3.1: Berücksichtigte biologische und chemische Prozesse

Man erkennt sofort einen geschlossenen Materiekreislauf für den Phosphor im Wasser und Sediment. Dieser Kreislauf ist patankar-konservativ nach Definition 2.5 und ermöglicht somit eine Anwendung des bereits vorgestellten modifizierten Patankar Euler Verfahrens und seiner Erweiterungen. Die Biomassen sind nicht geschlossen modelliert und erfordern daher eine gezielte Berücksichtigung konservativer und nicht konservativer Wechselwirkungen unter der generellen Forderung

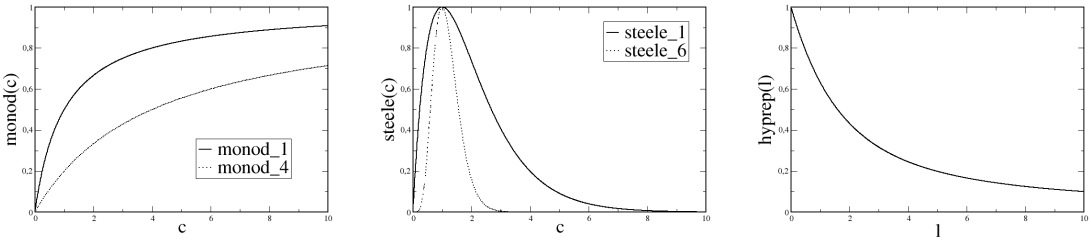

Abbildung 3.2: Typische Kontrollfunktionen: *monod*, *steele* und *hyprep*

der Positivität.

3.2.3 Mechanik der Modellierung

An dieser Stelle soll nur auf einige für ökologische Modellierung typische Ausdrücke eingegangen werden. Für eine allgemeine und umfassende Darstellung siehe [RM08] oder auch [SG85].

Für Sättigungsprozesse in Abhängigkeit einer Größe c_i ist ein häufig verwendeter nach Monod[1] benannter Modellierungsterm gegeben durch

$$\text{monod}_a(c_i) = \frac{c_i}{a + c_i} \text{ für } a, c_i > 0. \tag{3.5}$$

Es gilt offensichtlich $0 < \text{monod}_a(c_i) < 1$ bzw. $\text{monod}_a(c_1) < \text{monod}_a(c_2) \Leftrightarrow c_1 < c_2$ und $\text{monod}_a(a) = \frac{1}{2}$.

Für Prozesse mit einem festgelegten Maximum max für eine Größe c_i bietet sich eine Formulierung nach Steele [Ste62]

$$\text{steele}_a(c) = \left(ce^{1-c}\right)^a \text{ für } max, c_i, a > 0 \tag{3.6}$$

mit $c = \frac{c_i}{max}$ an. Auch hier gilt $0 < \text{steele}(c) \leq 1$. Da

$$\text{steele}'(c) = (1 - c)\,\frac{(ce^{1-c})^a a}{c}$$

nur eine Nullstelle hat, $\text{steele}'(1) = 0$, gibt es nur ein Maximum von $\text{steele}(1) = 1$. Beides sind also Funktionen, die als Steuerungselemente dämpfend wirken. Der Parameter a bei monod respektive steele steuert, wie schnell nahezu ungedämpfte Aufnahme oder Umsetzung bzw. wie schnell die Dämpfung um das Maximum

[1] Jacques Lucien Monod, franz. Biochemiker, Nobelpreisträger, 1910-1976

herum stattfindet. Die Funktion steele erzeugt eine nicht symmetrische Steigerung vor bzw. Abnahme der Prozessstärke nach dem Maximum $c_i = max$. Eine Funktion zur Darstellung hyperbelartiger Reduktion für eine Größe l liefert

$$\text{hyprep}(l) = \frac{1 - e^{-l}}{l} \text{ für } l > 0. \tag{3.7}$$

Nach l'Hospital gilt $\text{hyprep}(l) \stackrel{l \to 0}{=} 1$ und hyprep ist streng monoton fallend.

Die Abbildung 3.2 zeigt alle beschriebenen Kurventypen. Ihre Anwendungen befinden sich in den detaillierten Prozessbeschreibungen des Phosphormodells (3.4) im Anhang A.

3.3 Das mathematische Gesamtsystem

Führt man zusammen, was in Unterkapitel 3.1.1 und 3.2.2 beschrieben wurde und erweitert sämtliche im Wasser modellierten Komponenten des Systems (3.4) durch advektiven Transport und die gelösten Phosphorfraktionen des Wassers und des Sediments um diffusiven Transport, so ergibt sich das vollständige erweiterte West Lake Modell. Es wird nur für die gelösten Fraktionen des Phosphors Diffusion betrachtet, da der Effekt für partikuläre Stoffe als deutlich geringer angenommen wird. Die Struktur ist ähnlich zur swe (3.1) erweitert um den diffusiven Transport,

$$\partial_t \mathbf{u} + \partial_{x_1} \left(\mathbf{f}_1^c(\mathbf{u}) - \mathbf{f}_1^d(\mathbf{u}) \right) + \partial_{x_2} \left(\mathbf{f}_2^c(\mathbf{u}) - \mathbf{f}_2^d(\mathbf{u}) \right) = \mathbf{r}(\mathbf{u}). \tag{3.8}$$

Die rechte Seite $\mathbf{r}(\mathbf{u})$ ist -nach entsprechender Sortierung- eine, ohne weitere Modifikationen durchgeführte Zusammensetzung der rechten Seiten

$$\mathbf{r}(\mathbf{u}) = \begin{pmatrix} \mathbf{r}_{swe}(\mathbf{u}_{swe}) \\ \mathbf{r}_p(\mathbf{u}_p) \end{pmatrix}.$$

Vergleiche dazu (3.2) und (3.4). Allerdings sind die Komponenten der Gleichung (3.8), explizit dargestellt

$$\mathbf{u} = \begin{pmatrix} \Phi \\ \Phi v_1 \\ \Phi v_2 \\ PS \\ PD \\ PE_I \\ PE_O \\ PA_1 \\ PA_2 \\ PA_3 \\ PA_4 \\ PZ \\ BZ \\ BA_1 \\ BA_2 \\ BA_3 \\ BA_4 \end{pmatrix}, \mathbf{f}_j^c(\mathbf{u}) = \begin{pmatrix} \Phi v_j \\ \Phi v_1 v_j + \frac{\delta_{1j}}{2}\Phi^2 \\ \Phi v_2 v_j + \frac{\delta_{2j}}{2}\Phi^2 \\ v_j PS \\ v_j PD \\ 0 \\ 0 \\ v_j PA_1 \\ v_j PA_2 \\ v_j PA_3 \\ v_j PA_4 \\ v_j PZ \\ v_j BZ \\ v_j BA_1 \\ v_j BA_2 \\ v_j BA_3 \\ v_j BA_4 \end{pmatrix} \text{ und } \mathbf{f}_j^d(\mathbf{u}) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \text{Diff}_{PS}\partial_{x_j} PS \\ 0 \\ \text{Diff}_{PE_I}\partial_{x_j} PE_I \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{3.9}$$

mit zwei Konstanten, $\text{Diff}_{PS} = 8.64 \cdot 10^{-5} \frac{\text{m}^2}{\text{d}}$ und $\text{Diff}_{PE_I} = 6.83 \cdot 10^{-5} \frac{\text{m}^2}{\text{d}}$, deutlich umfangreicher als in Form (3.1).

3.4 Rand- und Anfangsbedingungen

Im Allgemeinen wird der Rand des Rechengebiets in die drei Klassen, feste Wand, Ein- und Ausströmrand, eingeteilt. Dabei ist $\mathbf{n}$ auf dem Rand immer ein nach außen gerichteter Einheitsnormalenvektor. An der festen Wand wird

$$\mathbf{vn} = 0$$

mit $\mathbf{v} = (v_1, v_2)^T$ gesetzt. Für den Ausströmrand werden Neumann Randbedingungen durch

$$\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = \mathbf{0} \tag{3.10}$$

festgelegt. Am Einströmrand schreibt man mit Dirichlet Randbedingungen jeder Komponente einen festen Wert

$$\mathbf{u} = \varphi,\ \varphi \in \mathbb{R}^{17} \tag{3.11}$$

vor. Eine Anfangsverteilung wird mittels einer Funktion

$$\mathbf{u}_0(x_1, x_2) \tag{3.12}$$

festgelegt.

Die Wahl eines Szenarios geschieht also über die Festlegung des Rechengebiets, des Eingangszustands φ und der Anfangsverteilung $\mathbf{u}_0$.

Um die Eignung des Gesamtverfahrens zu demonstrieren, wird neben speziellen Tests für die einzelnen Teile des Lösers auch ein Gesamtszenario präsentiert.

Dafür wird ein See, siehe Abbildung 3.3, über einen Jahreszyklus simuliert. Der See ist mit 1889 Boxen diskretisiert und hat eine physikalische Kantenlänge von etwa 10km x 5km. Das linke Ende des unter dem See befindlichen Kanals ist der einzige Einströmrand, das rechte Ende des Kanals der einzige Ausströmrand. Alle anderen Ränder sind feste Wand.

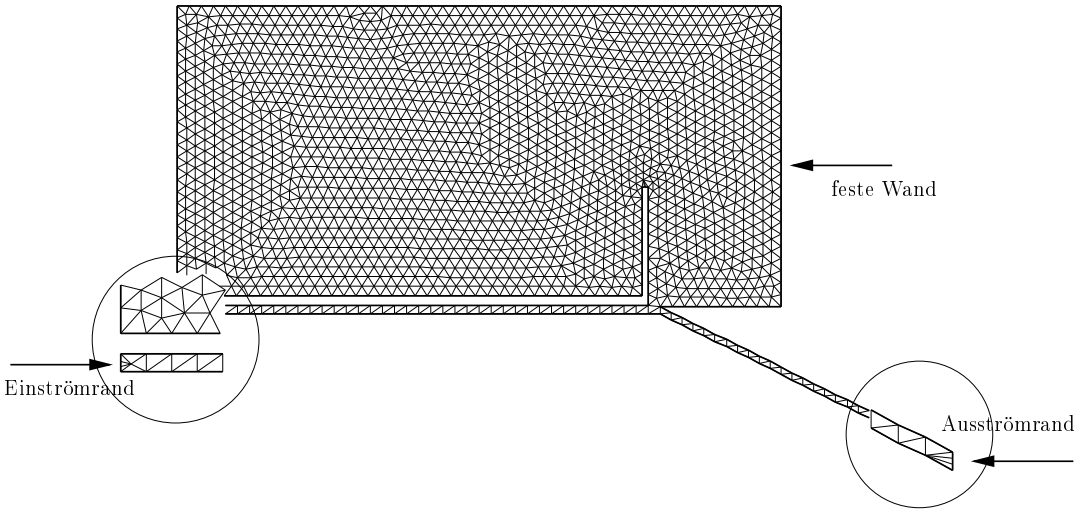

Abbildung 3.3: Das Netz des Sees

Es wird der Effekt eines partikelfreien Zustroms betrachtet. Daher ist

$$\varphi = \begin{pmatrix} 1.56\ \mathrm{m} \\ 200\ \frac{\mathrm{m}}{\mathrm{d}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.13}$$

Als Startwerte werden die Ergebnisse aus der Ecobas Rechnung von Unterkapitel 6.1 für die Zeit $t = 180d$, d.h. dem ersten Sommer der vorgenannten Rechnung, verwendet:

$$\mathbf{u}_0 = \begin{pmatrix} H \\ v_x \\ v_y \\ BA_A \\ BA_B \\ BA_C \\ BA_D \\ PA_A \\ PA_B \\ PA_C \\ PA_D \\ BZ \\ PZ \\ PS \\ PD \\ PE_I \\ PE_O \end{pmatrix}_0 = \begin{pmatrix} 1.56\ \mathrm{m} \\ 0.\ \frac{\mathrm{m}}{\mathrm{d}} \\ 0.\ \frac{\mathrm{m}}{\mathrm{d}} \\ 1.478178\,10^{+00}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 8.506500\,10^{-01}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 9.482997\,10^{-01}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 1.298752\,10^{+01}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 2.008258\,10^{-02}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 1.489144\,10^{-02}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 1.200959\,10^{-02}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 1.714340\,10^{-01}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 5.427954\,10^{+00}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 7.181461\,10^{-02}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 1.642674\,10^{-03}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 4.067037\,10^{-03}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 3.585960\,10^{-01}\ \frac{\mathrm{g}}{\mathrm{m}^3} \\ 4.446198\,10^{-02}\ \frac{\mathrm{g}}{\mathrm{m}^3} \end{pmatrix}.$$

Eine Darstellung der Ergebnisse der mit dem Solver aus Kapitel 4 berechneten Näherungen befindet sich in Unterkapitel 6.3.

Kapitel 4

Numerische Verfahren für Erhaltungsgleichungen

Für hyperbolisch parabolische Erhaltungsgleichungen,

$$\partial_t \mathbf{u} + \sum_{i=1,2} \partial_{x_i} \left(\mathbf{f}_i^c - \mathbf{f}_i^d \right) (\mathbf{u}) = \mathbf{r}(\mathbf{u}) \tag{4.1}$$

mit konvektivem Anteil $\mathbf{f}_i^c$, diffusivem Anteil $\mathbf{f}_i^d$ und Quellen bzw. Senken $\mathbf{r}(\mathbf{u})$, sind Finite Volumen Verfahren (kurz *FVV*) die nahe liegende Verfahrensklasse, da man über die integrale Formulierung sowohl die für Erhaltungsgleichungen gewünschte Konservativität direkt erhält, als auch die Verfahren beim Auftreten unstetiger Lösungen bis zu einem gewissen Grad numerisch robust sind. Da unstetige Lösungen in natürlicher Weise mögliche Lösungen nicht linearer Gleichungssysteme sind, ist letzteres vor allem in der Anwendung von großer Bedeutung.

Die folgenden Unterkapitel dienen der Beschreibung des verwendeten Strömungslösers. Der vorliegende Löser ist eine Adaption des in [Mei96, MS98, Mei94] beschriebenen Verfahrens. Es soll hier nur so wenig wie zur vollständigen Beschreibung des verwendeten Algorithmus nötig behandelt werden.

Für eine Einführung in das Gebiet der FVV für hyperbolische Erhaltungsgleichungen wird auf das exzellente Buch von LeVeque [Lev02] verwiesen. Als weiterführende Texte mit numerischem Schwerpunkt sei hier unter vielen noch auf [Lev06] vom selben Autor oder die umfassende Behandlung durch Morton und Sonar in [MS07] sowie auf das Buch [GR96] von Godlewski und Raviart hingewiesen. Eine theoretische Auseinandersetzung mit hyperbolischen Problemen findet sich beispielsweise in [Smo94] von Smoller.

4.1 Die integrale Form

Zur Verwendung eines FVV ist es nötig, die Differentialgleichung (4.1) im Raum zu integrieren.

Integration über ein Gebiet $\Omega \subset \mathbb{R}^2$ mit dem Rand $\partial\Omega$ und die Anwendung des Satzes von Gauß liefert

$$\partial_t \int_\Omega \mathbf{u}\, d\mathbf{x} = \int_{\partial\Omega} \left[\sum_{i=1,2} \left(\mathbf{f}_i^d - \mathbf{f}_i^c \right)(\mathbf{u}) n_i \right] d\mathbf{s} + \int_\Omega \mathbf{r}(\mathbf{u}) d\mathbf{x}.$$

Betrachtet man nun die Mittelwerte über dem Gebiet Ω, also

$$\mathbf{u}_\Omega(t) := \frac{1}{|\Omega|} \int_\Omega \mathbf{u}(\mathbf{x}, t)\, d\mathbf{x},$$

mit der Fläche $|\Omega|$ erhält man

$$\frac{d}{dt}\mathbf{u}_\Omega(t) = \frac{1}{|\Omega|} \int_{\partial\Omega} \left[\sum_{i=1,2} \left(\mathbf{f}_i^d - \mathbf{f}_i^c \right)(\mathbf{u}) n_i \right] d\mathbf{s} + \frac{1}{|\Omega|} \int_\Omega \mathbf{r}(\mathbf{u}) d\mathbf{x}.$$

Definiert man dann noch

$$\begin{aligned} \mathcal{L}^c &= -\frac{1}{|\Omega|} \int_{\partial\Omega} \sum_{i=1,2} \mathbf{f}_i^c(\mathbf{u}) n_i d\mathbf{s}, \\ \mathcal{L}^d &= \frac{1}{|\Omega|} \int_{\partial\Omega} \sum_{i=1,2} \mathbf{f}_i^d(\mathbf{u}) n_i d\mathbf{s} \text{ und} \\ \mathcal{R} &= \frac{1}{|\Omega|} \int_\Omega \mathbf{r}(\mathbf{u}) d\mathbf{x}, \end{aligned}$$

so erhält man zusammenfassend

$$\frac{d}{dt}\mathbf{u}_\Omega(t) = \mathcal{L}^c + \mathcal{L}^d + \mathcal{R}. \tag{4.2}$$

Legt man

$$\mathbf{u}(x, y, t) = \mathbf{u}_\Omega(t) \tag{4.3}$$

für $(x, y) \in \Omega \setminus \partial\Omega$ fest, d.h. es werden ab jetzt Ω-weit nur noch Raummittelwerte betrachtet, vereinfachen sich die Ausdrücke für $(x, y) \in \Omega \setminus \partial\Omega$ weiter zu

$$\mathcal{R} = \mathbf{r}(\mathbf{u}_\Omega).$$

Einsetzen liefert

$$d_t \mathbf{u}_\Omega = \mathcal{L}^c + \mathcal{L}^d + \mathbf{r}(\mathbf{u}_\Omega). \tag{4.4}$$

Für die Diskretisierung der Gleichung (4.4) wird ein Splitting Ansatz (siehe [Lev06]) verwendet. Man zerlegt dafür die rechte Seite von (4.4) in zwei Teile

$$d_t \mathbf{u}_\Omega = \mathcal{L}^c + \mathcal{L}^d \text{ und} \tag{4.5}$$

$$d_t \mathbf{u}_\Omega = \mathbf{r}(\mathbf{u}_\Omega). \tag{4.6}$$

Nun werden beide Gleichungen nach dem Schema

$$\mathbf{u}_\Omega^{n+\frac{1}{3}} = \mathbf{u}_\Omega^n + \frac{h}{2}\,\phi_1\left(\mathbf{u}_\Omega^n, h, \mathbf{r}\right) \tag{4.7}$$

$$\mathbf{u}_\Omega^{n+\frac{2}{3}} = \mathbf{u}_\Omega^{n+\frac{1}{3}} + h\,\phi_2\left(\mathbf{u}_\Omega^{n+\frac{1}{3}}, h, \mathcal{L}^c + \mathcal{L}^d\right) \tag{4.8}$$

$$\mathbf{u}_\Omega^{n+1} = \mathbf{u}_\Omega^{n+\frac{2}{3}} + \frac{h}{2}\,\phi_1\left(\mathbf{u}_\Omega^{n+\frac{2}{3}}, h, \mathbf{r}\right)$$

angenähert. Dies führt nach [Lev06] zu einem Verfahren zweiter Ordnung, falls die beiden untergeordneten Verfahren ϕ_1 und ϕ_2 zur Näherung der Gleichungen (4.5) respektive (4.6) jeweils auch zweiter Ordnung sind.

Unterkapitel 4.2 beschäftigt sich mit der Diskretisierung von $\mathcal{L}^c$, Unterkapitel 4.3 mit der zu $\mathcal{L}^d$ und für die Diskretisierung der Quellen und Senken, $\mathcal{R}$, wurden bereits eine Vielzahl von problemangepassten Verfahren in Kapitel 2 vorgestellt.

Die Netzstruktur

Sei das Gebiet $\Omega \subset \mathbb{R}^2$ von Interesse. Es habe den polygonalen Rand $\partial\Omega$. Der Rand sei gegliedert in drei Bereiche entsprechend den geltenden Randbedingungen, $\partial_e\Omega$ der Einströmrand, $\partial_a\Omega$ der Ausströmrand und $\partial_f\Omega$ die feste Wand. Es gelte $\partial\Omega = \partial_e\Omega \cup \partial_a\Omega \cup \partial_f\Omega$ und die Mengen $\partial_e\Omega \cap \partial_a\Omega$, $\partial_e\Omega \cap \partial_f\Omega$ und $\partial_f\Omega \cap \partial_a\Omega$ sind vom Lebesgue Maß Null bei Linienintegralen.

Das Gebiet Ω wird in eine endliche Anzahl von Dreiecken zerlegt. Seien $\triangle_i \subset \Omega, i = 1, \ldots, N_{tri}$ die Dreiecke und $\mathcal{T} = \{\triangle_i | i = 1, \ldots, N_{tri}\}$ die Menge aller Dreiecke der Zerlegung. Die Aufteilung erfolgt ohne hängende Knoten und überschneidungsfrei in dem Sinne von

$$(\triangle_i \setminus \partial\triangle_i) \cap (\triangle_j \setminus \partial\triangle_j) = \emptyset$$

für $i \neq j$ und es gilt

$$\cup_{i=1}^{N_{tri}} \triangle_i = \Omega.$$

Zur Berechnung dienen aber im Wesentlichen nicht diese Dreiecke, sondern mit ihrer Hilfe definierte polygonal berandete Boxen Ω_i. Man wählt aus dem Inneren

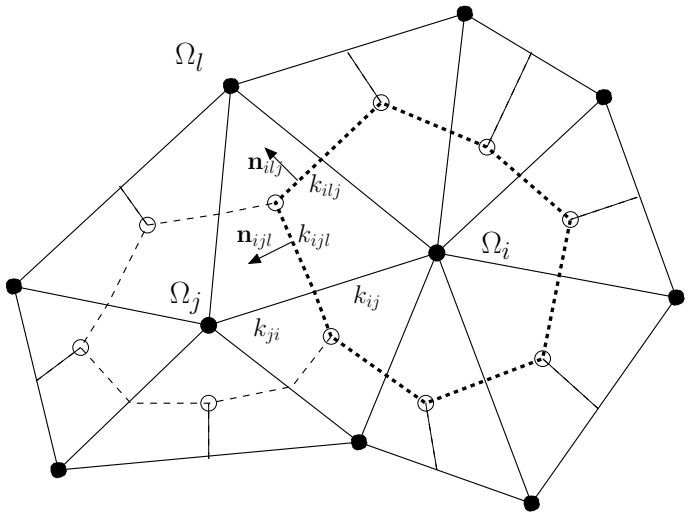

Abbildung 4.1: Ein Netzausschnitt - unstrukturiertes Netz mit Boxen

jedes Dreiecks einen Punkt und verbindet diesen durch ein Geradenstück jeweils mit den Seitenmittelpunkten der Dreiecke. Diese Geradenstücke bilden eine polygonale Berandung um die Ecken der Dreiecke und definieren in dieser Art die Boxen Ω_i. Es sei die Anzahl der Boxen mit N_{Box} bezeichnet. Siehe dazu Abbildung 4.1, wobei die Dreiecke durch durchgezogene und die Boxen durch unterbrochene Linien definiert sind. Die Dreiecksecken, die ausgefüllten Punkte, korrespondieren dabei mit den Boxen. So bezeichnet $\dot{\Omega}_i$ die Koordinaten der Dreiecksecke, die in Box Ω_i liegt, wohin gegen (Ω_i) den Schwerpunkt der Box Ω_i bezeichnet.

Bezeichne weiterhin $N(i) = \{j \in \{1, \ldots, N_{Box}\} \setminus \{i\} | \Omega_j \cap \Omega_i \neq \emptyset\}$ die Menge aller echten Nachbarn zur Box Ω_i.

Man erkennt sofort, dass es in dem Dreieck $\triangle_{ijl} = \triangle(\dot{\Omega}_i, \dot{\Omega}_j, \dot{\Omega}_l)$ genau drei Kanten von Boxen gibt, wobei j und l eingeschränkt sind auf $j \in N(i)$ und $l \in N(i) \cap N(j)$. Die Kanten in Dreieck $\triangle_{ijl}$ werden wie folgt bezeichnet. Die Kante in $\triangle_{ijl}$ zwischen Ω_i und Ω_j heißt k_{ijl}, die Kante zwischen Ω_i und Ω_l heißt k_{ilj} usw.

Zu den Kanten k werden normierte Normalenvektoren $\mathbf{n}$ benötigt. Ihre Bezeichnung erfolgt analog. $\mathbf{n}_{ijl}$ ist der Normalenvektor zur Kante k_{ijl} und zeigt von Ω_i nach Ω_j, folglich gilt $\mathbf{n}_{ijl} = -\mathbf{n}_{jil}$.

Es fehlen noch die Bezeichnungen für die Kanten k, die auf dem Rand des Rechengebiets $\partial\Omega$ liegen. Hier ist eine eindeutige Zuordnung über Dreiecke nicht mehr möglich. Diese Kanten liegen nicht innerhalb von Dreiecken sondern auf Dreieckskanten. Bezeichne daher k_{ij} die Kante zwischen $\dot{\Omega}_i$ und dem Mittelpunkt

der Strecke $\overline{\dot{\Omega}_i\dot{\Omega}_j}$. Das heißt $\overline{\dot{\Omega}_i\dot{\Omega}_j} = k_{ij} \cup k_{ji}$. Um den Formalismus im Folgenden zu vereinheitlichen, bestimme die Dreiecks- und Zellkanten auf dem Rand des Rechengebiets Ω in deren Unterteilung

$$k_{ij\zeta_1} = \begin{cases} k_{ij} & , \quad k_{ij} \subset \partial_e\Omega \\ \emptyset & , \quad \text{sonst} \end{cases}, \qquad k_{ij\zeta_2} = \begin{cases} k_{ij} & , \quad k_{ij} \subset \partial_a\Omega \\ \emptyset & , \quad \text{sonst} \end{cases}$$

$$\text{und } k_{ij\zeta_3} = \begin{cases} k_{ij} & , \quad k_{ij} \subset \partial_f\Omega \\ \emptyset & , \quad \text{sonst} \end{cases}.$$

Weiterhin definiere die Menge der erweiterten Nachbarn für alle Boxen i durch $N^+(i) = N(i) \cup \{\zeta_1, \zeta_2, \zeta_3\}$. Hierbei sind $\zeta_1, \zeta_2, \zeta_3$ keine Indizes im klassischen (nummerierten) Sinne. Es sind willkürlich gewählte Zeichen, die lediglich der Identifizierung als Randkante eines bestimmten Typs dienen und eine Summation ermöglichen. Die Details werden später verdeutlicht. Analog verfährt man mit den zugehörigen Normalenvektoren. Für $m = 1, 2, 3$ zeigt dabei $\mathbf{n}_{ij\zeta_m}$ von Ω_i und $\mathbf{n}_{ji\zeta_m}$ von Ω_j ausgehend jeweils aus dem Rechengebiet hinaus.

Zur Vermeidung von Definitionslücken wird noch

$$\mathbf{u}_{\zeta_1} = \mathbf{u}_{\zeta_2} = \mathbf{u}_{\zeta_3} = \mathbf{0}$$

gesetzt.

4.2 Die Diskretisierung der konvektiven Terme

In diesem Unterkapitel soll die numerische Approximation des Integrals

$$\mathcal{L}^c = -\frac{1}{|\Omega_i|} \int_{\partial\Omega_i} \sum_{m=1,2} \mathbf{f}_m^c(\mathbf{u}) n_m d\mathbf{s} \tag{4.9}$$

beschrieben werden. Aus der Struktur des Netzes weiß man, dass

$$\int_{\partial\Omega_i} \sum_{m=1,2} \mathbf{f}_m^c(\mathbf{u}) n_l d\mathbf{s} = \sum_{j\in N(i)} \sum_{l\in(N^+(i)\cap N^+(j))} \int_{k_{ijl}} \sum_{m=1,2} \mathbf{f}_m^c(\mathbf{u}) \left(n_{ijl}\right)_m d\mathbf{s} \tag{4.10}$$

ist.

Nutzt man nun die Rotationsinvarianz (siehe z.B. [Tor01b]) der gewöhnlichen swe bezogen auf die erweiterte swe (3.8) erhält man

$$\int_{k_{ijl}} \sum_{m=1,2} \mathbf{f}_m^c\left(\mathbf{u}\right) \left(n_{ijl}\right)_m d\mathbf{s} = \int_{k_{ijl}} T_{ijl}^{-1} \mathbf{f}_1^c\left(T_{ijl}\mathbf{u}\right) d\mathbf{s},$$

für Matrizen

$$T(\mathbf{n}) = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & n_1 & n_2 & 0 & \cdots & 0 \\ 0 & -n_2 & n_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

mit $T_{ijl} = T(\mathbf{n}_{ijl})$. Das Problem reduziert sich also auf die Lösung der Integrale

$$\int_{k_{ijl}} T_{ijl}^{-1} \mathbf{f}_1^c (T_{ijl}\mathbf{u})\, d\mathbf{s}.$$

Hierfür wird die Mittelpunktsregel aus [SK04], also

$$\int_{k_{ijl}} T_{ijl}^{-1} \mathbf{f}_1^c (T_{ijl}\mathbf{u})\, d\mathbf{s} \approx |k_{ijl}| \left[T_{ijl}^{-1} \mathbf{f}_1^c (T_{ijl}\mathbf{u}_{ijl}) \right] \tag{4.11}$$

verwendet. Statt aber eine Näherung $\mathbf{u}_{ijl}$ zu bestimmen und diese einzusetzen, bestimmt man direkt eine Näherung für den Kantenfluss $\mathbf{f}_1^c (T_{ijl}\mathbf{u}_{ijl})$.

Durch die Wahl der Mittelpunktsregel (die Gauß Quadratur zweiter Ordnung) gewährleistet man exakte Integration für Polynome erster Ordnung. Da die verwendeten Rekonstruktionstechniken maximal Polynome erster Ordnung als Rekonstruktionen liefern, ist eine Integration mit höherer Genauigkeit theoretisch nicht nötig.

4.2.1 Zellkanten im Inneren

Seien nun zur vereinfachten Schreibweise alle noch verbliebenen Indizes, ijl mit $l \neq \zeta_1, \zeta_2, \zeta_3$, als fest betrachtet und damit entbehrlich.

Da $l \neq \zeta_1, \zeta_2, \zeta_3$ ist, liegt k im Inneren des Rechengebiets. Es existieren also Ω_i und Ω_j mit $k \subset \Omega_i \cap \Omega_j$. Man kann entsprechend zwei unterschiedliche Zustandsvektoren $\mathbf{u}_i$ und $\mathbf{u}_j$, jeweils aus den Zuständen der Boxen Ω_i und Ω_j (sowie gegebenenfalls weiteren umliegenden Boxen), konstruieren.

Sei $\dot{\mathbf{k}}_{ijl}$ der Mittelpunkt der Strecke k_{ijl}. Nimmt man also zwei gegebene Zustände $\mathbf{u}_i$ und $\mathbf{u}_j$ für den Mittelpunkt $\dot{\mathbf{k}}$ als gegeben an, ist man noch mit dem Problem konfrontiert, aus diesen einen Zustandsvektor, $\mathbf{u}_{ijl}$, zu konstruieren, mit dessen Hilfe man dann den gesuchten Fluss in (4.11) approximieren kann.

Riemann Löser und Kantenflüsse

Für die Kantenflüsse löst man ein Riemannproblem mit den Zuständen $\mathbf{u}_i$ und $\mathbf{u}_j$. Eine detaillierte Beschreibung zu Riemannproblemen und Riemann Lösern findet sich z.B. in [Tor01a]. In der vorliegenden Umsetzung wird der HLL Riemann Löser (von Harten, Lax und van Leer) verwendet. Für eine theoretische Herleitung siehe ebenfalls in [Tor01a].

Hat man Schätzungen für die minimale, s_l, und die maximale, s_r, charakteristische Geschwindigkeit an $\dot{\mathbf{k}}$, definiert man

$$\mathbf{u}_{ijl} = \begin{cases} T_{ijl}\mathbf{u}_l & , \quad 0 \leq s_l \\ T_{ijl}\mathbf{u}_r & , \quad 0 \geq s_r \\ \frac{1}{s_r - s_l}\Big[s_r T_{ijl}\mathbf{u}_r - s_l T_{ijl}\mathbf{u}_l + \mathbf{f}_1(T_{ijl}\mathbf{u}_l) - \mathbf{f}_1(T_{ijl}\mathbf{u}_r) \Big] & , \quad \text{sonst} \end{cases} . \tag{4.12}$$

Für die Festlegung der minimalen und maximalen Geschwindigkeiten s_l und s_r folgt man der Orientierung der Normalenvektoren. Entsprechend korrelieren $\mathbf{u}_i$ und $\mathbf{u}_j$ mit $\mathbf{u}_l$ und $\mathbf{u}_r$ in dem Verhältnis, wie die Wahl der Ausrichtung fällt. Da die Normalenvektoren immer nach außen zeigen, gilt hier $\mathbf{u}_l = \mathbf{u}_i$ und $\mathbf{u}_r = \mathbf{u}_j$.

Die Gleichung (4.12) lässt sich verwenden, um in (4.11) die Auswertung des Integrals über die Kantenflüsse mittels des HLL Flusses

$$\mathbf{g}^{HLL}\left(\mathbf{u}_l, \mathbf{u}_r, \mathbf{n}_{ijl}\right) \approx \mathbf{f}_1^c\left(T_{ijl}\mathbf{u}_{ijl}\right) \tag{4.13}$$

zu approximieren. Man findet

$$\mathbf{g}^{HLL}\left(\mathbf{u}_l, \mathbf{u}_r, \mathbf{n}\right) \tag{4.14}$$
$$= \begin{cases} \mathbf{f}_1^c(T\left(\mathbf{n}\right)\mathbf{u}_l) & , \quad 0 \leq s_l \\ \mathbf{f}_1^c(T\left(\mathbf{n}\right)\mathbf{u}_r) & , \quad 0 \geq s_r \\ \frac{1}{s_r - s_l}\left(s_r \mathbf{f}_1^c(T\left(\mathbf{n}\right)\mathbf{u}_l) - s_l \mathbf{f}_1^c(T\left(\mathbf{n}\right)\mathbf{u}_r) + s_l s_r T\left(\mathbf{n}\right)\left(\mathbf{u}_r - \mathbf{u}_l\right) \right) & , \quad \text{sonst} \end{cases} .$$

Die charakteristischen Geschwindigkeiten s_l und s_r werden wie folgt geschätzt. Der Einsatz elementarer Linearer Algebra liefert die charakteristischen Geschwindigkeiten, die Eigenwerte der Jakobimatrix $\frac{d\mathbf{f}_1}{d\mathbf{u}}(\mathbf{u})$, als v und $v \pm \sqrt{\phi}$ mit $v = n_1 v_1 + n_2 v_2$. Pragmatisch wird

$$s_l = \min\left(v_l - \sqrt{\phi_l}, v_r - \sqrt{\phi_r}\right) \text{ und } s_r = \max\left(v_l + \sqrt{\phi_l}, v_r + \sqrt{\phi_r}\right)$$

gesetzt. In der Herleitung von (4.14) werden sechs Integrale bestimmt. Die Wahl der Geschwindigkeiten, s_l und s_r, ermöglicht das exakte Lösen von vier dieser

Integrale. Für die verbleibenden zwei Kanten verwendet man die Näherung aus (4.12).

Zur besseren Notation definiere für $l \neq \zeta_1, \zeta_2, \zeta_3$

$$\mathbf{g}_{ijl}^{HLL}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{g}^{HLL}(\mathbf{u}_i, \mathbf{u}_j, \mathbf{n}_{ijl}). \tag{4.15}$$

Rekonstruktion

Für die Bestimmung der Größen $\mathbf{u}_l$ und $\mathbf{u}_r$ werden zwei Möglichkeiten angeboten. Die erste Variante verwendet die jeweiligen Zellmittelwerte, $\mathbf{u}_{\Omega_i}$ und $\mathbf{u}_{\Omega_j}$, siehe dazu (4.3), der beiden Boxen. Dies entspricht

$$\mathbf{u}_i = \mathbf{u}_{\Omega_i} \text{ und } \mathbf{u}_j = \mathbf{u}_{\Omega_j}, \tag{4.16}$$

einer Rekonstruktion der Werte an den Kantenmittelpunkten mit konstanten Polynomen in der jeweiligen Box.

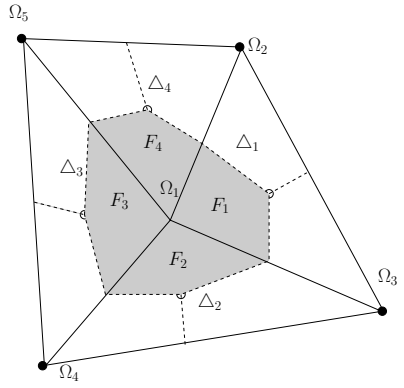

Abbildung 4.2: Gradienten über Dreiecke

Alternativ konstruiert man komponentenweise zu jeder Box unter Berücksichtigung aller direkten Nachbarzellen eine lineare Verteilung entsprechend der Form

$$\mathbf{u}_i = \mathbf{u}_{\Omega_i} + \sigma_{\Omega_i} \sum_{j=1,2} (\nabla \mathbf{u}_{\Omega_i})_j \left(\dot{k}_j - (\Omega_i)_j \right), \tag{4.17}$$

wobei

$$\nabla \mathbf{u}_{\Omega_i} = \frac{1}{|\Omega_i|} \sum_{j \in T(i)} |F_j| \nabla \mathbf{u}_{\triangle_{(j)}}$$

die Näherung an den Gradienten innerhalb einer Box meint. Sei $T(i) = \{j | \triangle_j \in \mathcal{T}, \Omega_i \cap \triangle_j \neq \emptyset\}$, die Indexmenge aller Dreiecke, die in Teilmengenbeziehung mit Ω_i stehen.

Für $j \in T(i)$ ist $F_j = \triangle_j \cap \Omega_i$ die zur Box Ω_i gehörende Menge des Dreiecks $\triangle_j$.

$\nabla \mathbf{u}_{\triangle_{(j)}}$ bezeichnet den eindeutig bestimmten Gradienten auf dem Dreieck, dessen Ecken durch die Schwerpunkte $(\Omega_{j_1}), (\Omega_{j_2})$ und (Ω_{j_3}) der drei Boxen bestimmt werden für die $\Omega_{j_m} \cap \triangle_j \neq \emptyset, m = 1, 2, 3$ gilt. Die konkrete Berechnungsvorschrift findet sich z.B. in [Lud99].

Die Näherungen an die Verteilung in der Zelle sind also mittels Polynomen erster Ordnung dargestellt. Der Faktor $\sigma_{\Omega_i} \in \left(\mathbb{R}^+ \cup \{0\}\right)^N$ soll das Erzeugen neuer Extrema in den während der Rechnung verwendeten Näherungen durch die Rekonstruktion verhindern.

Wie ist der Limitierungsfaktor, σ_{Ω_i}, des Gradienten, $\nabla \mathbf{u}_{\Omega_i}$, in (4.17) zu wählen? Es werden wiederum zwei Möglichkeiten angeboten, einerseits der klassische Limiter nach Barth und Jesperson [BJ89] und andererseits der Limiter von Venkatakrishnan [Ven95] in einer Formulierung von Aftosmis et al. [AGT95].

Limiter

Das Konstruktionsziel des Limiters nach Barth und Jesperson (LBJ) aus [BJ89] ist die Verhinderung von neuen Extrema im Verlauf der Rechnung durch den Prozess der Rekonstruktion. Dies wird angestrebt über eine Einschränkung der rekonstruierten Werte innerhalb jeder Box Ω_i, $\mathbf{u}_i$, an jeder Zellkante k_{ijl} zwischen den Zellmittelwerten aller anliegenden Zellen, $\mathbf{u}_{\Omega_j}, j \in N(i)$, und $\mathbf{u}_{\Omega_i}$ selbst.

Für eine Box Ω_i geht man dabei wie folgt vor. Für $m = 1, \ldots, N$ sei

$$(u_{\min})_m = \min_{k \in \{j \in N(i) \cup \{i\}\}} \left((\mathbf{u}_{\Omega_k})_m \right) \text{ und } (u_{\max})_m = \max_{k \in \{j \in N(i) \cup \{i\}\}} \left((\mathbf{u}_{\Omega_k})_m \right).$$

Mit Hilfe dieser Werte bestimmt man $\sigma_{\Omega_i} \in [0, 1]^N$ komponentenweise maximal, so dass nach (4.17)

$$\mathbf{u}_{\min} \leq \mathbf{u}_i \leq \mathbf{u}_{\max}$$

gilt.

Alternativ dazu bestimmt der Limiter nach Venkatakrishnan (LV) aus [Ven95] in einer Formulierung nach [AGT95] ein $\sigma_{\Omega_i} \in \left(\mathbb{R}^+ \cup \{0\}\right)^N$ für jede Zellkante, k_{ijl}, und wählt komponentenweise das Minimum.

Der Limiter verwendet an jeder inneren Kante, k_{ijl}, in jeder Komponente, m,

$$\left(\sigma_{\Omega_i}^{jl}\right)_m = \frac{dp_m dp_m + e1 + 2dp_m df_m}{dp_m dp_m + 2df_m df_m + dp_m df_m + e2}.$$

Dabei ist

$$\mathbf{df} = \sum_{l=1,2} (\nabla \mathbf{u}_{\Omega_i})_m \left(\dot{k}_m - (\Omega_i)_m \right)$$

(vergleiche (4.17)), $e1$ eine charakteristische Länge (Radius des größten Kreises um den Schwerpunkt, der noch vollständig in der Zelle Platz findet) in dritter Potenz, $e2 = 10^{-6}$ eine Konstante zur Verhinderung der Division durch Null und

$$\mathbf{dp} = \begin{cases} \mathbf{u}_{\Omega_j} - \mathbf{u}_{\Omega_i} & , \quad \text{sign}(\mathbf{u}_{\Omega_j} - \mathbf{u}_{\Omega_i}) = \text{sign}(\mathbf{df}) \\ 0 & , \quad \text{sonst} \end{cases} .$$

Abschließend wählt man

$$(\sigma_{\Omega_i})_m = \min_{j \in N(i)} \left(\left(\sigma^{jl}_{\Omega_i} \right)_m \right) .$$

Die beiden Limiter zeigen charakteristisch unterschiedliches Verhalten. Dies zeigt die Abbildung 4.3. Es ist zu beachten, dass die Graphiken jeweils Schnitte durch ein 2D Netz veranschaulichen. Die durchgehenden Linien sind die angedeuteten integralen Mittelwerte $\mathbf{u}_{\Omega_i}$ usw. Die unterbrochenen Linien stellen mögliche ungedämpfte Rekonstruktionen dar.

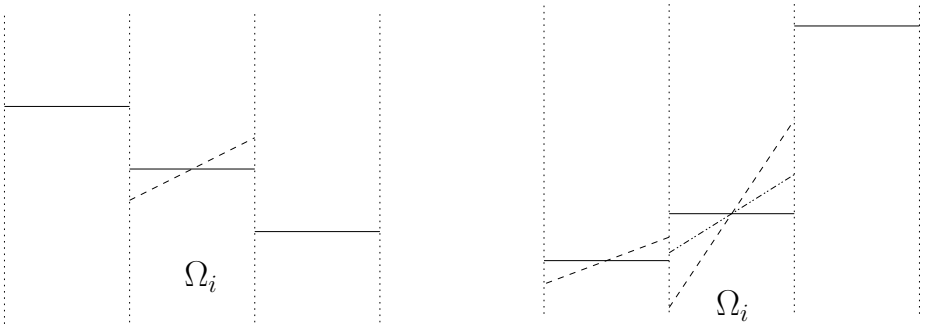

Abbildung 4.3: Limitervergleich

Der in der linken Graphik illustrierte Fall könnte durch 2D Effekte auftreten, wenn man den gewählten Schnitt entsprechend „geschickt“ wählt. Der LBJ würde hier nicht dämpfend wirken. Der LV würde in diesem Fall $\sigma_{\Omega_i} = \mathbf{0}$ setzen.

In der von der rechten Graphik dargestellten Situation würde, unter Vernachlässigung weiterer Einflüsse von nicht abgebildeten Zellen, der LV stärker dämpfen als der LBJ. Dort ist in der Zelle Ω_i auch noch die Limitierung durch den LV angedeutet. Der LBJ würde hier die Rekonstruktion am linken Rand der Zelle Ω_i bis auf das Niveau des integralen Mittelwertes der links liegenden Zelle anheben, aber

nicht darüber hinaus. Man erkennt deutlich, dass beide nicht garantieren können, dass die Box mit dem größeren Mittelwert auch in der Kantenrekonstruktion den größeren Wert erhält.

4.2.2 Zellkanten an den Rändern

Für die Kanten auf den Rändern des Rechengebiets mit $l = \zeta_1, \zeta_2, \zeta_3$ hat man nur einseitige Informationen aus der laufenden Rechnung oder den Anfangswerten bezüglich der Zustandsgrößen zur Lösung des Riemannproblems. Man kennt also in (4.14) nur $\mathbf{u}_l$ und nicht $\mathbf{u}_r$. Daher lässt sich (4.14) nicht direkt verwenden.

Für die Ein- und Ausströmränder wird ein Ghost Cell Ansatz verwendet, um dieses Informationsdefizit behandeln zu können.

An den Einströmrändern gibt es eine Funktion φ, vergleiche (3.11), welche hier Zustände vorschreibt. Man setzt $\mathbf{u}_r = \varphi$ und erhält für die Kantenflüsse auf Einströmrändern

$$\mathbf{g}^{HLL}_{ij\zeta_1}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{g}^{HLL}(\mathbf{u}_i, \varphi, \mathbf{n}_{ij\zeta_1}). \tag{4.18}$$

Für die Ausströmränder werden von Neumann Randbedingungen mit $\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = \mathbf{0}$ angesetzt. Hier wird entsprechend $\mathbf{u}_r = \mathbf{u}_l$ betrachtet und ergibt für die Kantenflüsse

$$\mathbf{g}^{HLL}_{ij\zeta_2}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{g}^{HLL}(\mathbf{u}_i, \mathbf{u}_i, \mathbf{n}_{ij\zeta_2}). \tag{4.19}$$

Für die Kanten k auf der festen Wand ist $\mathbf{vn} = 0$ festgelegt, d.h. es gibt keine Flüsse senkrecht zu einer festen (undurchdringlichen) Wand. Verwendet man dies in (3.9), so vereinfachen sich die relevanten konvektiven Flüsse zu

$$\mathbf{g}^{HLL}_{ij\zeta_3}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{f}^c_1\left(T(\mathbf{n}_{ij\zeta_3})\,\mathbf{u}_i\right) \overset{\mathbf{vn}=0}{=} \begin{pmatrix} 0 \\ \frac{1}{2}\phi^2 n_1 \\ \frac{1}{2}\phi^2 n_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.20}$$

Man erhält abschließend für (4.9) durch sukzessives Einsetzen von (4.10), (4.11) und (4.13) und den Definitionen (4.15), (4.18), (4.19) und (4.20)

$$\mathcal{L}^c = -\frac{1}{|\Omega|}\left(\sum_{j\in N(i)} \sum_{l\in(N^+(i)\cap N^+(j))} |k_{ijl}|\left[T^{-1}_{ijl}\mathbf{g}^{HLL}_{ijl}(\mathbf{u}_i, \mathbf{u}_j)\right]\right). \tag{4.21}$$

4.3 Die Diskretisierung der viskosen Terme

Zur numerischen Lösung des Integrals

$$\begin{aligned}\mathcal{L}^d &= \frac{1}{|\Omega_i|}\int_{\partial\Omega_i}\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u})n_m d\mathbf{s}\\ &= \frac{1}{|\Omega_i|}\sum_{j\in N(i)}\sum_{l\in(N^+(i)\cap N^+(j))}\int_{k_{ijl}}\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u})\left(n_{ijl}\right)_m d\mathbf{s}\end{aligned}\tag{4.22}$$

über einer Zelle Ω_i, unter Zuhilfenahme der Netzstruktur (vergleiche dazu auch (4.10)), ist es hilfreich, sich die konkrete Form von $\mathbf{f}_m^d(\mathbf{u})$ zu Nutze zu machen. Man findet in (3.9), dass

$$\left(\mathbf{f}_m^d(\mathbf{u})\right)_4 = \mathrm{Diff}_{PS}\partial_{x_m}PS \text{ und } \left(\mathbf{f}_m^d(\mathbf{u})\right)_6 = \mathrm{Diff}_{PE_I}\partial_{x_m}PE_I$$

ist. Für alle anderen Komponenten gilt $\left(\mathbf{f}_m^d(\mathbf{u})\right)_j = 0, j \in \{1,\ldots,17\}\setminus\{4,6\}$.

Das Integral wird mittels Mittelpunktsregel, siehe (4.11), durch direktes Einsetzen am Kantenmittelpunkt

$$\int_{k_{ijl}}\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u})\left(\mathbf{n}_{ijl}\right)_m d\mathbf{s} \approx |k_{ijl}|\left(\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u}_{ijl})\left(\mathbf{n}_{ijl}\right)_m\right)\tag{4.23}$$

ausgewertet, dabei ist $\mathbf{u}_{ijl} \approx \mathbf{u}\left(\dot{\mathbf{k}}_{ijl}\right)$. Einsetzen für die Komponente PS liefert

$$|k_{ijl}|\left(\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u}_{ijl})\left(\mathbf{n}_{ijl}\right)_m\right)_4 = \mathrm{Diff}_{PS}|k_{ijl}|\left(\partial_{x_1}PS\left(\mathbf{n}_{ijl}\right)_1 + \partial_{x_2}PS\left(\mathbf{n}_{ijl}\right)_2\right).$$

Man benötigt nun Schätzungen für die Gradienten in den Größen PS und PE_I an den Integrationspunkten.

4.3.1 Viskose Flüsse über innere Kanten

Es gelte $l \neq \zeta_1, \zeta_2, \zeta_3$. Die Integrationspunkte liegen für alle Kanten, die nicht auf dem Rand des Rechengebiets liegen, nach Konstruktion des Netzes innerhalb eines Dreiecks. Die Gradienten $\nabla_{\triangle_{ijl}}PS$ auf den Dreiecken $\triangle_{ijl}$ in der Größe PS sind eindeutig bestimmt über eine Lokalisierung der Zellmittelwerte an den Dreiecksecken, vergleiche dazu das auf (4.17) Folgende. Auch hier entspricht die Berechnung der Formel aus [Lud99]. Direktes Einsetzen liefert

$$|k_{ijl}|\left(\sum_{m=1,2}\mathbf{f}_m^d(\mathbf{u}_{ijl})\left(\mathbf{n}_{ijl}\right)_m\right)_4 = \mathrm{Diff}_{PS}|k_{ijl}|\left(\nabla_{\triangle_{ijl}}PS\cdot\mathbf{n}_{ijl}\right)$$

und analog

$$|k_{ijl}| \left(\sum_{m=1,2} \mathbf{f}_m^d(\mathbf{u}_{ijl}) \left(\mathbf{n}_{ijl}\right)_m \right)_6 = \mathrm{Diff}_{PE_I} |k_{ijl}| \left(\nabla_{\triangle_{ijl}} PE_I \cdot \mathbf{n}_{ijl} \right).$$

Setze den diffusiven Fluss für $l \neq \zeta_1, \zeta_2, \zeta_3$

$$H_{ijl}\left(\mathbf{u}_{\Omega_i}, \mathbf{u}_{\Omega_j}, \mathbf{u}_{\Omega_l}\right) = \mathrm{Diff}_{\mathbf{u}} \left(\nabla_{\triangle_{ijl}} \left(\mathbf{u}_{\Omega_i}, \mathbf{u}_{\Omega_j}, \mathbf{u}_{\Omega_l}\right) \cdot \mathbf{n}_{ijl} \right), \tag{4.24}$$

dabei ist $\mathrm{Diff}_{\mathbf{u}} = (0,\ 0,\ 0,\ \mathrm{Diff}_{PS},\ 0,\ \mathrm{Diff}_{PE_I},\ 0,\ 0,\ 0, 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)^T$ der Vektor der Diffusionskoeffizienten. Man lokalisiert zur Bestimmung des Gradienten $\nabla_{\triangle_{ijl}} \left(\mathbf{u}_{\Omega_i}, \mathbf{u}_{\Omega_j}, \mathbf{u}_{\Omega_l}\right)$ in nahe liegender Weise die Werte $\mathbf{u}_{\Omega_i}, \mathbf{u}_{\Omega_j}, \mathbf{u}_{\Omega_l}$ an den korrespondierenden Punkten $\dot{\Omega}_i, \dot{\Omega}_j, \dot{\Omega}_l$.

4.3.2 Viskose Flüsse über Randkanten

Für Kanten, die auf Ausströmrändern liegen oder auf der festen Wand, gibt es auf Grund der Randbedingungen keine diffusiven Flüsse. Daher wird für $l = \zeta_2, \zeta_3$

$$H_{ijl}(\mathbf{u}_i, \mathbf{u}_j, \mathbf{u}_l) = \mathbf{0} \tag{4.25}$$

gesetzt.

Für die Randkanten, welche auf einem Einströmrand liegen, ist das bereits beschriebene Vorgehen nicht direkt umsetzbar. Natürlich will man aber die Form (4.24) nutzen. Zu klären ist also, welche Form die Funktion H haben soll. Auch hierbei wird sich der Formulierung von Ludwig aus [Lud99] bedient. Allerdings ist es nötig, nochmals Notation einzuführen.

Definition 4.1 *Zu jeder Box Ω_i mit wenigstens einer Einströmkante sei*

$$\mathbf{p}_i = \dot{\Omega}_i + \frac{\sqrt{3}}{8} \mathbf{r} \min \left(\left\| \dot{\Omega}_p - \dot{\Omega}_i \right\|_2, \left\| \dot{\Omega}_q - \dot{\Omega}_i \right\|_2 \right).$$

Dabei sind $\dot{\Omega}_p, \dot{\Omega}_i, \dot{\Omega}_n, \dot{\Omega}_j$ und $\dot{\Omega}_q$ die Koordinaten der jeweiligen Ecken der Dreiecke, vergleiche Abbildung 4.4 und $\mathbf{r}$ bezeichnet die normierte Winkelhalbierende zu dem Winkel $\angle\left(\dot{\Omega}_n, \dot{\Omega}_i, \dot{\Omega}_j\right)$. Beachte dabei, dass $\dot{\Omega}_p$ und $\dot{\Omega}_q$ die Ecken zu dem Dreieck sind, in welches $\mathbf{r}$ zeigt.

Nun hat man ausreichend definierte Koordinaten zur Verfügung, um Dreiecke zu wählen, über die eine sinnvolle Gradientenschätzung möglich ist. Setze daher für die Einströmränder, also $l = \zeta_1$,

$$H_{ij\zeta_1}\left(\mathbf{u}_{\Omega_i}, \mathbf{u}_{\Omega_j}, \mathbf{u}_{\Omega_{\zeta_1}}\right) = \mathrm{Diff}_{\mathbf{u}} \left(\nabla_{\triangle_{\left(\mathbf{p}_i, \dot{\Omega}_i, \mathbf{x}_{ij}\right)}} \left(\mathbf{u}_{\Omega_i}, \varphi, \varphi\right) \cdot \mathbf{n}_{ij\zeta_1} \right). \tag{4.26}$$

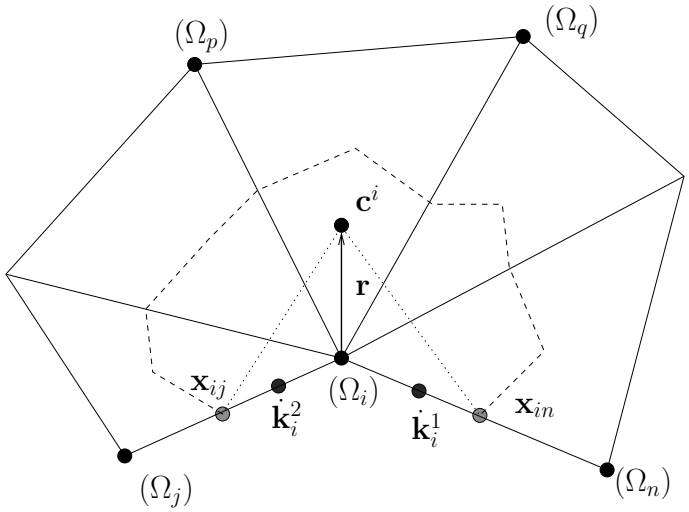

Abbildung 4.4: Randbehandlung: Diffusion

Man bestimmt also den Gradienten, $\nabla_{\triangle_{\left(\mathbf{p}_i,\dot{\Omega}_i,\mathbf{x}_{ij}\right)}}(\mathbf{u}_{\Omega_i},\varphi,\varphi)$, durch eine Lokalisierung der Werte von $\mathbf{u}_{\Omega_i}$ an $\mathbf{p}_i$ und der Werte von φ an den beiden Punkten auf dem Rand des Rechengebiets $\dot{\Omega}_i$ und $\mathbf{x}_{ij}$.

Insgesamt liefert einsetzen von (4.23) und den Definitionen von H (4.26), (4.25) und (4.24) in (4.22)

$$\mathcal{L}^d = \frac{1}{|\Omega_i|} \sum_{j\in N(i)} \sum_{l\in N^+(i)\cap N^+(j)} |k_{ijl}|\, H_{ijl}(\mathbf{u}_i,\mathbf{u}_j,\mathbf{u}_l). \tag{4.27}$$

4.4 Implizite Flüsse

Legt man $t = t^n$ fest, so ergeben sich durch das bisher beschriebene (Unterkapitel 4.2 und 4.3) explizite Berechnungsvorschriften. Dies bringt aber auch starke Einschränkung der numerisch zulässigen Zeitschrittweite, h, mit sich, da aus Stabilitätsgründen der numerische Abhängigkeitsbereich den physikalischen umfassen muss. Daher wird durch die konvektiven Flüsse für die Schrittweite, h, des expliziten Verfahrens für jede Box Ω_i eine Beschränkung in der Form

$$h \leq CFL \frac{x_{min}}{|v| + \Phi} \tag{4.28}$$

angewendet, dabei ist $CFL \in (0,1)$, x_{min} der Radius des größten Kreises, der vollständig in der Box Ω_i liegt, Φ das Geopotential und $|v| = \sqrt{v_1^2 + v_2^2}$.

Unter Zuhilfenahme der gleichen Notation werden für die Schrittweite des expliziten Verfahrens durch jede Box Ω_i Einschränkungen durch die viskosen

Flüsse in der Form

$$h \leq CFL \frac{\left(\frac{x_{min}}{2}\right)^2}{d} \tag{4.29}$$

angenommen, wobei $d = \max(\text{Diff}_{PS}, \text{Diff}_{PE_I})$ das Maximum aus den beiden Diffusionskoeffizienten ist.

Insgesamt wird für explizite Verfahren das maximale h gewählt, welches für alle Boxen Ω_i den Bedingungen aus (4.28) und (4.29) genügt.

Um dieser Schrittweitenschranke theoretisch nicht unterworfen zu sein, bieten sich implizite Verfahren an. Dafür werden in den Flüssen $\mathbf{g}^{HLL}$ und H die Argumente entsprechend gesetzt. Das explizite Verfahren ist gegeben durch

$$\mathbf{g}^{HLL} = \mathbf{g}^{HLL}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right),$$

das Implizite durch

$$\mathbf{g}^{HLL} = \mathbf{g}^{HLL}\left(\mathbf{u}_l^{n+1}, \mathbf{u}_r^{n+1}, \mathbf{n}\right).$$

Zu beachten ist, dass n bzw. $(n+1)$ hier den Zeitschritt meint und nicht den Normalenvektor $\mathbf{n}$. Da aus offensichtlichen Gründen $\mathbf{u}_l^{n+1}$ und $\mathbf{u}_r^{n+1}$ aber noch nicht bekannt sind, verwendet man abgebrochene Taylorentwicklungen der Art

$$\begin{aligned} &\mathbf{g}^{HLL}\left(\mathbf{u}_l^{n+1}, \mathbf{u}_r^{n+1}, \mathbf{n}\right) \\ &\quad\approx \mathbf{g}^{HLL}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right) + J^{HLL}_{\mathbf{u}_l}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right)\left(\triangle \mathbf{u}_l^n\right) + J^{HLL}_{\mathbf{u}_r}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right)\left(\triangle \mathbf{u}_r^n\right) \\ &\quad\approx \mathbf{g}^{HLL}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right) + J^{HLL}_{\mathbf{u}_l}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right)\left(\triangle \mathbf{u}_{\Omega_l}^n\right) + J^{HLL}_{\mathbf{u}_r}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right)\left(\triangle \mathbf{u}_{\Omega_r}^n\right). \end{aligned} \tag{4.30}$$

Hierbei bezeichnen $\left(J^{HLL}_{\mathbf{u}_l}\right)_{ij}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right) = \frac{\partial g_i^{HLL}}{\partial (\mathbf{u}_l)_j}\left(\mathbf{u}_l^n, \mathbf{u}_r^n, \mathbf{n}\right)$ respektive $J^{HLL}_{\mathbf{u}_r}$ die jeweiligen Jakobimatrizen und $\triangle \mathbf{u}_l^n = \mathbf{u}_l^{n+1} - \mathbf{u}_l^n$ bzw. $\triangle \mathbf{u}_r^n$ die Differenzenvektoren der Zustandsgrößen oder, alternativ formuliert, die Updates der Zustandsvektoren. Die Jakobimatrizen werden durch Differenzenquotienten bestimmt.

Genauso kann man mit den viskosen Flüssen H verfahren,

$$\begin{aligned} &H_{ijl}\left(\mathbf{u}_{\Omega_i}^{n+1}, \mathbf{u}_{\Omega_j}^{n+1}, \mathbf{u}_{\Omega_l}^{n+1}\right) \\ &\quad\approx H_{ijl}\left(\mathbf{u}_{\Omega_i}^n, \mathbf{u}_{\Omega_j}^n, \mathbf{u}_{\Omega_l}^n\right) + J^H_{\mathbf{u}_{\Omega_i}}\left(\mathbf{u}_{\Omega_i}^n, \mathbf{u}_{\Omega_j}^n, \mathbf{u}_{\Omega_l}^n\right)\left(\triangle \mathbf{u}_{\Omega_i}^n\right) \\ &\qquad + J^H_{\mathbf{u}_{\Omega_j}}\left(\mathbf{u}_{\Omega_i}^n, \mathbf{u}_{\Omega_j}^n, \mathbf{u}_{\Omega_l}^n\right)\left(\triangle \mathbf{u}_{\Omega_j}^n\right) + J^H_{\mathbf{u}_{\Omega_l}}\left(\mathbf{u}_{\Omega_i}^n, \mathbf{u}_{\Omega_j}^n, \mathbf{u}_{\Omega_l}^n\right)\left(\triangle \mathbf{u}_{\Omega_l}^n\right). \end{aligned} \tag{4.31}$$

4.5 Das verwendete Gesamtverfahren

Das Gesamtverfahren für die Gleichung (3.8) besteht, wie in (4.5) und (4.6) dargelegt, in zwei Berechnungsschritten. Diese werden identifiziert durch die Ausdrücke

ϕ_1 und ϕ_2 aus (4.7) und (4.8). Ihre konkrete Gestalt ist eine Zusammenführung der Überlegungen aus Kapitel 2 sowie den vorhergehenden Betrachtungen dieses Kapitels.

4.5.1 Die Gestalt von ϕ_1

Die vollständige rechte Seite des Systems (3.8) stellt stark unterschiedliche Anforderungen an die zur Lösung verwendeten Verfahren. Die Biomassen (BZ, BA_i, $i = 1,\ 2,\ 3,4$) sowie die Wasserhöhe in Φ (und damit auch Φ selbst) sind in natürlicher Weise positiv. Für diese Größen wird das Verfahren von Patankar aus [Pat80] in seiner dort beschriebenen ursprünglichen Form verwendet. Dies garantiert positive Näherungen.

Alle beteiligten Größen des Phosphorzyklus ($PS, PD, PE_I, PE_O, PZ, PA_i, i = 1,2,3,4$) sind nicht nur positiv, sondern auch konservativ bzw. Masse erhaltend nach Definition 2.5. Daher finden die in Unterkapitel 2.2 vorgestellten modifizierten Patankar Verfahren Anwendung.

Die Geschwindigkeiten (und damit auch die Größen Φv_1 und Φv_1) sind nicht eingeschränkt. Für sie werden explizite Runge Kutta Verfahren verwendet, wie sie in jedem Textbuch zur Numerik von gewöhnlichen Differentialgleichungen ausführlich besprochen werden (klassische Monographien, die sich nur mit diesem Thema beschäftigen, sind [HNW91, HWN02]).

4.5.2 Die Gestalt von ϕ_2

Für die Zeitintegration des Strömungsanteils wird im Fall des expliziten Verfahrens zur Wahrung der Positivität der Ergebnisse das Emini Verfahren verwendet. Für das implizite Verfahren wird das explizite Euler für die Zeitintegration verwendet. Falls negative Näherungen berechnet werden, wird die Schrittweite reduziert und ein erneuter Schritt ausgeführt.

Setzt man (4.21) und (4.27) für eine Box Ω_i zusammen, so erhält man für explizites ϕ_2 aus (4.8) zum Zeitpunkt t^n

$$\begin{aligned}\phi_2 &= \left(\mathcal{L}^c + \mathcal{L}^d\right)^n_{i,ex} \\ &= -\frac{1}{|\Omega|}\left(\sum_{j\in N(i)}\ \sum_{l\in(N^+(i)\cap N^+(j))} |k_{ijl}|\left[T^{-1}_{ijl}\mathbf{g}^{HLL}_{ijl}\left(\mathbf{u}^n_i,\mathbf{u}^n_j\right)\right]\right)\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{|\Omega_i|}\sum_{j\in N(i)}\sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|H_{ijl}(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{u}_l^n)\\
=\;&\frac{1}{|\Omega_i|}\sum_{j\in N(i)}\sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\left[H_{ijl}(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{u}_l^n)-\;T_{ijl}^{-1}\mathbf{g}_{ijl}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n\right)\right].
\end{aligned}
$$

Möchte man die Zeischrittweitenrestriktion aufheben und wendet das implizite Verfahren an, wird daraus unter Verwendung von (4.30) und (4.31)

$$
\begin{aligned}
&\left(\mathcal{L}^c+\mathcal{L}^d\right)_{i,impl}^n\\
=\;&\frac{1}{|\Omega_i|}\sum_{j\in N(i)}\sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\Bigg[\\
&H_{ijl}\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)+J_{\mathbf{u}_{\Omega_i}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)\left(\triangle\mathbf{u}_{\Omega_i}^n\right)\\
&+J_{\mathbf{u}_{\Omega_j}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)\left(\triangle\mathbf{u}_{\Omega_j}^n\right)+J_{\mathbf{u}_{\Omega_l}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)\left(\triangle\mathbf{u}_{\Omega_l}^n\right)\\
&-T_{ijl}^{-1}\Bigg[\mathbf{g}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\\
&+J_{\mathbf{u}_i}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\left(\triangle\mathbf{u}_{\Omega_i}^n\right)+J_{\mathbf{u}_j}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\left(\triangle\mathbf{u}_{\Omega_j}^n\right)\Bigg]\Bigg]\\
=\;&\frac{1}{|\Omega_i|}\sum_{j\in N(i)}\sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\Bigg[\\
&H_{ijl}\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)-T_{ijl}^{-1}\mathbf{g}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\\
&+\left[J_{\mathbf{u}_{\Omega_i}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)-\;T_{ijl}^{-1}J_{\mathbf{u}_i}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\right]\left(\triangle\mathbf{u}_{\Omega_i}^n\right)\\
&+\left[J_{\mathbf{u}_{\Omega_j}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)-\;T_{ijl}^{-1}J_{\mathbf{u}_j}^{HLL}\left(\mathbf{u}_i^n,\mathbf{u}_j^n,\mathbf{n}\right)\right]\left(\triangle\mathbf{u}_{\Omega_j}^n\right)\\
&+\left[J_{\mathbf{u}_{\Omega_l}}^H\left(\mathbf{u}_{\Omega_i}^n,\mathbf{u}_{\Omega_j}^n,\mathbf{u}_{\Omega_l}^n\right)\right]\left(\triangle\mathbf{u}_{\Omega_l}^n\right)\Bigg].
\end{aligned}
$$

Löst man nun Gleichung (4.5) mittels des expliziten Euler Verfahrens, ergibt sich für Zelle Ω_i

$$
\mathbf{u}_{\Omega_i}^{n+1}=\mathbf{u}_{\Omega_i}^n+h\;\left(\mathcal{L}^c+\mathcal{L}^d\right)_{i,ex}^n
$$

oder im Falle des impliziten Verfahrens

$$
\triangle\mathbf{u}_{\Omega_i}^n=h\;\left(\mathcal{L}^c+\mathcal{L}^d\right)_{i,impl}^n.
$$

Einsetzen und Umstellen ergibt für die impliziten Updates für Zelle Ω_i

$$\begin{aligned}
&\left[1-\tfrac{h}{|\Omega_i|}\sum_{j\in N(i)}\ \sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\left[J^H_{\mathbf{u}_{\Omega_i}}\left(\mathbf{u}^n_{\Omega_i},\mathbf{u}^n_{\Omega_j},\mathbf{u}^n_{\Omega_l}\right)-T^{-1}_{ijl}\,J^{HLL}_{\mathbf{u}_i}\left(\mathbf{u}^n_i,\mathbf{u}^n_j,\mathbf{n}\right)\right]\right]\left(\triangle\mathbf{u}^n_{\Omega_i}\right)\\
&-\left[\tfrac{h}{|\Omega_i|}\sum_{j\in N(i)}\ \sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\left[J^H_{\mathbf{u}_{\Omega_j}}\left(\mathbf{u}^n_{\Omega_i},\mathbf{u}^n_{\Omega_j},\mathbf{u}^n_{\Omega_l}\right)-T^{-1}_{ijl}\,J^{HLL}_{\mathbf{u}_j}\left(\mathbf{u}^n_i,\mathbf{u}^n_j,\mathbf{n}\right)\right]\right]\left(\triangle\mathbf{u}^n_{\Omega_j}\right)\\
&-\left[\tfrac{h}{|\Omega_i|}\sum_{j\in N(i)}\ \sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\left[J^H_{\mathbf{u}_{\Omega_l}}\left(\mathbf{u}^n_{\Omega_i},\mathbf{u}^n_{\Omega_j},\mathbf{u}^n_{\Omega_l}\right)\right]\right]\left(\triangle\mathbf{u}^n_{\Omega_l}\right)\\
&=\tfrac{h}{|\Omega_i|}\sum_{j\in N(i)}\ \sum_{l\in N^+(i)\cap N^+(j)}|k_{ijl}|\left[H_{ijl}\left(\mathbf{u}^n_{\Omega_i},\mathbf{u}^n_{\Omega_j},\mathbf{u}^n_{\Omega_l}\right)-T^{-1}_{ijl}\,\mathbf{g}^{HLL}\left(\mathbf{u}^n_i,\mathbf{u}^n_j,\mathbf{n}\right)\right].
\end{aligned}$$

Zusammensetzen über alle Zellen, $\Omega_1,\ldots,\Omega_{N_{Box}}$, liefert eine schwach besetzte Matrix L mit $(N_{Box}\times N)^2$ Einträgen. Man erhält insgesamt das System

$$L\triangle\mathbf{u}^n = h\ \left(\mathcal{L}^c+\mathcal{L}^d\right)^n_{ex}$$

mit $\left(\mathcal{L}^c+\mathcal{L}^d\right)^n_{ex}=\left(\left(\mathcal{L}^c+\mathcal{L}^d\right)^n_{1,ex},\ldots,\left(\mathcal{L}^c+\mathcal{L}^d\right)^n_{N_{Box},ex}\right)^T$ und $\triangle\mathbf{u}^n=\left(\triangle\mathbf{u}^n_{\Omega_1},\ldots,\triangle\mathbf{u}^n_{\Omega_{N_{Box}}}\right)^T$.

Das Lösen dieses Gleichungssystems geschieht mit Hilfe iterativer Löser auf die an dieser Stelle nicht weiter eingegangen werden soll. Für den interessierten Leser sei hier auf [Mei01, Mei99, Bir05] verwiesen. Ein zentraler Bestandteil des impliziten Algorithmus ist die Invertierung von Matrizen der Dimension N. Im hier betrachteten Fall gilt $N=17$. Für diese Invertierung wird das Paket Lapack aus [Fou09] verwendet.

Die Ergebnisse des Gesamtsystemlösers finden sich in Unterkapitel 6.3.

Zur Berechnung des kreisförmigen Dammbruchproblems (Unterkapitel 3.1.2) wird mit der nahe liegenden Einschränkung von $\mathbf{u}$ auf $\mathbf{u}_{swe}$ ebenso der hier vorgestellte Algorithmus verwendet. Die Ergebnisse zu diesem Testfall befinden sich in 6.2.

Teil III

Ergebnisse

Kapitel 5

Ergebnisse für die gewöhnlichen Differentialgleichungssysteme aus Kapitel 1

In diesem Kapitel werden alle Ergebnisse zu den Problemen aus Kapitel 1 und den Verfahren aus Kapitel 2 gezeigt und besprochen. Das erste Unterkapitel widmet sich der Darstellung zum Problem aus Beispiel 1.1. Das zweite Unterkapitel widmet sich unter Verwendung des Problems aus Beispiel 1.1 und der dazu gehörenden analytischen Lösung einer Ordnungsanalyse aller vorgestellten Verfahren aus Kapitel 2.

Das Unterkapitel 5.3 zeigt die Ergebnisse zu dem System aus Beispiel 1.2 und den dazu passenden Verfahren aus Kapitel 2. Im letzten Unterkapitel wird das Modell aus Beispiel 1.4 numerisch behandelt.

5.1 Numerische Bearbeitung des linearen Systems (1.1) durch die Verfahren aus Kapitel 2

Im folgenden Abschnitt werden die numerischen Ergebnisse zum System aus Beispiel 1.1 präsentiert. Das System beschreibt den Masseaustausch von Komponente c_1 zu c_2 bis zu einem asymptotisch angestrebten Endzustand des Systems. Abbildung 1.1 zeigt die analytische Lösung.

Eine quantitative Untersuchung der Fehler und eine einhergehende Ordnungsbetrachtung aller vorgestellten Verfahren findet in Unterkapitel 5.2 statt, daher

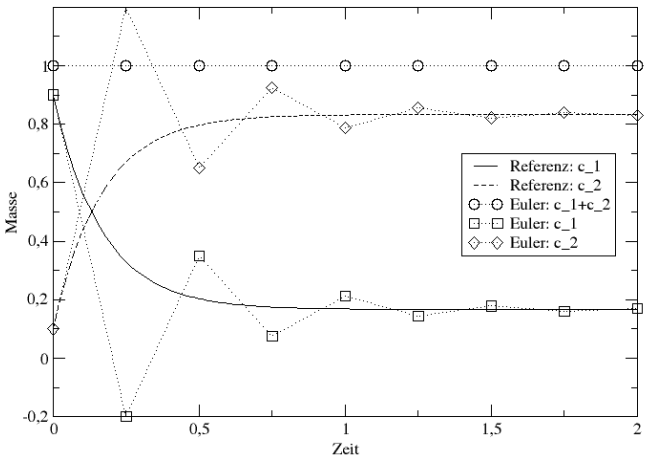

Abbildung 5.1: Das Euler-Verfahren mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

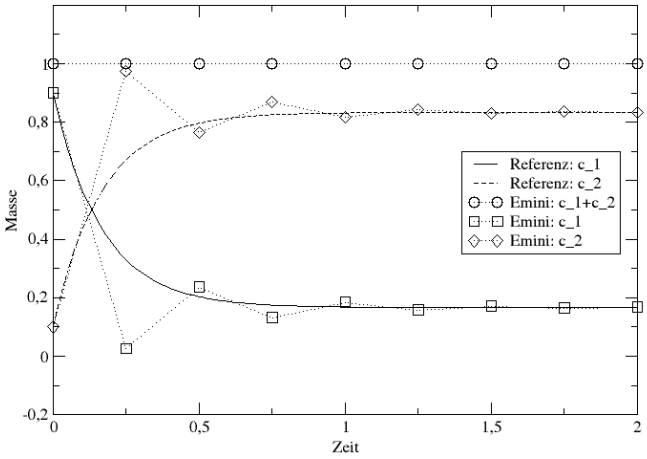

Abbildung 5.2: Das Emini Verfahren mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

findet in diesem Unterkapitel keine Quantifizierung der Fehler statt.

Das Euler Verfahren (Abbildung 5.1) liefert bei der gewählten Schrittweite im ersten Berechnungsschritt negative Ergebnisse. Es ist aber in der Lage, den Endzustand des Systems korrekt wiederzugeben.

Das Emini Verfahren (Abbildung 5.2) liefert keine negativen Ergebnisse. Al-

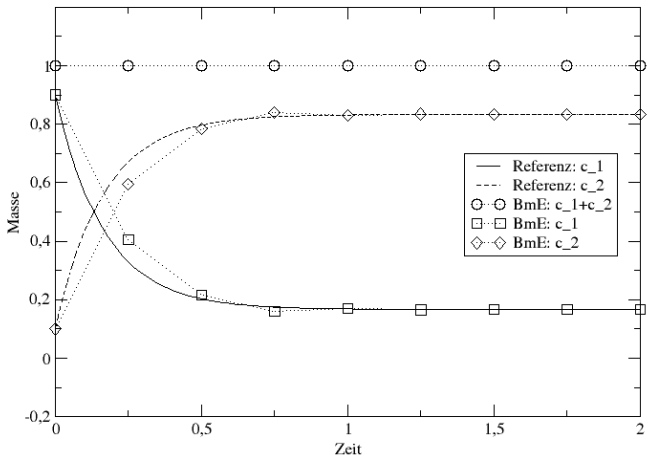

Abbildung 5.3: Das BmE Verfahren mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

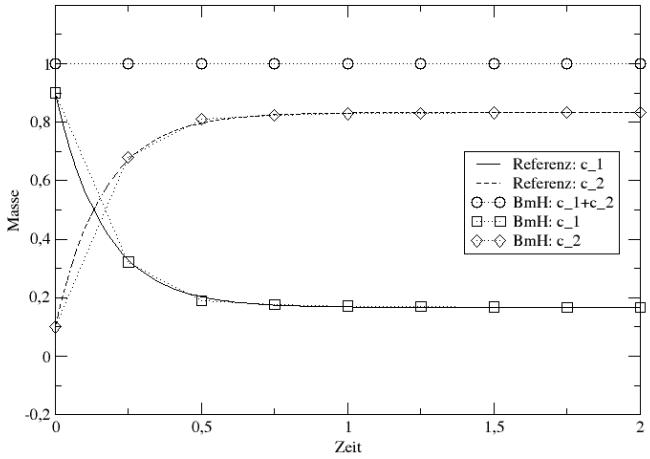

Abbildung 5.4: Das BmH mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

lerdings oszillieren die berechneten Lösungen stark um die analytische Lösung. Den Endzustand gibt auch das Emini Verfahren zuverlässig wieder.

Die Bruggeman Verfahren (BmE Abbildung 5.3 und BmH Abbildung 5.4) liefern, wie erwartet, ausschließlich positive Ergebnisse. Die berechneten Näherungen liegen in unmittelbarer Nähe der analytischen Lösung, wobei die Näherungen

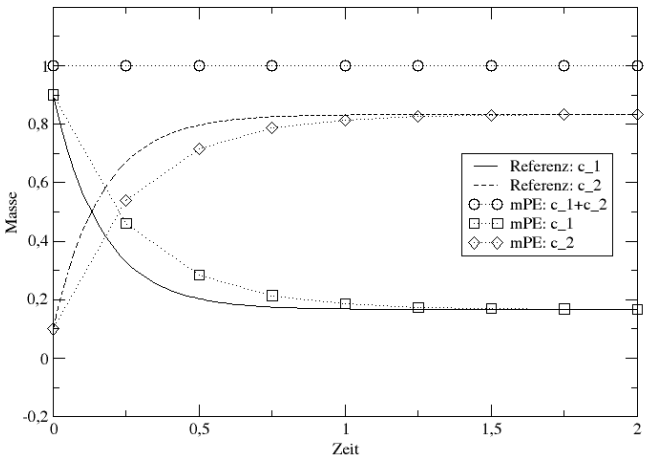

Abbildung 5.5: Das mPE Verfahren mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

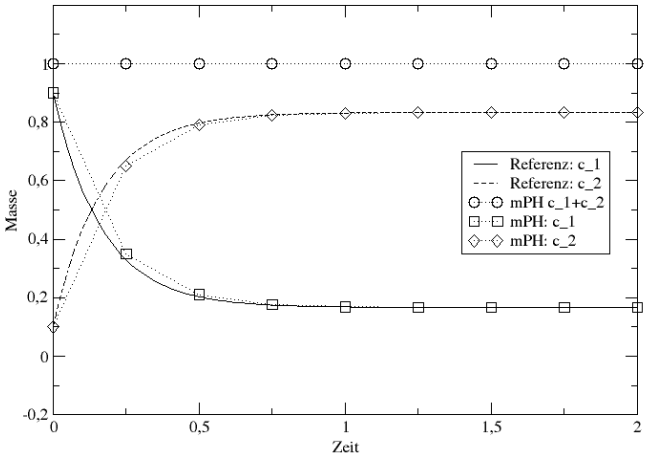

Abbildung 5.6: Das mPH Verfahren mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

zweiter Ordnung des BmH noch erkennbar besser sind.

Das mPE (Abbildung 5.5) und das mPH (Abbildung 5.6) liefern ebenso nur positive Ergebnisse. Allerdings dämpfen sie die Änderungen sichtlich stärker als die Bruggeman Verfahren und verzögern so den stattfindenden Austausch. Entsprechend liefern sie schlechtere Ergebnisse für dieses Problem.

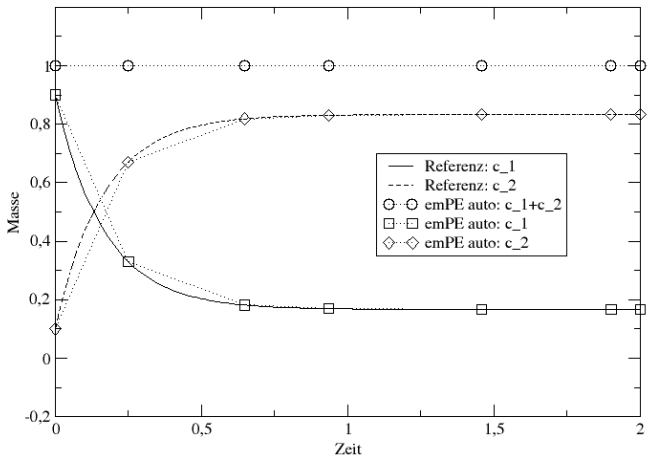

Abbildung 5.7: Das emPE Verfahren mit automatischer Schrittweitensteuerung und die analytische Referenzlösung für das lineare System.

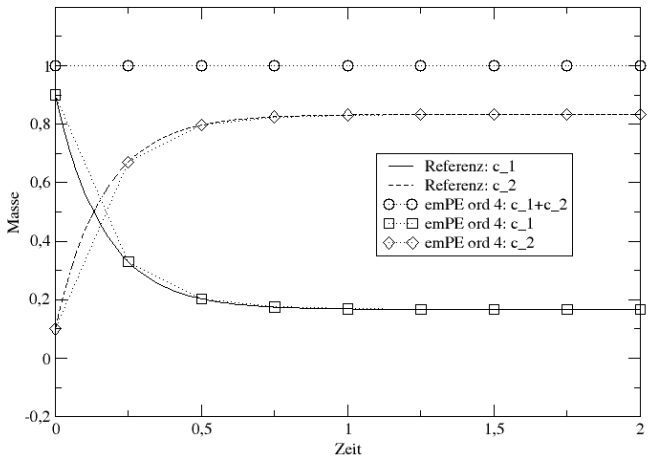

Abbildung 5.8: Das emPE Verfahren vierter Ordnung mit Schrittweite $h = 0.25$ und die analytische Referenzlösung für das lineare System.

Die beiden gezeigten emPE Verfahrensvarianten (emPE auto Abbildung 5.7 und emPE ord 4 Abbildung 5.8) zeigen diese charakteristische Verzögerung der mPE, mPH, BmE und BmH Verfahren nicht. Sie liefern sehr gute (und nur positive) Näherungen.

5.2 Vergleich der Verfahren und praktische Fehlerordnung

In der Abbildung 5.9 sind die Varianten aller vorgestellten Verfahren für gewöhnliche Differentialgleichungen charakterisiert. Jeder Punkt steht für eine vollständige Berechnung des linearen Problems (1.1) mit fester Schrittweite $h(i) = 2^i\, 10^{-4}$ für $i = 0, \ldots, 14$ und $h(15) = 2$. Zu jedem Rechendurchlauf wurde jeweils der Skalar

$$grgf = \frac{1}{N} \sum_{j=1}^{N} \sqrt{\frac{\sum_{k=1}^{D} \left[c_j\big(h(i)\cdot k\big) - c_j^k\right]^2}{\sum_{k=1}^{D} \left[c_j\big(h(i)\cdot k\big)\right]^2}} \tag{5.1}$$

mit Hilfe der berechneten und der analytischen Lösung bestimmt. Dabei ist $N = 2$ die Anzahl der Komponenten, $D \in \mathbb{N}, \frac{2}{h(i)} \leq D < \frac{2}{h(i)} + 1$ die Anzahl der verwendeten Rechenschritte und der Exponent 2 meint jeweils die zweite Potenz und steht in diesem seltenen Fall nicht für einen Zeitschritt. Dieser gemittelte relative Gesamtfehler (grgf) ist über der verwendeten Schrittweite aufgetragen.

Die Graphik liefert folgende Anschauung. Man erkennt die prognostizierte Ordnung aller Verfahren in wenigstens Teilen der untersuchten Schrittweitenbereiche.

Vor allem die extrapolierten mPE Verfahren hoher Ordnung, Ordnung acht und aufwärts, sind nur für große Schrittweiten in der Lage, also $h > 0.1$, die theoretisch zu erzielende Ordnung auch numerisch zu gewinnen. Das scheint für dieses Problem grundsätzlich nicht von Nachteil, da realistische Rechnungen sicherlich mit Schrittweiten in diesen Bereichen arbeiten. Allerdings ist der Aufwand enorm. Für eine Näherung 12. Ordnung benötigt man 78 $\left(= \sum_{j=1}^{12} j\right)$ mPE Verfahrensschritte. Wird eine solch hohe Genauigkeit in diesem Schrittweitenbereich benötigt, ist das Verfahren 12. Ordnung allerdings das effizienteste im Vergleich aller mPE Varianten.

Zu den Bruggeman Verfahren (BmE und BmH) lässt sich festhalten, dass sie im Bereich ihrer Ordnung, also das BmE verglichen mit dem Euler und dem mPE Verfahren und das BmH Verfahren verglichen mit dem mPH Verfahren, für dieses Problem die kleineren Fehler liefern.

Insgesamt ergibt sich folgendes Bild. Ist das Problem einfach genug (z.B. das lineare System), ist, trotz möglicher Positivitätsforderungen, die Anwendung der modifizierten Patankar Verfahren und deren Erweiterungen nur zu empfehlen, wenn gleichzeitig auch noch sehr hohe Anforderungen an die Größe der Zeitschrit-

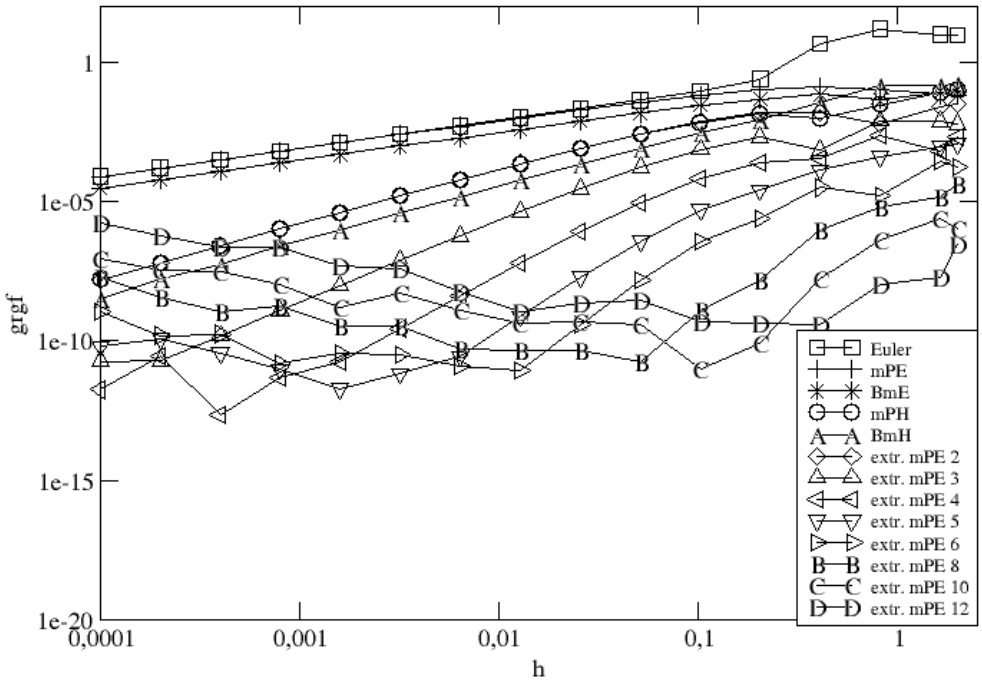

Abbildung 5.9: Die Ordnungsgraphik für alle ODE Löser berechnet mit Hilfe des linearen Problems (1.1) mit den Schrittweiten $h(i) = 2^i \cdot 10^{-4}$ für $i = 0, \ldots, 14$ und $h(15) = 2$.

te und die Güte der Näherungen vorliegt. Für geringe Ordnungsanforderungen sind die Bruggeman Verfahren zu empfehlen.

Für anspruchsvolle, steife Probleme (z.B. das Orego Problem oder der in [BDM03, Zar05] behandelte Robertson Testfall) sind sowohl die Bruggeman Verfahren als auch die hier nur kurz gestreiften expliziten Runge Kutta Verfahren ungeeignet. Diese lieferten in allen numerischen Tests des Autors keine Lösungen. Hier liegt die Stärke der modifizierten Patankar Verfahren und ihrer Varianten. Die Verfahren erster und zweiter Ordnung berechnen die Lösungen zwar mit einer zeitlichen Verzögerung, sie geben aber die Lösungen quantitativ sehr gut wieder.

Die extrapolierten Verfahren wirken dieser Verzögerung entgegen und zeigen sie in sehr viel geringerem Maße und erlauben zudem eine beliebig hohe Ordnung.

5.3 Numerische Bearbeitung des geobiologischen Systems (1.3) durch die Verfahren aus Kapitel 2

In diesem Unterkapitel werden die Näherungen verschiedener Verfahren für das System Beispiel 1.2 betrachtet. Dieses System (Beispiel 1.2) beschreibt die Aufnahme der Nährstoffe C und N durch Phytoplankton P und das Absterben des Phytoplanktons zu toter Materie D. Zu Beginn ist der beherrschende Prozess das Wachstum des Phytoplanktons. Dies dauert bis zur Erschöpfung der vorhandenen Rohstoffe C und N an. Nachdem der erste Rohstoff (bei der Wahl der Konstanten in diesem Fall C) erschöpft ist, wächst P nicht mehr und das Absterben des Phytoplanktons bleibt, solange noch P vorhanden ist, der aktive Prozess. Das Euler

Verfahren	Euler	Emini	BmE	BmH	vmPE
grgf	0.54	0.58	0.76	0.40	0.64

Tabelle 5.1: grgf (5.1) der Verfahren zu Problem (1.3)

Verfahren (Abbildung 5.10) liefert wiederum, wie in den Näherungen zum Beispiel 1.1 (Abbildung 5.1) negative Schätzungen. Ebenso kann man Verzögerungen der Nährstoffaufnahme erkennen. Das erreichte Maximum der Phytoplanktonmasse liegt deutlich über dem der Referenzlösung. Den Endzustand prognostiziert das Euler Verfahren aber korrekt. Insgesamt ergibt sich ein grgf von 0.54 (Tabelle 5.1).

Das Emini Verfahren (Abbildung 5.11) bietet ein vergleichbares Bild zum Euler Verfahren mit dem wesentlichen Unterschied, dass es keine negativen Näherungen bestimmt und einen etwas schlechteren grgf von 0.58 (Tabelle 5.1) aufzuweisen hat.

Die Bruggeman Verfahren (das BmE: Abbildung 5.12 und das BmH: Abbildung 5.13) liefern eine maximale Phytoplanktonmasse, die der Referenzlösung ähnlich ist. Die Verzögerung der Prozesse sieht man deutlich. Auch hier sieht man eine klare Verbesserung der Näherungen durch die Anwendung des Verfahrens zweiter Ordnung. Für diese beiden Verfahren beläuft sich der grgf auf 0.76 respektive 0.40 (Tabelle 5.1).

Das vmPE Verfahren (Abbildung 5.14) erbt die Eigenschaften des mPE Verfahrens. Es verzögert deutlich die Prozesse und liefert ein Phytoplanktonmaxi-

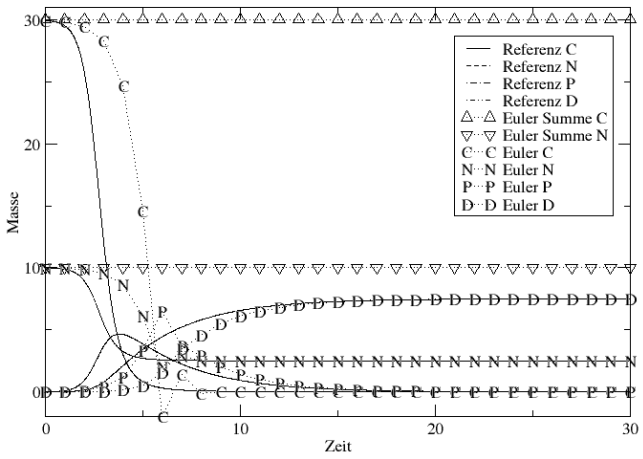

Abbildung 5.10: Das Euler-Verfahren mit Schrittweite $h = 1$ und die Referenzlösung für das geobiologische System (1.3).

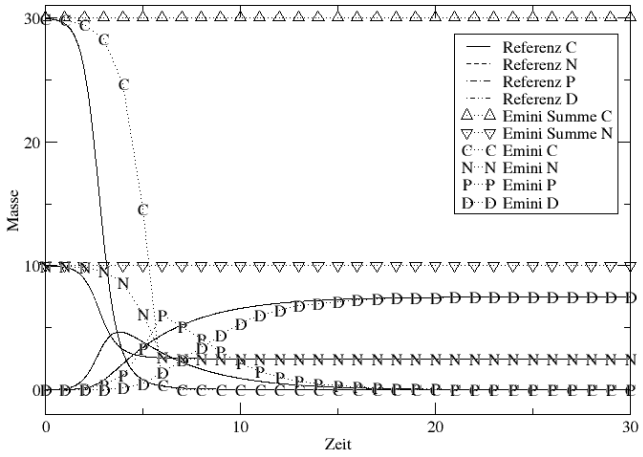

Abbildung 5.11: Das Emini Verfahren mit Schrittweite $h = 1$ und die Referenzlösung für das geobiologische System (1.3).

mum unter dem Niveau der Referenzlösung. Man sieht aber die Wahrung der Konservativität im Sinne der Definition 2.36, welches es vom mPE Verfahren unterscheidet. Ebenso liegt der grgf mit 0.64 (Tabelle 5.1) im Vergleich der Verfahren erster Ordnung zwischen dem der Euler bzw. Emini Verfahren einerseits und dem BmE Verfahren andererseits.

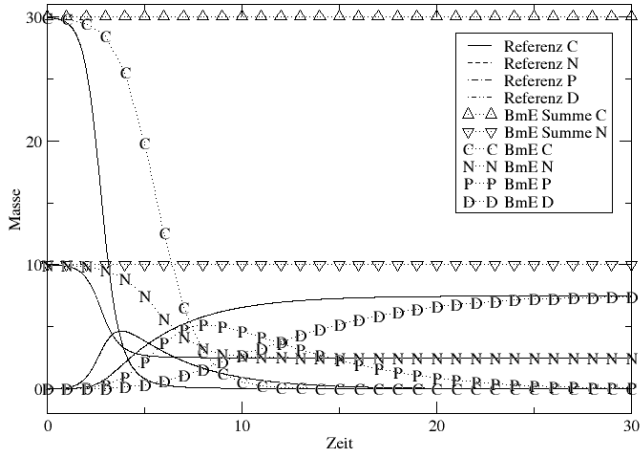

Abbildung 5.12: Das BmE Verfahren mit Schrittweite $h = 1$ und die Referenzlösung für das geobiologische System.

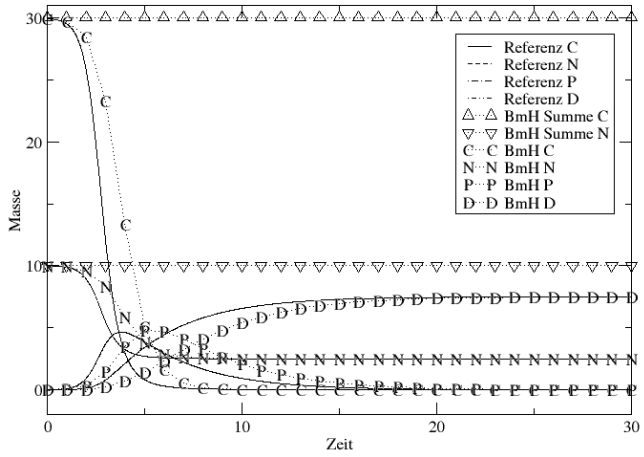

Abbildung 5.13: Das BmH Verfahren mit Schrittweite $h = 1$ und die Referenzlösung für das geobiologische System (1.3).

In der letzten Graphik zu diesem System (Abbildung 5.15) sind die Verfahren direkt miteinander für die Größe D verglichen. Man erkennt, dass das Euler Verfahren das Ende des Anstiegs am besten nachvollzieht. Bevor die Ressource C knapp wird ($t < 7$), zeigt das einzige hier präsentierte Verfahren zweiter Ordnung, das BmH, die beste Näherung. Zur Wahrung der Positivität (von C) aber dämpft

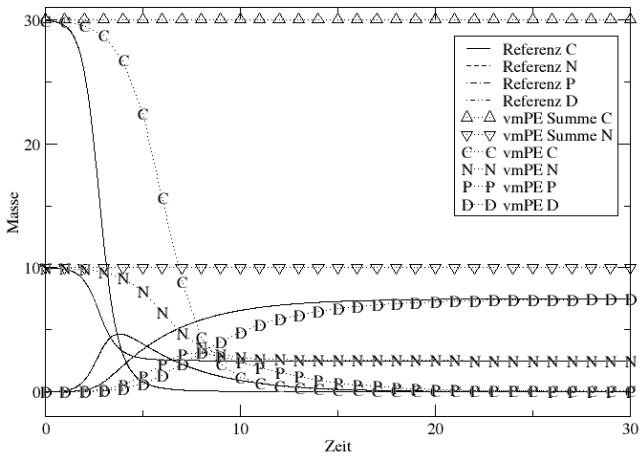

Abbildung 5.14: Das vmPE Verfahren mit Schrittweite $h = 1$ und die Referenzlösung für das geobiologische System (1.3).

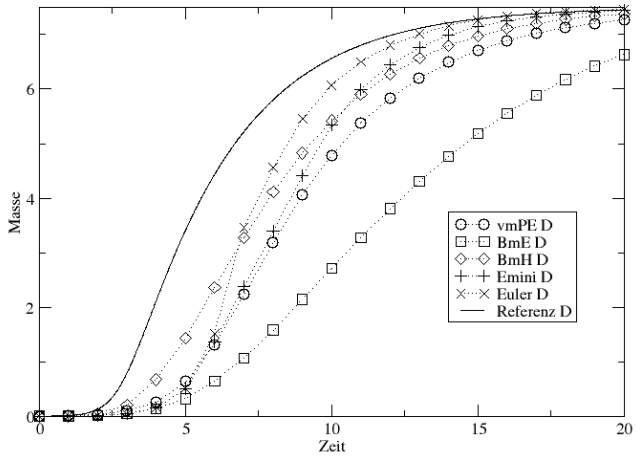

Abbildung 5.15: Ein Ausschnitt aller Verfahren mit Schrittweite $h = 1$ und der Referenzlösung für Komponente D des geobiologischen Systems (1.3).

es so stark, dass das Euler Verfahren die bessere Näherung liefert. Das Emini Verfahren ist vergleichbar gut. Das vmPE Verfahren ist ähnlich, aber schlechter als die anderen bereits besprochenen Verfahren. Das BmE Verfahren dämpft sehr viel stärker und liegt abgesehen von den Anfangs- und Endzuständen (nicht mehr im gewählten Ausschnitt zu sehen) weit von der Referenzlösung.

5.4 Numerische Bearbeitung des Orego Problems (1.4) durch die Verfahren aus Kapitel 2

Das Modell aus Beispiel 1.4 ist ein sehr steifes Differentialgleichungssystem und beschreibt eine zyklisch ablaufende chemische Reaktion. Abbildung 1.4 zeigt die Referenzlösung.

Das vmPE Verfahren ist grundsätzlich in der Lage, das Problem zu bewältigen. Es liefert einen grgf (5.1) von 0.34. In der Gesamtansicht, Abbildung 5.16, scheinen die Näherungen mit der festen Schrittweite $h = 0.015$ die Referenzlösung sehr gut zu treffen. Ein Zoom auf den zweiten starken Impuls (Abbildung 5.17) zeigt aber deutlich wieder die für modifizierte Patankar Verfahren typische Verzögerung der Reaktionen. Andererseits kann man aber ebenso die Ähnlichkeit der Kurven beobachten. Die Maxima und Steigungen sind vergleichbar.

Die Abbildung 5.18 zeigt die vierte Komponente des erweiterten Orego Problems. Sie ist nicht in der vorhergehenden Graphik enthalten, da sie negative Werte hat und diese eine Darstellung mit logarithmischer Skala unmöglich machen würden.

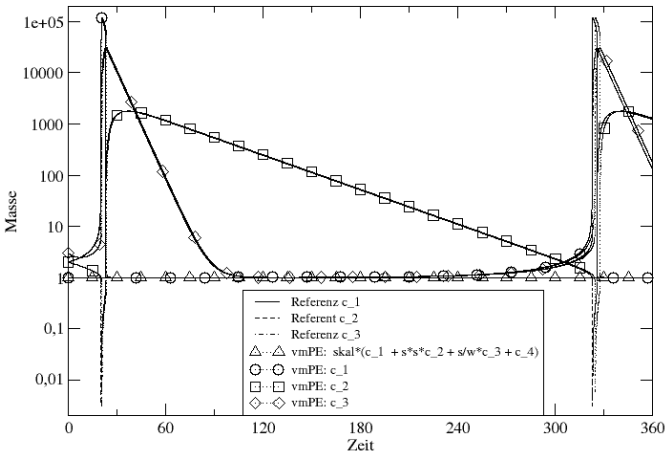

Abbildung 5.16: Das vmPE Verfahren mit Schrittweite $h = 0.015$ für die ersten drei Komponenten des Systems (1.4) und die Referenzlösung für das Orego Problem.

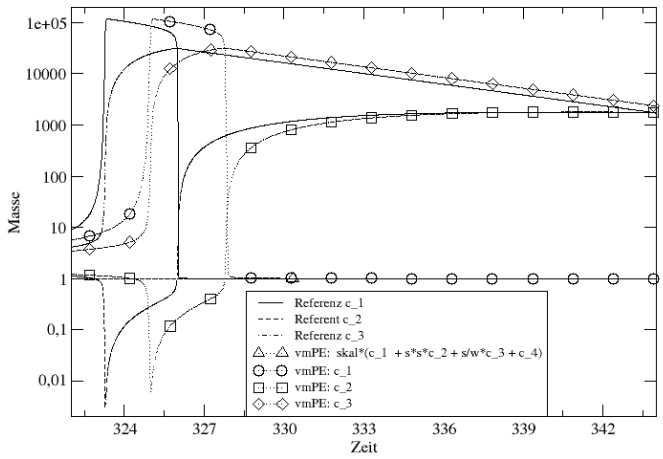

Abbildung 5.17: Ein Ausschnitt des vmPE Verfahrens mit Schrittweite $h = 0.015$ für die ersten drei Komponenten des Systems (1.4) und der Referenzlösung für das Orego Problem.

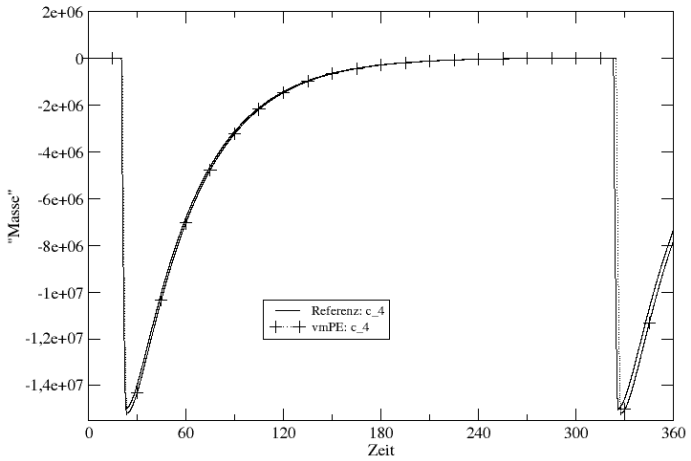

Abbildung 5.18: Das vmPE Verfahren mit Schrittweite $h = 0.015$ für die vierte Komponente des Systems (1.4) und die Referenzlösung für das Orego Problem.

Das BmH war mit einer Schrittweite von $h = 10^{-8}$, also deutlich kleiner als die Schrittweite der Referenzlösung, $h = 10^{-5}$, nicht in der Lage, das Problem zu bewältigen. Auch größere Schrittweiten brachten keinen Erfolg.

Kapitel 6

Ergebnisse für alle Testfälle aus Kapitel 3

6.1 Numerischer Vergleich des Phosphorzyklus (3.4) für Ecobas und die modifizierten Patankar Verfahren

Im Folgenden finden sich die Darstellungen von Näherungen zum Modell (3.4) des Phosphorkreislaufs ohne Strömung. Gezeigt sind die Berechnungen für einen Zeitraum von acht Jahren. Hier dargestellt sind nicht alle Größen der Gleichung (3.4). Vielmehr wurden zwei besonders interessante, PS und BA_A, ausgewählt.

Um den Fehler zu jeder Abbildung angeben zu können, wird hier noch ein vereinfachter Fehler definiert. Für den Fall, dass man nur eine Größe c_j betrachtet, wird aus dem grgf (5.1) unter Beibehaltung der dortigen Notation

$$rgf = \sqrt{\frac{\sum_{k=1}^{D} \left[c_j\big(h(i)\cdot k\big) - c_j^k\right]^2}{\sum_{k=1}^{D} \left[c_j\big(h(i)\cdot k\big)\right]^2}}. \tag{6.1}$$

Eine Zusammenfassung aller Fehler befindet sich in Tabelle 6.1.

Die ersten beiden Graphiken (Abbildung 6.1 und 6.2) zeigen die Näherungen des mPH Verfahrens. Sie sind mit einer sehr hohen Auflösung, $h = \frac{1}{64}$, gerechnet. Die Lösungen stimmen über weite Teile mit den Vergleichslösungen von Ecobas überein (rgf $<$0.036, vergleiche dazu Tabelle 6.1). Die einzige wahrnehmbare Differenz ist das erreichte Maximum der Algenbiomasse BA_A im dritten Jahr

Abbildung	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8
Verfahren	mPH	mPH	mPH	mPH	Emini	Emini	Emini	Emini
Größe	PS	BA_A	PS	BA_A	PS	BA_A	PS	BA_A
h	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{2}$	$\frac{1}{2}$
rgf	0.010	0.036	0.077	0.702	0.010	0.036	0.088	0.772

Tabelle 6.1: rgf (6.1) aller Abbildungen 6.1 bis 6.8

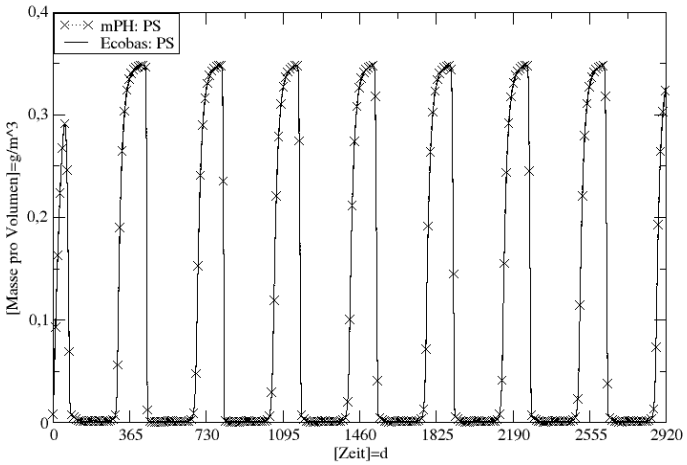

Abbildung 6.1: Das mPH Verfahren mit Schrittweite $h = \frac{1}{64}$ und die Ecobaslösung für die Größe PS des Systems (3.4).

($730 < t < 1095$). Die anderen Größen werden in gleicher zufriedenstellender Weise berechnet.

Die beiden nächsten Graphiken, Abbildung 6.3 und 6.4, liefern ein anderes Bild. Wiederum sind Ergebnisse des mPH Verfahrens gezeigt. Ebenfalls sieht man wieder die Näherungen für die Größen PS und BA_A. Hier wurde relativ grob aufgelöst, $h = \frac{1}{2}$. Die Größe PS wird in optimaler Übereinstimmung getroffen (rgf = 0.077), aber die Biomasse BA_A ist quantitativ sehr weit von der Vergleichslösung von Ecobas entfernt (rgf = 0.702). Insgesamt sind alle Biomassen der Algen, BA_A, BA_B, BA_C und BA_D stark abweichend von den Näherungen der hoch aufgelösten Rechnung bzw. Ecobas. Trotzdem ist die Masse des gelösten

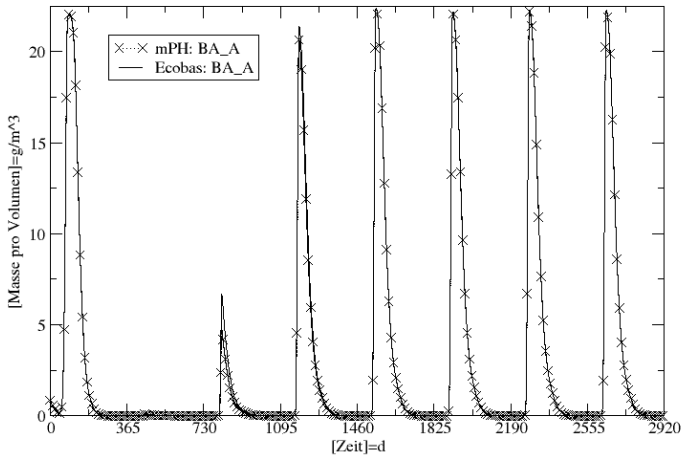

Abbildung 6.2: Das mPH Verfahren mit Schrittweite $h = \frac{1}{64}$ und die Ecobaslösung für die Größe BA_A des Systems (3.4).

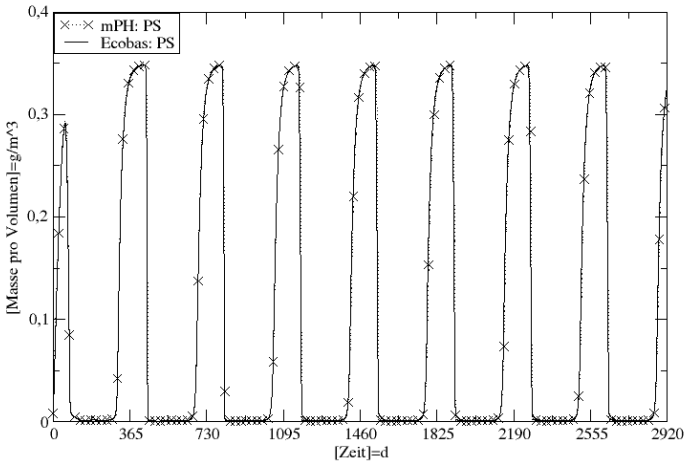

Abbildung 6.3: Das mPH Verfahren mit Schrittweite $h = \frac{1}{2}$ und die Ecobaslösung für die Größe PS des Systems (3.4).

Phosphors, PS, in allen drei Rechnungen übereinstimmend.

Die nächsten beiden Darstellungen (Abbildungen 6.5 und 6.6) zeigen den Vergleich des Emini Verfahrens mit der Schrittweite von $h = \frac{1}{64}$ und der von Ecobas

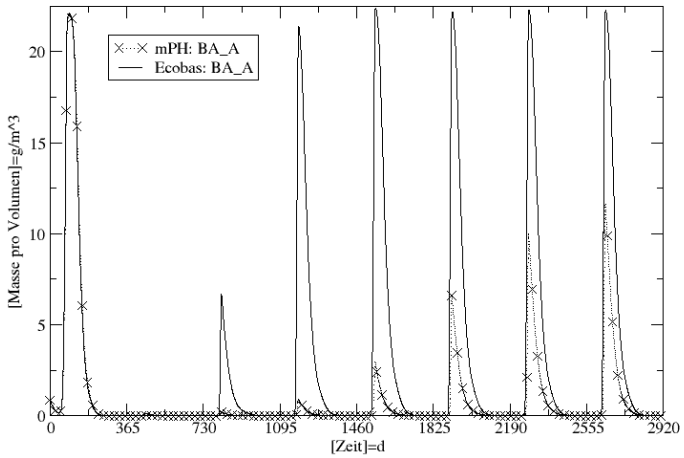

Abbildung 6.4: Das mPH Verfahren mit Schrittweite $h = \frac{1}{2}$ und die Ecobaslösung für die Größe BA_A des Systems (3.4).

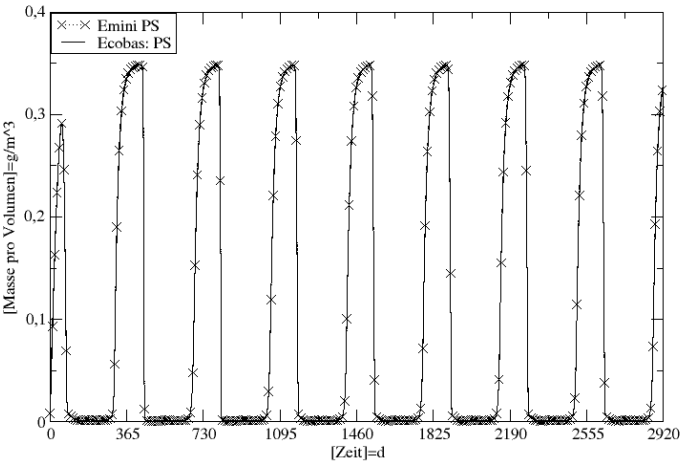

Abbildung 6.5: Das Emini Verfahren mit Schrittweite $h = \frac{1}{64}$ und die Ecobaslösung für die Größe PS des Systems (3.4).

bestimmten Näherungen. Die Abweichungen (rgf) sind identisch zu denen der Rechnung des mPH Verfahrens bei gleicher Schrittweite.

Abschließend, Abbildung 6.7 und 6.8, sind noch die Ergebnisse des Emini

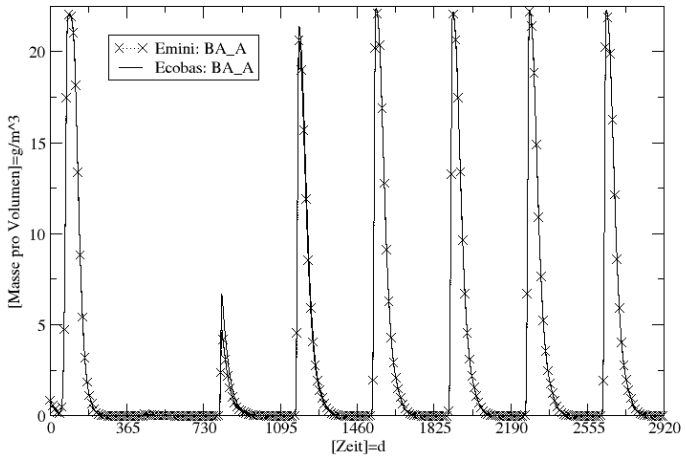

Abbildung 6.6: Das Emini Verfahren mit Schrittweite $h = \frac{1}{64}$ und die Ecobaslösung für die Größe BA_A des Systems (3.4).

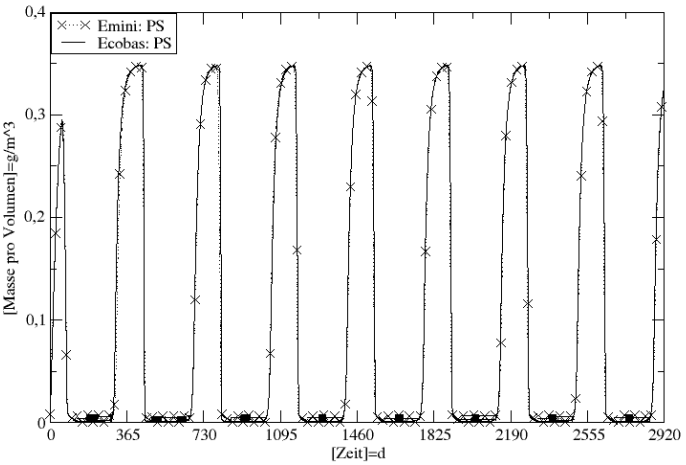

Abbildung 6.7: Das Emini Verfahren mit Schrittweite $h = \frac{1}{2}$ und die Ecobaslösung für die Größe PS des Systems (3.4).

Verfahrens mit ebenfalls sehr grober Schrittweite, $h = 0.5$, präsentiert. Auch hier werden die Maxima des gelösten Phosphors sehr gut getroffen. Im Bereich der Minima (den Sommermonaten) oszilliert die Näherung stark, was auf Grund der

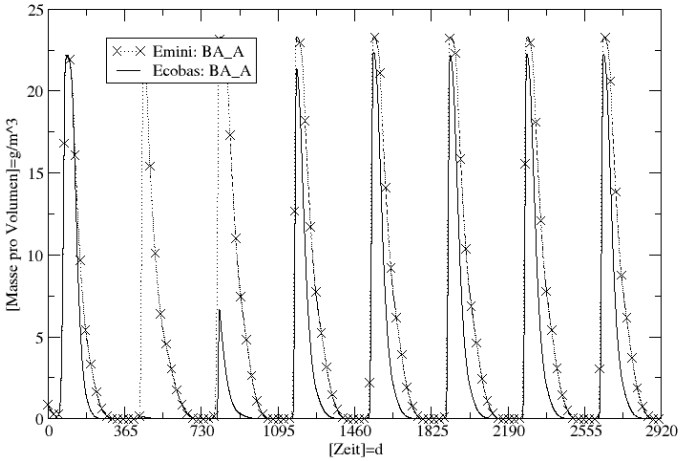

Abbildung 6.8: Das Emini Verfahren mit Schrittweite $h = \frac{1}{2}$ und die Ecobaslösung für die Größe BA_A des Systems (3.4).

Auflösung nur zu erahnen ist (rgf = 0.088). Die Ergebnisse für die Größe BA_A werden in entgegengesetzter Richtung verfehlt (rgf = 0.772). Die nahezu völlige Auslöschung der Algengruppe A, welche von allen anderen Rechnungen gezeigt wird, wird nicht abgebildet. Insgesamt ähneln die Prognosen für die letzten vier Jahre aber, unabhängig von zuvor gezeigten Unterschieden, den Näherungen von Ecobas.

Die Beobachtung deckt sich mit den Folgerungen aus dem Kapitel 5. Für große Zeitschritte bieten sich die modifizierten Patankar Verfahren an. Sie liefern nicht oszillierende, positive und konservative Näherungen. Allerdings wird dies bei vielen Beispielen mit einer Dämpfung im Vergleich zu Referenzlösungen erzielt. Außerdem verlangen sie das Lösen linearer Gleichungssysteme für jeden Berechnungsschritt und sind damit im Vergleich zu expliziten Verfahren in der Anwendung relativ aufwändig.

6.2 Das kreisförmige Dammbruchproblem

Im folgenden Abschnitt werden die Ergebnisse des in Kapitel 4 vorgestellten Strömunsglösers mit den Ergebnissen des Strömungslösers von Toro [Tor01b] zu dem kreisförmigen Dammbruchproblem aus Unterkapitel 3.1.2 verglichen.

Der eigene Strömungslöser wird in zwei Varianten verwendet. Sie unterscheiden sich lediglich in der Wahl des Limiters (dem Limiter nach Barth und Jesperson (LBJ) und dem nach Venkatakrishnan (LV), beide beschrieben in Unterkapitel 4.2.1). Ansonsten sind Beide höherer Ordnung und explizit. Eine Übersicht zu den Fehlern aller gezeigten Darstellungen befindet sich in Tabelle 6.2.

Abbildung	6.9		6.10		6.11		6.12	
Limiter	LV		LV		LBJ		LBJ	
Zeitpunkt in s	0.4		4.7		0.4		4.7	
Größe	H	vx	H	vx	H	vx	H	vx
rgf	0.049	0.060	0.028	0.738	0.038	0.086	0.024	0.087

Tabelle 6.2: rgf (6.1) zu den gezeigten Abbildungen

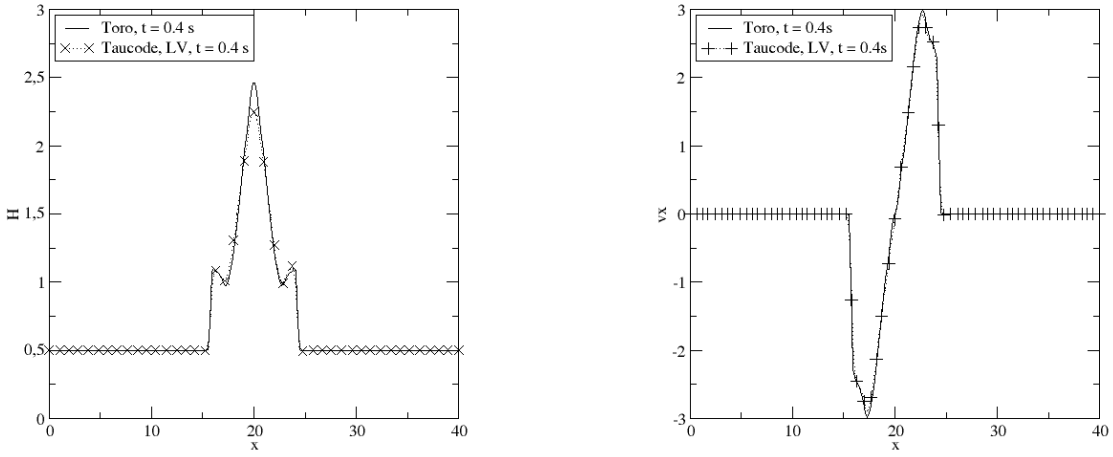

Abbildung 6.9: Näherung des Strömungslösers mit dem Limiter nach Venkatakrishnan für die Wasserhöhe H und die Geschwindigkeit v_1 des Dammbruchproblems zum Zeitpunkt $t = 0.4s$ und die Vergleichslösung nach [Tor01b].

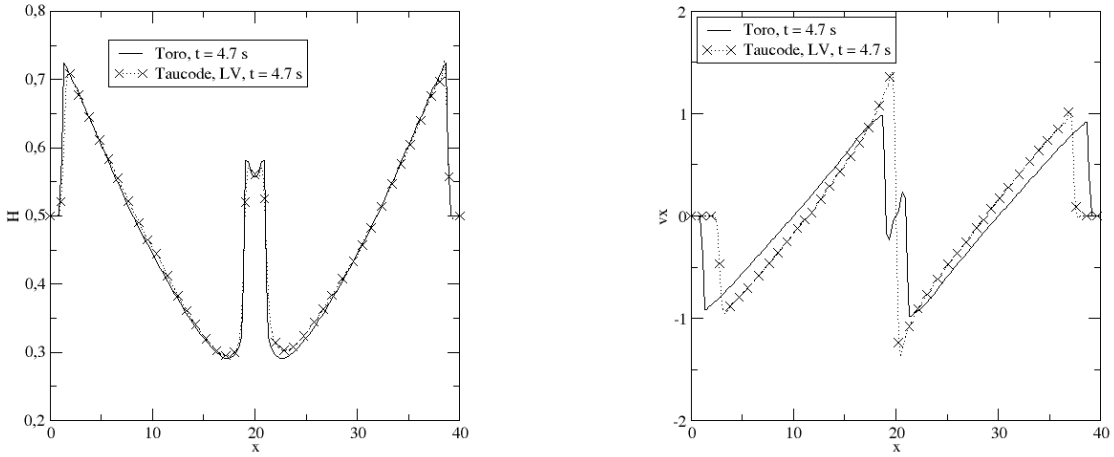

Abbildung 6.10: Näherung des Strömungslösers mit dem Limiter nach Venkatakrishnan für die Wasserhöhe H und die Geschwindigkeit v_1 des kreisförmigen Dammbruchproblems zum Zeitpunkt $t = 4.7s$ und die Vergleichslösung nach [Tor01b].

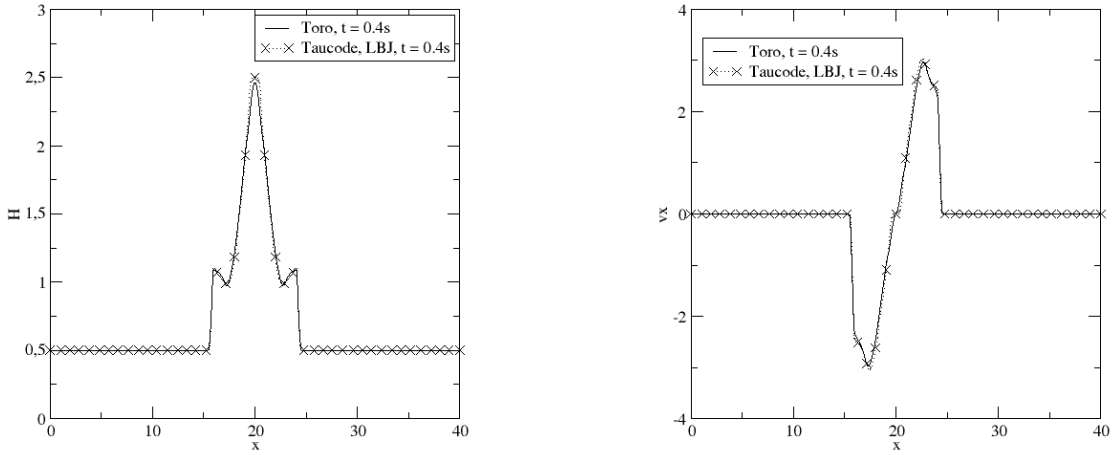

Abbildung 6.11: Näherung des Strömungslösers mit dem Limiter nach Barth und Jesperson für die Wasserhöhe H und die Geschwindigkeit v_1 des kreisförmigen Dammbruchproblems zum Zeitpunkt $t = 0.4s$ und die Vergleichslösung nach [Tor01b].

Die bestimmten Näherungen zum Zeitpunkt $t = 0.4s$ (Abbildung 6.9 und 6.11) zeigen beide eine gute Übereinstimmung mit der Vergleichslösung (jeweils

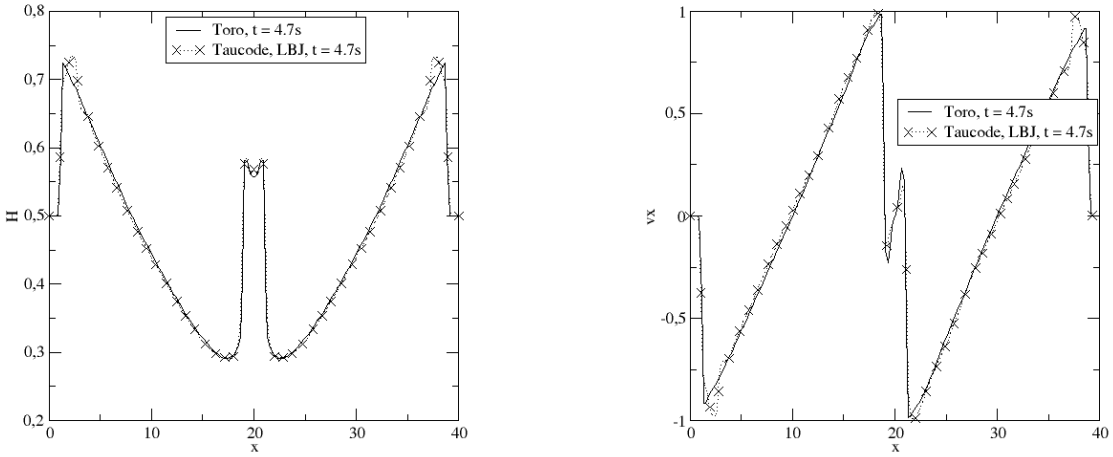

Abbildung 6.12: Näherung des Strömungslösers mit dem Limiter nach Barth und Jesperson für die Wasserhöhe H und die Geschwindigkeit v_1 des kreisförmigen Dammbruchproblems zum Zeitpunkt $t = 4.7s$ und die Vergleichslösung nach [Tor01b].

rgf<0.09), siehe Tabelle 6.2. Sie unterscheiden sich allerdings in der Höhe des Maximums der Wasserhöhe. Hier liegt die Variante mit dem LBJ über dem Maximum der Vergleichslösung nach Toro [Tor01b], wohingegen die durch den LV modifizierte Variante dieses Maximum unterschätzt. Die Variante des Taucodes benötigte 48 (im Wesentlichen äquidistante) Rechenschritte bis zum Zeitpunkt $0.4s$, wohingegen der Code nach Toro 20 Schritte benötigte.

Zum Zeitpunkt $t = 4.7s$ liegen die Näherungen sehr viel weiter auseinander (Abbildung 6.10 und 6.12). Beide Limiter liefern eine passable Übereinstimmung in der prognostizierten Wasserhöhe, wobei hier der LBJ die Näherungen liefert, welche näher an der Vergleichslösung sind (grf = 0.024 im Gegensatz zu 0.028).

Bei der Schätzung der Geschwindigkeiten liegen erhebliche Differenzen zwischen den Näherungen mit dem LV und der Vergleichslösung (grf = 0.738). Zu diesem Zeitpunkt liefern die Vergleichslösung und die Näherung mit dem LBJ (grf = 0.087) bereits eine zweite von innen nach außen laufende Welle; diese fehlt in der Näherung mit dem LV vollständig. Dafür ist die äußere Welle noch nicht soweit wie in der Vergleichslösung.

Es lässt sich festhalten, dass die Variante mit dem LBJ insgesamt einen sehr hohen Grad an Übereinstimmung zur Vergleichslösung liefert. Die den LV nutzende Variante liefert deutliche Differenzen. In der Prognose der Wasserhöhe liegen alle Verfahren sehr dicht beieinander.

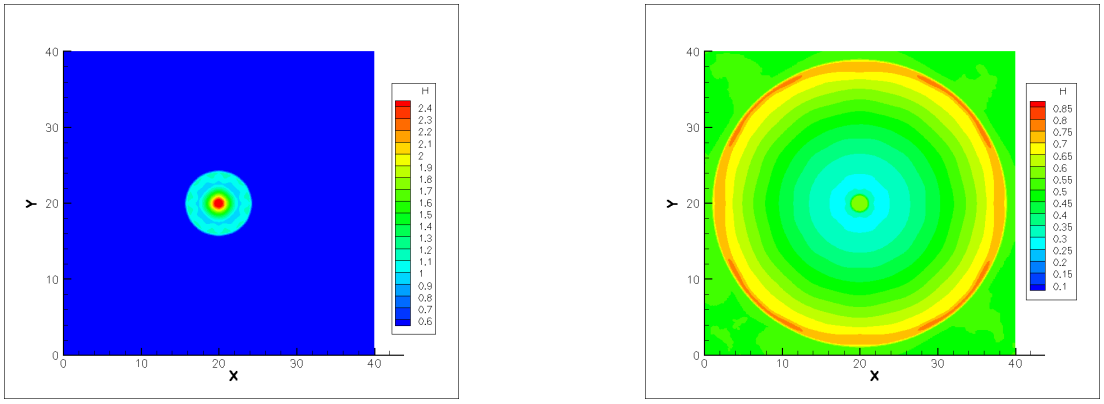

Abbildung 6.13: Näherung des Strömungslösers mit dem Limiter nach Barth und Jesperson für die Wasserhöhe H in Metern zum Zeitpunkt $t = 0.4s$ (links) und $t = 4.7s$ (rechts).

Abbildung 6.13 zeigt jeweils eine 2D Ansicht der Wasserhöhe. Es sind die mit dem LBJ berechneten Näherungen zum Zeitpunkt $t = 0.4s$ links und $t = 4.7s$ rechts. Man erkennt deutlich die erwartete kreissymmetrische Form der Wellen.

6.3 Anwendung des Gesamtverfahrens

Im nun folgenden Abschnitt werden die Ergebnisse des Gesamtverfahrens aus Unterkapitel 4.5 betrachtet. Da es keine Messdaten gibt, beschränkt sich der Vergleich auf die Lösungen von Ecobas für das System (3.4). Dieses System behandelt keinen Transport (weder advektiv noch diffusiv). Es können also nur strukturelle Anhaltspunkte aus dem Vergleich gewonnen werden.

Die beiden Graphiken (Abbildung 6.14 und 6.15) zeigen diesen Vergleich, d.h. sie zeigen einerseits die von Ecobas berechneten Verläufe der beiden bereits in Unterkapitel 6.1 untersuchten Größen PS und BA_A. Andererseits werden die Verläufe des Verfahrens aus Unterkapitel 4.5 (inkl. Strömung - advektiv wie diffusiv) zu jeweils zwei Zellen dargestellt, wobei hier nochmal in Erinnerung gerufen wird, dass Ecobas keinen Transport in Betracht zieht.

Die Zelle 1 liegt auf dem Einströmrand, die Box 150 in dem dem Ausströmrand zugewandten Arm des Kanals. Die Daten sind im Abstand von etwa 30 Tagen ohne weitere Zwischenwerte erfasst.

Betrachtet man die Abbildung 6.14, erkennt man große quantitative Unterschiede in den Näherungen des Gesamtverfahrens zu den Ecobas Näherungen.

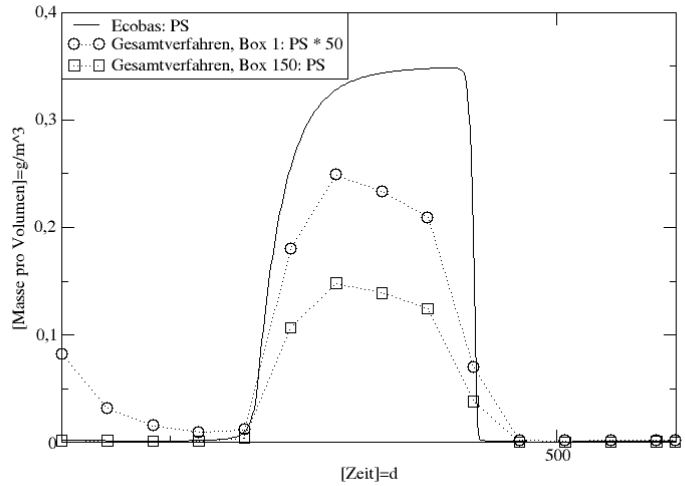

Abbildung 6.14: Vergleich zweier Boxen und der Ecobas Näherung für die Größe PS des Gesamtsystems (3.8).

Diese lassen sich mit dem Zufluss von Wasser ohne gelösten Phosphor erklären. Dieser Prozess ist nahe liegender Weise am Einströmrand, Zelle 1, deutlich ausgeprägter als bei Zelle 150. Daher wurde die Größe PS der Zelle 1 auch mit dem Faktor 50 vergrößert dargestellt. Das qualitative Verhalten der Kurven ist relativ ähnlich, womit die Dynamik des gesamten Systems auch bei geringeren Massekonzentrationen erhalten zu bleiben scheint.

Die Betrachtung der folgenden Graphik 6.15 liefert weniger leicht zu deutende Informationen. Festzuhalten bleibt zuerst, dass das qualitative Verhalten wiederum gut übereinstimmt zwischen den Näherungen von Ecobas und den Näherungen des Gesamtverfahrens.

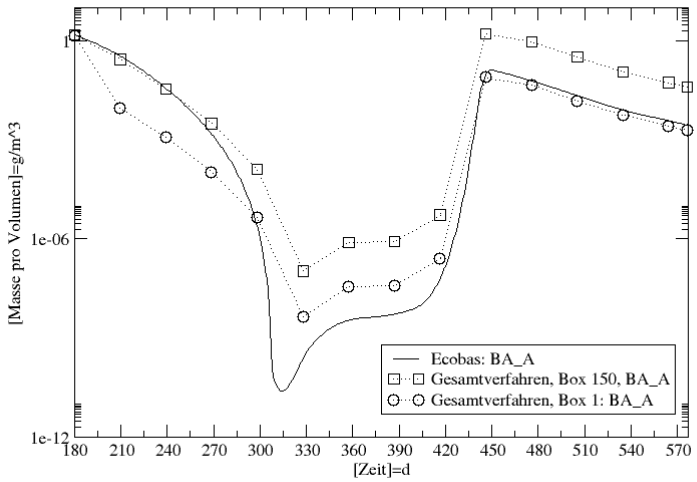

Abbildung 6.15: Vergleich zweier Boxen und der Ecobas Näherung für die Größe BA_A des Gesamtsystems (3.8).

Die absoluten Konzentrationen aber liefern eine Verkehrung des Arguments zur Erklärung der geringeren Konzentrationen des gelösten Phosphors. Die Berechneten Konzentrationen der Biomasse der Algengruppe A sind in den Näherungen des Gesamtverfahrens im Wesentlichen größer als die entsprechende Lösung von Ecobas. Dies lässt sich als Beleg für die Komplexität des Systems deuten.

Die Algengruppe A ist in der Vergleichsrechnung von Ecobas nahezu ausgelöscht. Dies geschieht auf Grund spezieller nicht periodischer Verhältnisse der anderen Systemgrößen; die Konzentration der Algengruppe A erholt sich in den

Rechnungen des reinen Phosphorsystems, vergleiche dazu die Ergebnisse aus Unterkapitel 6.1. Diese extreme Reduktion der Algengruppe A scheint so sensibel, dass durch die Verringerung der Konzentration durch das Ausspülens die Entwicklungsumstände für diese Algengruppe besser und nicht schlechter zu werden scheinen.

Die Abbildung 6.16 zeigt die auskonvergierten Absolutgeschwindigkeiten. Im Bereich des Sees findet keine große Bewegung statt. Die Strömung im Kanalzufluss und -abfluss liegt in der Größenordnung der Geschwindigkeit am Einströmrand (3.13). Dieser Zustand ist nach einem Monat berechnete Zeit erreicht. Die Strömung ändert sich anschließend nicht mehr. Dieses Szenario wurde gewählt, damit die berechneten Näherungen um nicht noch mehr unabschätzbare Unterschiede zu den Vergleichslösungen von Ecobas differieren.

Die beiden Graphiken in Abbildung 6.17 schlüsseln die Geschwindigkeiten in die beiden Richtungskomponenten auf. Dabei ist für die linke Darstellung (vx) die abgebildete obere Grenze der Skala, die Zehn, nicht das Maximum der im See berechneten Geschwindigkeiten. Vielmehr wurde dieser Maßstab gewählt, um die interessanten Teile der Strömung, das Gebiet im Einflussbereich des Sees differenziert auflösen zu können. Für die rechte Darstellung (vy) gilt mit dem gleichen Gedanken, dass die untere Grenze, die minus Zehn, nicht das Minimum ist, sondern auch hier der Differenzierungsbereich gewählt wurde, um das Gebiet im Seezulauf besser aufzulösen.

Es bleibt anzumerken, dass über die Nutzung des impliziten Strömungslösers eine Rechenzeitersparnis von 50% erzielt werden konnte. Dabei hatte der verwendete Zeitschritt die Form

$$h = 2 \cdot 10^3 \cdot \min\left(\frac{x_{min}}{|v| + \Phi}, \frac{\left(\frac{x_{min}}{2}\right)^2}{d}\right).$$

Vergleiche dazu auch (4.28) und (4.29).

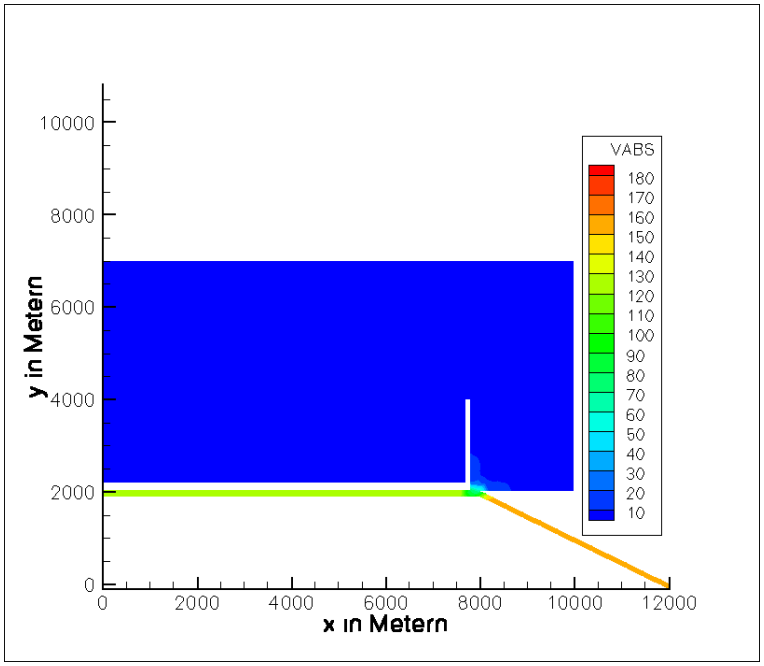

Abbildung 6.16: Die auskonvergierten Absolutgeschwindigkeiten in Metern pro Tag des Gesamtsystems (3.8).

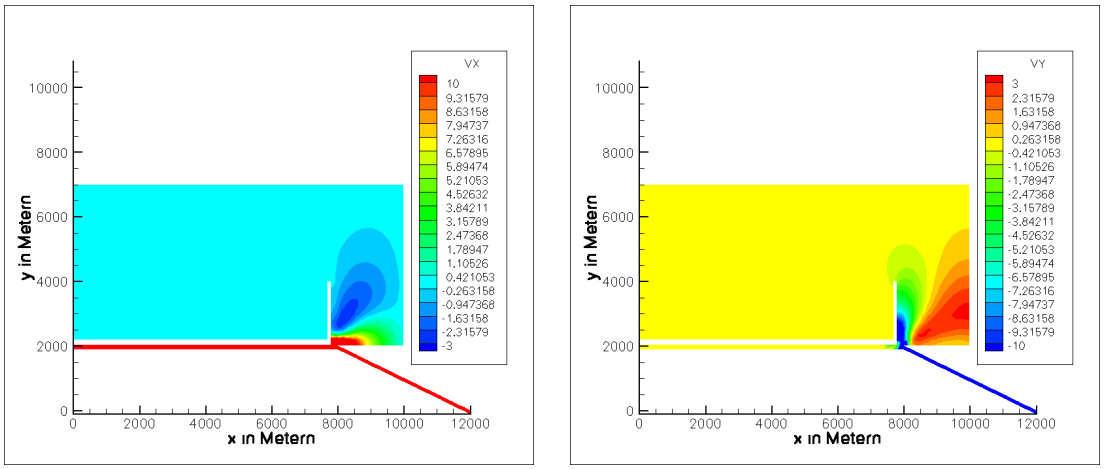

Abbildung 6.17: Die auskonvergierten Geschwindigkeiten in x und y Richtung in Metern pro Tag des Gesamtsystems (3.8).

Kapitel 7

Zusammenfassung und Ausblick

Im ersten Teil dieser Arbeit wurde eine vereinheitlichende Theorie für numerische Verfahren für gewöhnliche Differentialgleichungen im Kontext chemischer und biologischer Reaktionen und Prozesse vorgestellt. Dabei wurde besonderes Augenmerk auf Verfahren gelegt, die die in diesen Bereichen zentralen Begriffe Konservativität und Positivität von Systemen beachten und deren mathematischen Eigenschaften erhalten.

Die existierenden Verfahren, welche in unterschiedlicher Weise diesen Themenbereichen zuzuordnen sind, wurden beschrieben. Mit Hilfe der allgemeinen hier entwickelten Theorie, konnten inhärente Schwächen dieser Verfahren aufgezeigt werden, z.B. das Ordnungsmaximum von zwei, die Schwierigkeiten beim Lösen steifer Differentialgleichungen oder der Erhalt von Konservativität ausschließlich in sehr restriktiver Form.

Anschließend wurden Verfahrensverallgemeinerungen und Varianten entwickelt, die diese Mängel beheben: Die vorgestellten extrapolierten Verfahren sind in Abhängigkeit des betrachteten Problems von beliebiger Ordnung. Ebenso konnte eine Verallgemeinerung der bereits bekannten modifizierten Patankar Verfahren gezeigt werden, die die komplexe Form der Konservativität von Systemen erhält und steife Differentialgleichungen lösen kann. Für dieses Verfahren wurde an konkreten Beispielen die Wirksamkeit und Effektivität sowie die Erfüllung aller gewünschten Eigenschaften sowohl theoretisch als auch praktisch gezeigt.

Mögliche zukünftige Forschung könnten in der genaueren Untersuchung des verallgemeinerten modifizierten Patankar Verfahrens bestehen. Besonders hervorzuheben wäre dabei die Frage, ob sich das verallgemeinerte modifizierte Patankar Verfahren ebenso wie das modifizierte Patankar Euler Verfahren als Kernalgorith-

mus zur Extrapolation eignet. In gleicher Weise wäre auch die Frage interessant, welchen Eigenschaften ein System genügen muss, um die uneingeschränkte Positivität des verallgemeinerten modifizierten Patankar Verfahrens zu garantieren.

Im zweiten Abschnitt der Arbeit ging es um eine Untersuchung möglicher Schwierigkeiten und Effekte einer Erweiterung des bestehenden Modellierungswerkzeugs Ecobas auf spezielle Typen von partiellen Differentialgleichungen. Dabei wurde aber noch nicht Ecobas selbst erweitert, sondern in einem Grundlagenansatz, die in dieser Arbeit entwickelten Algorithmen auf ein Problem angewendet, welches in seiner Art der Formalisierung und Komplexität von einem erweiterten Ecobas sinnvoll angenähert werden können soll. Dafür wurde ein ökologisches Problem von relevanter Komplexität mit der bekannten zweidimensionalen Flachwassergleichung gekoppelt.

Zur Lösung wurde eine Variante des Taucodes weiterentwickelt und mit den modifizierten Patankar Verfahren aus dem ersten Abschnitt dieser Arbeit kombiniert. Die einzelnen Teile des Algorithmus wurden in einzelnen Tests auf ihre Tauglichkeit für diese Aufgabe untersucht. Weiterhin wurde ein Gesamttest präsentiert. Die vorgestellten Verfahren konnten die Testfälle ausnahmslos zufriedenstellend bewältigen. Damit wäre eine entsprechende Weiterentwicklung von Ecobas in einer Folgearbeit möglich.

Zum konkurrenzfähigen Einsatz wäre aber eine automatische Schrittweitensteuerung und Verfahrensauswahl für die gewöhnlichen Differentialgleichungslöser ein großer Schritt. Da die modifizierten Patankar Verfahren und deren Varianten in ihrem Einsatzfeld sehr stark sind (d.h. sehr robust, in der Extrapolationsvariante sehr genau usw.), dafür aber numerisch auch sehr aufwendig, gilt es für einen effizienten Einsatz, die Nutzung auf die notwendigen Stellen zu reduzieren. Daher sind die in dieser Arbeit gemachten Annahmen (zumeist feste Schrittweite, immer nur Einsatz eines Verfahrens für ein Problem bzw. eine Komponente usw.) für einen Einsatz in der Praxis nur begrenzt geeignet.

Ebenso lässt sich der implizite Strömungslöser weiter optimieren. Besonders das Lösen der linearen Gleichungssysteme mittlerer Dimension verbraucht einen Großteil der Rechenzeit. Hierfür wäre ein Löser vorstellbar, der z.B. Informationen über die Struktur der Matrizen ausnutzt.

Anhang A

Das ausformulierte ökologische Modell

Das vollständige System besteht aus den 14 gewöhnlichen Differentialgleichungen ($i = 1, 2, 3, 4$ respektive A, B, C, D)

$$
\begin{aligned}
\partial_t BA_i &= growth_i - res_i - sink_i - graz_i \\
\partial_t PA_i &= uptba_i - resp_i - setpa_i - gsinkp_i - assimp_i \\
\partial_t BZ &= \sum_{j=1}^{4} assim_j - zres - zmor \\
\partial_t PZ &= \sum_{j=1}^{4} assimp_j - zresp - zmorp \\
\partial_t PS &= exchp_{ps} - exchp_{pe_i} - \sum_{j=1}^{4} uptba_j + minpd + zresp + \sum_{j=1}^{4} resp_j \\
\partial_t PD &= -minpd - setpd + zmorp \\
\partial_t PE_I &= -exchp_{ps} + exchp_{pe_i} + minpe \\
\partial_t PE_O &= -minpe + setpd + \sum_{j=1}^{4} (setpa_j + gsinkp_j).
\end{aligned}
$$

Sämtliche verwendete Konstanten sind in den Tabellen A.1, A.2, A.3, A.4 und A.5 beschrieben. Die nummerischen Werte sind der Veröffentlichung [HJ02] entnommen.

Im Folgenden werden alle Funktionen der rechten Seite in Konstanten, Zustandsgrößen und Inputparameter (vom Anwender vorgegebene, möglicherweise

Konstante	Wert	Einheit	Beschreibung
$growth_{max_i}$	3.3, 5.97, 6.43, 10.37	$\frac{1}{d}$	Maximale Wachstumsrate von Algengruppe i
DP_{S_i}	0.027, 0.016, 0.016, 0.018	$\frac{g}{m^3}$	monod a zur Phosphoraufnahme von Algengruppe i
TOP_i	30, 23, 28, 21	C°	optimale Wachstumstemperatur von Algengruppe i
TP_i	4,12,9.3,19	-	steele Modifikator von Algengruppe i
LOP_i	310, 340, 350, 310	Lux	Optimale Wachstumslichtintensität von Algengruppe i

Tabelle A.1: Beschreibung aller Konstanten, nummerischen Werte und Einheiten - Teil 1

variable Größen, Wassertemperatur T, Sedimenttemperatur TE und Lichtintensität I) aufgeschlüsselt. Für die gezeigten Rechnungen wurden Lufttemperatur- und Niederschlagsdaten von [Wea08c, Wea08b, Wea08a] verwendet. Lösen einer eindimensionalen Wärmeleitungsgleichung liefert die Temperaturverteilung wie in der Abbildung A.1 gezeigt. Ebenso ist dort die Lichtintensität angegeben. Es wird die Näherung

$$I = 268 + \frac{783 - 268}{2}\left(1 - \cos\left(\frac{2\pi}{365}(t + 10)\right)\right)$$

mit dem Lichtintensitätsmaximum, 768 Lux, und -minimum, 268 Lux, verwendet. Die Zahlen stammen ebenso wie die anderen Konstanten aus der Arbeit [HJ02]. Zuerst wird die Aufsplittung der Terme zur Dynamik der Biomassen der Algen beschrieben.

- Der Zuwachs der Biomasse einer Algengruppe i ist festgelegt durch ihre maximale Wachstumsrate $growth_{max_i}$, die limitierenden Faktoren für Temperatur, Licht und verfügbaren Phosphor

$$growth_i = growth_{max_i}\, TT_i\, PP_i\, LL_i\, BA_i.$$

  - Die Limitierung des Algenwachstums durch den verfügbaren Phosphor wird beschrieben nach Monod (3.5) durch

$$PP_i = \frac{PS}{DP_{s_i} + PS}.$$

  - Die Auswirkung der Temperatur und der „gemittelten“ Lichtintensität auf das Algenwachstum wird nach Steele (3.6) modelliert durch

$$TT_i = \left(\frac{T}{TOP_i}\, e^{\frac{TOP_i - T}{TOP_i}}\right)^{TP_i} \text{ und } LL_i = \frac{L}{LOP_i}\, e^{\frac{LOP_i - L}{LOP_i}}.$$

Konstante	Wert	Einheit	Beschreibung
K_1	1.5	$\frac{1}{d}$	Extinktionskoeffizient des Wassers
K_2	1	$\frac{1}{d}$	Extinktionskoeffizient der Algen
RO_i	0.007, 0.003, 0.003, 0.003	$\frac{1}{d}$	Stoffwechselverlustfaktor von Algengruppe i
$TCOEF$	0.038	$\frac{1}{°C}$	Q10 Koeffizient; Temperaturvergleichskoeffizient
K_{set_i}	0.035, 0.052, 0.048, 0.073	$\frac{m}{d}$	Sedimentierungsgeschwindigkeit von Algengruppe i
GR_{max}	0.09	$\frac{1}{d}$	Maximale Grazingrate für Zooplankton
R_i	0.8, 0.45, 0.3, 0.5	-	Präferenzfaktor des Grazings von Algengruppe i

Tabelle A.2: Beschreibung aller Konstanten, nummerischen Werte und Einheiten - Teil 2

  - Die mittlere Lichtintensität, bestimmt über die Wassertiefe H und die Extinktion K, wird berechnet unter Verwendung der Funktion hyprep (3.7) durch
$$L = I \left(\frac{1 - e^{-K\,H}}{K\,H} \right).$$
  - Die Extinktion wird abhängig von der Wassertrübung und den Algenmassen und deren Beschattungseffekt bestimmt
$$K = K_1 + K_2 \sum_j BA_j.$$

- Die Biomasseverluste einer Algengruppe durch Stoffwechselprozesse sind modelliert als
$$res_i = RO_i\, e^{TCOEF\,T}\, BA_i.$$

- Die Sedimentation einer Algengruppe und der damit einhergehende Verlust an Biomasse wird quantifiziert durch
$$sink_i = BA_i\, \frac{K_{set_i}}{H}.$$

- Die Reduktion einer Algengruppe durch Zooplanktongrazing wird berücksichtigt durch
$$graz_i = GR_{max}\, FF \frac{R_i\, BA_i}{F} BZ.$$

Konstante	Wert	Einheit	Beschreibung
F_{min}	0.05	$\frac{g}{m^3}$	minimale Nahrungskonzentration für Zooplankton
F_s	0.25	$\frac{g}{m^3}$	monod a für die Zooplanktonnahrungsaufnahme
$util_i$	0.8, 0.8, 0.8, 0.8	-	Effizienzkoeffizient der jeweiligen Algenaufnahme
ZRO	0.02	$\frac{1}{d}$	Stoffwechselrate, Masseerhalt von Zooplankton
ZRM	0.04	-	Stoffwechselrate, Verdauung von Zooplankton

Tabelle A.3: Beschreibung aller Konstanten, nummerischen Werte und Einheiten - Teil 3

  - Dabei spielen die Präferenz und die verfügbaren Biomassen aller Algenarten eine wesentliche Gewichtungsrolle

$$F = \sum_j R_j \; BA_j.$$

  - Die Gesamtverfügbarkeit von Algenarten als Nahrung für das Zooplankton gehen in einer Formulierung nach Monod (3.5) ein

$$FF = \begin{cases} 0.0 & F_{min} > F \\ \frac{F - F_{min}}{F_s + F - F_{min}} & F_{min} \leq F \end{cases}.$$

Jetzt wird die Dynamik der Biomasse des Zooplanktons präzisiert.

- Der Biomassenzuwachs für Zooplankton richtet sich nach den entsprechenden Verlusten der Algen $graz_i$ reduziert um einen Effizienzfaktor

$$assim_i = util_i \; graz_i.$$

- Die Verluste der Biomassen der Algen, die durch Grazing entstehen, aber auf Grund von ineffizienter Verwertung nicht durch Zooplankton aufgenommen werden, werden als fester Bestandteil quantifiziert

$$gsink_i = (1. - util_i) \; graz_i,$$

  sie werden aber nur in der Phosphorbilanz weiter verwendet.

- Der Stoffwechselverlust an Biomasse des Zooplanktons ist die Summe aus den Beiträgen zum Erhalt der bestehenden Biomasse und der verrichteten Verdauungsarbeit

$$zres = ZRO \; e^{TCOEF\,T} \; BZ + ZRM \; \sum_j graz_j.$$

Konstante	Wert	Einheit	Beschreibung
mor	0.005	$\frac{1}{d}$	Sterberate von Zooplankton
K_{ex}	0.03	$\frac{1}{d}$	Diffusionskoeffizient des gelösten Phosphors, PS und PE_I
V_s	0.125	$\frac{m}{d}$	Sedimentationsrate von PD

Tabelle A.4: Beschreibung aller Konstanten, nummerischen Werte und Einheiten - Teil 4

- Die Sedimentation reduziert die Biomasse des Zooplanktons durch

  $$zmor = mor\ BZ,$$

  wobei an dieser Stelle die Sedimentationsrate mor höher angesetzt wird als aus Beobachtungen hervorgeht, um die Nichtberücksichtigung höherer trophischer Ebenen in der Modellierung auszugleichen, siehe [HJ02].

Die Funktionen, welche die nicht an Organismen gekoppelten Phosphorfraktionen miteinander in Verbindung setzen, werden nun beschrieben.

- Der diffusive Austausch der gelösten Phosphorfraktionen PS und PE_I wird beschrieben durch

  $$exchp = K_{ex}\ (PE_I - PS).$$

  In dieser Form lässt sich die Funktion $exchp$ aber nicht im Kontext der modifizierten Patankar Verfahren verwenden. Hier ist es nötig, für jeden Term das Vorzeichen zu kennen. Daher wird die Aufschlüsselung

  $$exchp = exchp_{PS} - exchp_{PE_I}$$

  mit den Definitionen

  $$exchp_{PS} = \begin{cases} exchp & , \quad PS \leq PE_I \\ 0 & , \quad \text{sonst} \end{cases} \qquad \text{und}$$

  $$exchp_{PE_I} = \begin{cases} -exchp & , \quad PS > PE_I \\ 0 & , \quad \text{sonst} \end{cases}$$

  verwendet.

- Die Sedimentation von PD beschreibt

  $$setpd = \frac{PD}{H}\ V_s.$$

Konstante	Wert	Einheit	Beschreibung
K_{m1}	0.01	$\frac{1}{d}$	Mineralisierungsrate von PD
S_{m1}	0.8	-	Temperaturkoeffizient der Mineralisierung von PD
K_{m2}	0.178	$\frac{1}{d}$	Mineralisierungsrate von PE_O
S_{m2}	1.08	-	Temperaturkoeffizient der Mineralisierung von PE_O
PUP_{max_i}	0.07, 0.1, 0.07, 0.07	$\frac{1}{d}$	Maximale Phosphoraufnahmerate von Algengruppe i
Pin_{min_i}	0.005, 0.005, 0.005, 0.005	$\frac{g}{m^3}$	Minimale Phosphorkonzentration von Algengruppe i
Pin_{max_i}	0.015, 0.020, 0.015, 0.015	$\frac{g}{m^3}$	Maximale Phosphorkonzentration von Algengruppe i

Tabelle A.5: Beschreibung aller Konstanten, nummerischen Werte und Einheiten - Teil 5

- Die Mineralisierung des organisch gebundenen Phosphors im Wasser respektive im Sediment ist modelliert durch

$$minpd = K_{m1} S_{m1}^{T-20}\ PD \text{ bzw. } minpe = K_{m2} S_{m2}^{TE-20}\ PE_O.$$

Als letzter Komplex bleiben noch die Funktionen, die die Zusammenhänge zwischen den Phosphorfraktionen der Algen bzw. des Zooplanktons und den nicht an Organismen gekoppelten Phosphorfraktionen beschreiben.

- Die Phosphoraufnahme durch Algengruppe i wird beschrieben durch

$$uptba_i = PUP_{max_i}\ PC_i\ BA_i\ PP_i.$$

  - Die Aufnahme wird gehemmt durch bereits in den Zellen enthaltenen Phosphor in der Form nach Monod (3.5)

$$PC_i = \begin{cases} 1 & PCON_i < Pin_{min_i} \\ 0 & PCON_i > Pin_{max_i} \\ \frac{Pin_{max_i} - PCON_i}{Pin_{max_i} - Pin_{min_i}} & sonst \end{cases} \quad \text{mit } PCON_i = \frac{PA_i}{BA_i}.$$

- Die Stoffwechselverluste des Phosphors von Algengruppe i sind direkt an die entsprechenden Stoffwechselverluste der Biomasse gekoppelt ebenso auch die Verluste durch Sedimentation

$$resp_i = res_i\ PCON_i \text{ und } setpa_i = sink_i\ PCON_i.$$

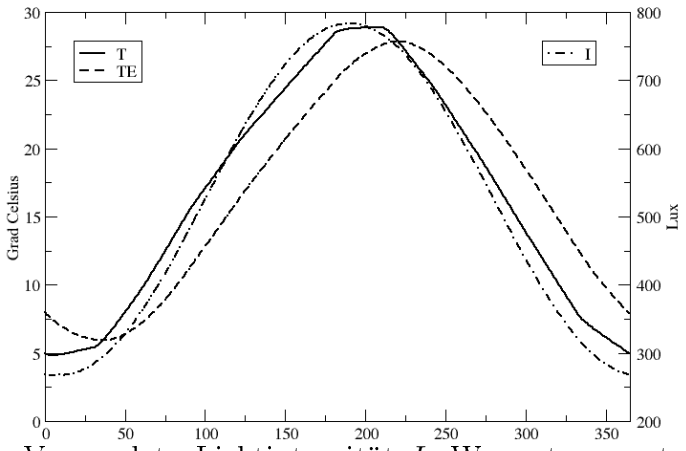

Abbildung A.1: Verwendete Lichtintensität I, Wassertemperatur T und Sedimenttemperatur TE

- Der Phosphor, der durch Grazing den Algen entzogen wird, wird zum einen Teil über den Effizienzfaktor $util_i$ gesteuert, der Phosphormasse des Zooplanktons und zum anderen Teil direkt dem organisch gebundenen Phosphor im Sediment via

$$gsinkp_i = gsink_i \; PCON_i \text{ und } assimp_i = assim_i \; PCON_i$$

  zugeführt.

- Ebenso wie im Falle der Stoffwechselverluste und der Sedimentation der Algen wird auch für das Zooplankton ein dem aktuellen Verhältnis zwischen Bio- und korrespondierender Phosphormasse angepasster Verlust verwendet

$$zresp = zres \; PCON \text{ und } zmorp = zmor \; PCON$$

$$\text{mit } PCON = \frac{PZ}{BZ}.$$

Literaturverzeichnis

[AGT95] Aftosmis, M., D. Gaitonde und T. S. Tavares: *Behavior of Linear Reconstruction Techniques on Unstructured Meshes.* AIAA Journal, 33(11):2038–2049, 1995.

[AO03] Arheimer, B. und J. Olsson: *Integration and Coupling of Hydrological Models with Water Quality Models.* Technischer Bericht 75, Swedish Meteorological and Hydrological Institute (SMHI), 2003.

[AWB+96] Asshoff, M., W. Windhorst, J. Benz, M. Knorrenschild, R. Lenz und O. Springstubbe: *Informationssysteme für ökologische Modelle: Konzepte zur Dokumentation und Verwaltung.* EcoSys, 4:297–316, 1996.

[BBKS07] Bruggeman, J., H. Burchard, B. W. Kooi und B. Sommeijer: *A second-order, unconditionally positive, mass-conserving integration scheme for biochemical systems.* Applied Numerical Mathematics, 57:36–58, 2007.

[BD83] Bader, G. und P. Deuflhard: *A Semi-Implicit Mid-Point Rule for Stiff Systems of Equations.* Numerische Mathematik, 41:373–399, 1983.

[BDM03] Burchard, H., E. Deleersnijder und A. Meister: *A high-order conservative Patankar-type discretisation for stiff systems of production-destruction equations.* Applied Numerical Mathematics, 47:1–30, 2003.

[Ben94] Benz, J.: *ECOBAS - Dokumentation mathematischer Beschreibungen ökologischer Prozesse.* In: Keller, H.B., R. Grützner und J. Benz (Herausgeber): *3. Treffen des Ak 5 "Werkzeuge für Simulation und Modellbildung in Umweltanwendungen". 28.10.- 29.10.93 in Kassel/Witzenhausen. Berichte des Kernforschungszentrums, KfK 5310*, Seiten 55 – 63. Gesellschaft für Informatik, 1994.

[Ben09] Benz, J.: *Ecobas Homepage*. http://ecobas.org/, 10.10.2009.

[BGH99] Benz, J., T. Gabele und R. Hoch: *Standardization of Model Documentation. Part II.* ECOMOD, Jun. 99:1 – 11, 1999.

[BH07] Bryhn, A. C. und L. Hakonson: *A Comparison of Predictive Phosphorus Load-Concentration Models for Lakes*. Ecosystems, 10:1084–1099, 2007.

[BHG98] Benz, J., R. Hoch und T. Gabele: *Standardization of Model Documentation*. ECOMOD, Sep. 98:19 – 20, 1998.

[BHL01] Benz, J., R. Hoch und T. Legovic: *ECOBAS - modelling and documentation*. Ecological Modelling, 138(1-3):3 – 15, 2001.

[Bir05] Birken, P.: *Numerical Simulation of Flows at Low Mach Number with Heat Source*. Dissertation, Universität Kassel, 2005.

[BJ89] Barth, T. J. und D. Jesperson: *The design and application of upwind schemes on unstructured meshes*. AIAA, 89-0366, 1989.

[Boe99] Boegman, L.: *Application of a Two-Dimensional Hydrodynamic and Water Quality Model to Lake Erie*. Diplomarbeit, University of Toronto, 1999.

[BRBM07] Broekhuizen, N., G. J. Rickard, J. Bruggeman und A. Meister: *An improved and generalized second order, unconditionally positive, mass conserving integration scheme for biochemical systems*. Applied Numerical Mathematics, Seite doi:10.1016/j.apnum.2006.12.002, 2007.

[Deu83] Deuflhard, P.: *Order and Stepsize Control in Extrapolation Methods*. Numerische Mathematik, 41:399–422, 1983.

[Deu85] Deuflhard, P.: *Recent Progress in Extrapolation Methods for Ordinary Differential Equations*. SIAM Review, 27(4):505–535, 1985.

[DP04] Dahl, M. und B. C. Pers: *Comparison of four models simulating phosphorus dynamics in Lake Vänern, Sweden*. Hydrology and Earth System Sciences, 8(6):1153–1163, 2004, ISSN 1027-5606.

[Fou09] Foundation, National Science: *LAPACK Homepage*. http://www.netlib.org/lapack/, 10.08.2009.

[GR96] Godlewski, E. und P. A. Raviart: *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Springer New York, 1996.

[Gra65] Gragg, W. B.: *On Extrapolation Algorithms for Ordinary Initial Value Problems*. SIAM Journal on Numerical Analysis, Serie B 2(3):384–403, 1965.

[Gra02] Gray, C. R.: *An Analysis of the Belousov-Zhabotinskii Reaction*. Undergraduate Math Journal, 3, 2002.

[Hen62] Henrici, P.: *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons, 1962.

[HGB98] Hoch, R., T. Gabele und J. Benz: *Towards a standard for documentation of mathematical models*. Ecological Modelling, 113:3 – 12, 1998.

[HJ02] Hongping, P. und M. Jianyi: *Study on the algal dynamic model for West Lake, Hangzhou*. Ecological Modelling, 148:67 – 77, 2002, ISSN 0304-3800.

[HNW91] Hairer, E., S. P. Nørsett und G. Wanner: *Solving Ordinary Differential Equations I*. Springer-Verlag, 1991.

[HP09] Hindmarsh, A. und L. Petzhold: *ODEPACK Homepage*. https://computation.llnl.gov/casc/odepack/odepack_home.html, 10.08.2009.

[HWN02] Hairer, E., G. Wanner und S. P. Nørsett: *Solving Ordinary Differential Equations II*. Springer-Verlag, 2002.

[JA90] Janse, J.H. und T. Aldenberg: *PCLoos: A eutrophication model of the Loosdrecht Lakes*. Technischer Bericht no. 714502001, National Institute of Public Health and Environmental Protection, P.O. Box 1, 3720 Bilthoven, January 1990.

[Jan05] Janse, J.H.: *Model studies on the eutrophication of shallow lakes and ditches*. Dissertation, Wageningen Universiteit, 2005.

[Lev02] Leveque, R. J.: *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[Lev06] Leveque, R. J.: *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, 2006.

[Lud99] Ludwig, M.: *Die numerische Simulation der Temperatur in Früh- und Neugeborenen*. Diplomarbeit, Universität Hamburg, 1999.

[MBMP06] Malmaeus, J.M., T. Blenckner, H. Markensten und I. Persson: *Lake phosphorus dynamics and climate warming: A mechanistic model approach*. Ecological Modelling, 190(1-2):1 – 14, 2006, ISSN 0304-3800.

[Mei94] Meister, A.: *Ein Beitrag zum DLR-τ-Code: Ein explizites und implizites Finite-Volumen-Verfahren zur Berechnung instationärer Strömungen auf unstrukurierten Gittern*. Technischer Bericht, Deutsche Forschungsanstalt für Luft- und Raumfahrt e.V., 1994.

[Mei96] Meister, A.: *Zur zeitgenauen numerischen Simulation reibungsbehafteter, kompressibler, turbulenter Strömungsfelder mit einer impliziten Finite-Volumen-Methode vom Box-Typ*. Dissertation, Deutsche Forschungsanstalt für Luft- und Raumfahrt, Göttingen, 1996.

[Mei99] Meister, A.: *Numerik linearer Gleichungssysteme*. Vieweg, 1999.

[Mei01] Meister, A.: *Analyse und Anwendung Asymptotik-basierter numerischer Verfahren zur Simulation reibungsbehafteter Strömungen in allen Mach-Zahlbereichen*, Habilitation, Universität Hamburg, 2001.

[MH04] Malmaeus, J.M. und L. Håkanson: *Development of a Lake Eutrophication model*. Ecological Modelling, 171(1-2):35 – 63, 2004, ISSN 0304-3800.

[MS98] Meister, A. und T. Sonar: *Finite-volume schemes for compressible fluid flow*. Surveys on Mathematics for Industry, 8:1–36, 1998.

[MS07] Morton, K. W. und T. Sonar: *Finite volume methods for hyperbolic conservation laws*. Acta Numerica, 16:155–238, 2007.

[ORF01] Omlin, M., P. Reichert und R. Forster: *Biogeochemical model of Lake Zürich: model equations and results.* Ecological Modelling, 141:77–103, 2001.

[Pat80] Patankar, S. V.: *Numerical Heat Transfer and Fluid Flows.* McGraw-Hill, New York, 1980.

[Per02] Pers, C.: *Model description of BIOLA - a biogeochemical lake model (including literature review).* Technischer Bericht Report RH 16, SMHI, SMHI, Norrköping, Sweden, 2002.

[Per05] Pers, C.: *Modeling the Response of Eutrophication Control Measures in a Swedish Lake.* AMBIO: A Journal of the Human Environment, 34(7):552–558, 2005.

[Per06] Pers, C.: *The BIOLA model.* http://www.smhi.se/sgn0106/if/hydrologi/biola.htm, 2006.

[Pla04] Plato, R.: *Numerische Mathematik kompakt.* Vieweg, 2. Auflage, 2004.

[RIE01] Reynolds, C. S., A. E. Irish und J. A. Elliott: *The ecological basis for simulating phytoplankton responses to environmental change (PROTECH).* Ecological Modelling, 140(3):271 – 291, 2001, ISSN 0304-3800.

[RM08] Reichert, P. und J. Mieleitner: *Lake Models.* Encyclopedia of Ecology, 3:2068 – 2080, 2008.

[SBKB08] Strube, T., J. Benz, S. Kardaetz und R. Brüggemann: *ECOBAS - A tool to develop ecosystem models exemplified by the shallow lake model EMMO.* Ecological Informatics, 3(2):154 – 169, April 2008.

[SG85] Straškraba, M. und A. Gnauk: *Freshwater Ecosystems.* Elsevier, 1985.

[SK04] Schwartz, H. R. und N. Köckler: *Numerische Mathematik.* Teubner, 5. Auflage, 2004.

[SMB+04] Saloranta, T. M., O. Malve, T. H. Bakken, A. S. Ibrekk und J. Moe: *Lake Water Quality Models and Benchmark Criteria.* Technischer Bericht BMW WP6 Delivery Report, NIVA, SYKE, 2004.

[Smo94] Smoller, J.: *Shock Waves and Reaction-Diffusion Equations.* Springer-Verlag, second Auflage, 1994.

[Ste62] Steele, J. H.: *ENVIRONMENTAL CONTROL OF PHOTOSYNTHESIS IN THE SEA.* Limnology and Oceanography, 7, 1962.

[Sto57] Stoker, J. J.: *Water Waves.* Interscience Publisher, New York, 1957.

[Str06] Strube, T.: *Auswirkungen des globalen Wandels auf das Ökosystem Müggelsee - Entwicklung eines Gewässergütemodells mit der Prozessdatenbank ECOBAS.* Dissertation, Mathematisch-Naturwissenschaftliche Fakultät II, Humboldt-Universität zu Berlin, Berlin, 2006.

[Tes09] *Test Set for IVP Solvers.* http://pitagora.dm.uniba.it/~testset/problems/orego.php, June 4. 2009.

[Tor01a] Toro, E. F.: *Riemann Solvers and Numerical Methods for Fluid Dynamics.* John Wiley & Sons, Ltd, Chichester, 2001.

[Tor01b] Toro, E. F.: *Shock-Capturing Methods for Free-Surface Shallow Flows.* John Wiley, 2001.

[VC07] Vázquez-Cendón, M. E.: *Depth Averaged Modelling of Turbulent Shallow Water Flow with Wet-Dry Fronts.* Archieves of Computational Methods in Engineering, 14(3):303–341, 2007.

[Ven95] Venkatakrishnan, V.: *Convergence to Steady State Solutions of the Euler Equations on Unstructured Grids with Limiters.* Journal of Computational Physics, 118:120–130, 1995.

[Wea08a] *Weather in China I.* http://www.chinatoday.com.cn/English/chinatours/hangzhou.htm, May 23. 2008.

[Wea08b] *Weather in China II.* http://www.ilec.or.jp/database/asi/asi-53.html, May 23. 2008.

[Wea08c] *Weather in China III.* http://www.chinahighlights.com/hangzhou/weather.htm, May 23. 2008.

[Zar05] Zardo, P. A.: *Konservative und positive Verfahren für autonome gewöhnliche Differentialgleichungssysteme.* Diplomarbeit, Universität Kassel, 2005.

[ZCB08] Zhang, Hongyan, David A. Culver und Leon Boegman: *A two-dimensional ecological model of Lake Erie: Application to estimate dreissenid impacts on large lake plankton populations.* Ecological Modelling, 214(2-4):219 – 241, 2008, ISSN 0304-3800.

Eidesstattliche Erklärung zur eingereichten Dissertation

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig und ohne unerlaubte Hilfe angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen sind, habe ich als solche kenntlich gemacht. Kein Teil dieser Arbeit ist in einem anderen Promotions- oder Habilitationsverfahren verwendet worden.