Proceedings of the Sixty-Seventh European Study Group Mathematics with Industry

Wageningen, The Netherlands, 26-30 January 2009

Editors:
Jaap Molenaar
Karel Keesman
Joost van Opheusden
Timo Doeswijk

ISBN: xxxx-xxxx-xx

Financial support:

NWO
STW
CWI
ESGI
Biometris

Contents

Chapter 1
Dynamical Models of Extreme Rolling of Vessels in Head Waves

Claude Archer[1] Ed F.G. van Daalen[2] Sören Dobberschütz[3] Marie-France Godeau[1] Johan Grasman[4] Michiel Gunsing[2] Michael Muskulus[5] Alexandr Pischanskyy[6] Marnix Wakker[6]

abstract:
Rolling of a ship is a swinging motion around its length axis. In particular vessels transporting containers may show large amplitude roll when sailing in seas with large head waves. The dynamics of the ship is such that rolling interacts with heave being the motion of the mass point of the ship in vertical direction. Due to the shape of the hull of the vessel its heave is influenced considerably by the phase of the wave as it passes the ship. The interaction of heave and roll can be modeled by a mass-spring-pendulum system. The effect of waves is then included in the system by a periodic forcing term. In first instance the damping of the spring can be taken infinitely large making the system a pendulum with an in vertical direction periodically moving suspension. For a small angular deflection the roll motion is then described by the Mathieu equation containing a periodic forcing. If the period of the solution of the equation without forcing is about twice the period of the forcing then the oscillation gets unstable and the amplitude starts to grow. After describing this model we turn to situation that the ship is not anymore statically fixed at the fluctuating water level. It may move up and down showing a motion modeled by a damped spring. One step further we also allow for pitch, a swinging motion around a horizontal axis perpendicular to the ship. It is recommended to investigate the way waves may directly drive this mode and to determine the amount of energy that flows along this path towards the roll mode. Since at sea waves are a superposition of waves with different wavelengths, we also pay attention to the properties of such a type of forcing containing stochastic elements. It is recommended that as a measure for the occurrence of large deflections of the roll angle one should take the expected time for which a given large deflection may occur instead of the mean amplitude of the deflection.
KEYWORDS: *Mathieu equation, ship dynamics, roll, spring-pendulum systems, stochastic waves*

[1]Ecole Royale Militaire, Belgium
[2]MARIN, Wageningen, The Netherlands
[3]Universität Bremen, Germany
[4]Wageningen University and Research Centre, The Netherlands
[5]Leiden University, The Netherlands
[6]Delft University of Technology, The Netherlands

1.1 Introduction

On October 20th, 1998, a post-Panamax C11 class cargo ship sailed on the Pacific from Taiwan to Seattle. While traversing a heavy storm, the vessel began an extreme rolling motion (transversal swinging) with an angle of up to 40 degrees to each side. After the storm had settled, the crew examined the status of the cargo and found that one third of the containers were lost and another third heavily damaged, making this incident the greatest container casualty known so far (cf. France et al. [9] for a detailed account of the events).

The ship experienced a phenomenon known as "parametric roll" or "parametric resonance": During only a few roll cycles, the roll angle increases far above what would be considered normal (mostly up to 10 degrees). This behaviour of a ship had been known from the 1950s, but only considered to be relevant for smaller vessels in following seas (cf. [26]).

After the October 1998 incident, interest has been renewed. It has been suggested by Shin et al. [26] that the hull shape of modern container ships might increase the risk of parametric roll. In order to enlarge the load capacity while keeping the water resistance small, the length and width of ships increased and a wide, flat stern and pronounced bow flares appeared. This had an effect on the ship's stability when encountering waves.

Possible countermeasures to avoid heavy rolling include the attachment of stabilizing fins to the outer hull of the ship or the installation of active water tanks in the interior of the ship. However, these actions increase the fuel consumption and lessen the number of containers which can be carried. That is why these techniques are not used for modern cargo ships (cf. [26]).

Obviously, the estimation of the risk of parametric resonance for a given ship geometry and load characteristic is of great importance to ship owners and constructors. Research centres such as the Marine Research Institute Netherlands (MARIN) therefore try to predict the probability of the occurrence of parametric roll by using computer simulations and model tests.

This paper reviews models for describing the excitation of the rolling motion. Methods for analyzing autoparametric resonance in mechanical systems ([30]) are applied to three different models: the variable length pendulum model (Section 2), the spring-pendulum model (Section 3) and the spring-double pendulum model, see Section 4. These models describe a ship as a force-driven dynamical system of springs and pendulums with one up to three degrees of freedom. However, these models only consider a single encounter frequency. To account for the more realistic situation of a superposition of different waves

we add up different waves containing stochastic elements in a way that an appropriate type of spectrum is composed, see Section 5.

1.1.1 Shape of a vessel and its metacentric height

In this part, we discuss the main causes of this sudden large amplitude rolling motion [26], such as the new design of the hull of container ships influencing the stability under heavy weather conditions and the phenomenon of parametric resonance as it applies to roll dynamics.

In the seventies, ships had a payload of about 2000 containers and were able to sail at about 25 knots. Nowadays, some container ships are built to carry 10000 containers, or even more, and are still sailing at the same speed. To achieve this without increasing the fuel consumption, it was necessary to give thinner shapes to hulls. Consequently:

- Ships are longer and their length now approximately corresponds to the length of waves as they are met in the Pacific and the North Atlantic Ocean.
- The bow and stern shapes are thinner and more extended to the centre of the ship, the section in the centre with a fixed U-shaped cross-section of the hull is now forming a much smaller proportion of the ship, see Figure 1.

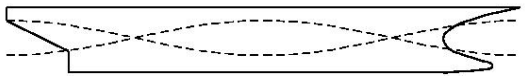

Figure 1.1: Change in waterline surface of a ship at different phases of the passing wave.

These two aspects have as consequence that, when the ship encounters waves, its changed dynamical characteristics may bring about a critical response. Most important factor is the varying metacentric height of the ship, this is a vector pointing to the centre of gravity of the vessel with the centre of buoyancy as origin (the centre of buoyancy is the gravity centre of the displaced body of water). The averaged transverse cross-section of the vessel yields the transverse component (GM) of the metacentric height, see Figure 2. This component determines for a large part the roll amplitude. If the ship length is about the length of the incoming wave, the ship periodically passes two extremes with the middle of the ship at a crest or in a trough. From Figure 1 it is seen that this makes for a large

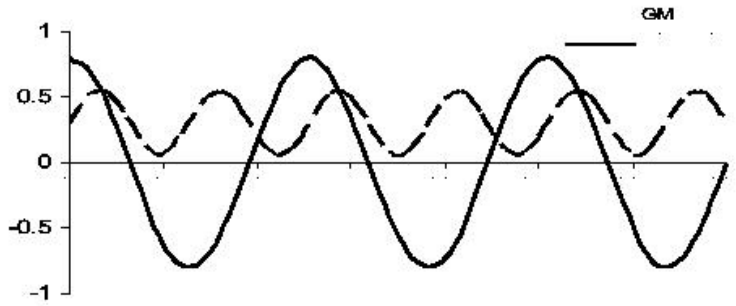

Figure 1.2: Roll angle in the course of time in the resonant state. Note that the period is twice the period of the wave. The dotted line denotes the variation of the metracentric height GM from the waves.

difference in buoyancy. Consequently, the restoring force varies a lot. The GM represents indirectly the righting lever $GZ = GMsin\phi$, where ϕ is the roll angle, see Figure 2.

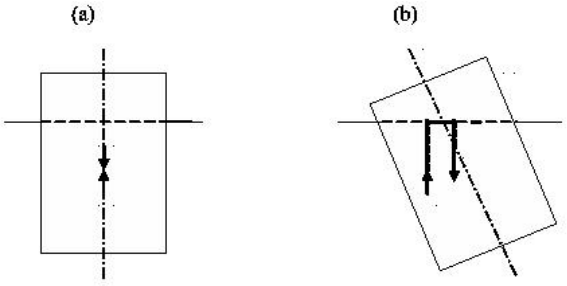

Figure 1.3: Cross section of a floating container. The line that connects the centre of buoyancy B with the gravity centre G of the container represents the metacentric height. (a) The rest state; note that the gravity force is compensated by the upward force of the displaced body of water. (b) The container is out of balance; the centre of buoyancy has moved to the left. The two forces are balanced in the vertical direction, but cause a (clockwise) restoring moment GZ equal to the projection of the metacentric height GM (line BG) upon the horizontal axis.

When the ship is on the crest of a head wave whose length is about the length of the ship, the waterline surface of the ship (surface area defined by the intersection of the hull and the water surface) takes a minimum value. Correspondingly GZ as well as the roll stability are both at a minimum. Vice versa, GM and the stability are at their maximum when the ship is in the trough of the wave. These successive variations of stability can cause large roll motions. Since this phenomenon occurs with waves affecting the restoring force with a frequency being about twice the natural roll frequency and results in a large

roll amplitude, it is called parametric resonance. In Figure 3 the different stages of the roll cycle can be discerned. When the ship is away from the vertical (0 degrees angle) and in a deep trough, the righting lever is larger than in calm water, and the ship comes back faster to the vertical while accumulating kinetic energy. At the end of the first quarter of the roll period, the ship crosses the vertical and continues its move to the other side because of the inertia. During this second quarter of the roll period, the ship is on a crest, so the righting lever is smaller than in calm water, and the ship continues its movement to a larger roll angle, due to the accumulated kinetic energy. In the third quarter the action of the first quarter is repeated at the opposite side and the fourth quarter mirrors the second quarter.

1.1.2 Extreme rolling

The rolling motion may re-enforce itself so that the amplitude of the roll motion increases within a couple of roll periods from a few degrees to about 40 degrees. The resonant dynamics from a periodic change of the GM due to waves coming in with a frequency being twice the natural frequency is best understood from a pendulum whose length varies with time. The amplitude of such a pendulum can be increased if its length varied in such a way that it is smaller when it moves away from the vertical and larger when it gets close to the vertical. This means that the length must vary with a frequency that is twice the natural frequency of the pendulum. The GM of the ship can be seen as the length of the pendulum. This phenomenon is similar to a child sitting on a swing who tries to get higher by being seated when the swing gets close to the vertical and getting up when it moves away from the vertical. The centre of gravity of the child then rises (smaller pendulum length) and falls (larger pendulum length) successively making the swing a (parametricly) resonant system which we will discuss in Section 2.

For getting parametric roll in case of a rolling vessel, some conditions must be fulfilled:

- The length of an incoming wave is about the length of the ship
- The waterline surface of the ship varies considerably during a wave cycle
- The encounter frequency of the waves is about twice the natural roll frequency of the ship

1.1.3 Models with more than one degree of freedom (DOF)

In spite of these apparently strict conditions, it's still very difficult to accurately estimate the risk of occurrence of parametric roll. It is our aim to select mathematical models that

suitably can be used for numerical simulations based on naval architecture data. Eissa et al. [6] describe a 2-DOF nonlinear spring pendulum model. By multiple time scale perturbation techniques, approximations of the solutions up to 4th order are obtained, together with stability regions and solvability conditions.

A 3-DOF nonlinear model was developed by Neves [23], Neves and Rodriguez [24], using a Taylor-series expansion up to second (and later to third) order of the damping and restoring forces. The governing equations are found to be a coupled system of Hill equations ([12]), with the parameters given explicitly by the ship's characteristics and geometry. This model has been implemented in MATLAB by Holden et al. [13] and is now a part of the Marine Systems Simulator[1].

1.2 The Parametric Pendulum (1-DOF)

The swing analogy leads to the differential equation of a variable length pendulum. The general 1-DOF equation for the roll angle ϕ is

$$(I + A)\frac{d^2\phi}{dt^2} + B\frac{d\phi}{dt} + C(t)sin(\phi) = 0, \tag{1.1}$$

where $C(t)$ is the restoring force, I the ship inertia, A the added mass and B the damping in the roll direction ([17], [10]). The restoring force is $GZ = GMsin(\phi)$. For small angle it is linearised as $GZ = GM\phi$. Dunwoody [5] showed that in calm water GM oscillates around its mean value GM_m and also that the amplitude GM_a of the oscillation is proportional to the wave elevation. In the case of a single frequency wave with angular velocity ω, Dunwoody's results give

$$C(t) = \rho g\Delta(GM_m + GM_a cos(\omega t)), \tag{1.2}$$

where Δ is the displacement of the ship (water equivalent of the immersed volume of the ship), ρ the water density and g the gravity constant. If the pendulum angle ϕ is small Eq.(1.1) can be linearized taking the form of the periodically forced Mathieu equation, see Tondl [30]:

$$\frac{d^2\phi}{dt^2} + b\frac{d\phi}{dt} + (c + dcos(\omega t))\phi = 0, \tag{1.3}$$

where b is the damping coefficient, c the coefficient of the restoring force and d and ω respectively the amplitude and angular velocity or frequency of the periodic forcing. The natural frequency ($d = 0$) is

[1] Available under a GNU General Public License, see www.marinecontrol.org.

$$\omega_0 = \sqrt{c - b^2/4}. \tag{1.4}$$

for which resonance behavior for a forcing at twice the natural frequency is well known. More general the phenomenon may occur for any periodic forcing term. The corresponding differential equation is then referred to in the literature as a Hill equation, see Hochstadt [12].

1.2.1 A threshold for parametric roll

In [10] the threshold for parametric roll has been studied for both the original and the linearized restoring force term. We have seen before that the restoring force of the ship against rolling is larger when the wave trough is amidships than when a wave crest arrives at this point. This is due to the variation of GM. Let δGM be the difference of GM between these two extreme cases and let $p = \delta GM/GM_m$ be the proportion that represents this variation with respect to the calm water GM_m. A large ratio p corresponds to a high probability of having parametric roll. The threshold for parametric roll is then expressed as the critical minimal value of p for which a large response (resonance) occurs. For the Mathieu equation ([30]) this critical value can be estimated for a given ω. For a ship model it has been predicted in [10] that parametric roll starts when $p > 4\mu/\omega_0$ with damping ratio $\mu = \frac{1}{2}B\omega_0/((I + A)\omega_0)$ and ω_0 given by (1.4). An accurate determination of the damping ratio μ is crucial as it will strongly influence the model prediction.

In [17], this threshold has been validated against basin data for a simple hull form and compared with results from the full non-linear time domain seakeeping code PRETTI developed for improving the computation of the effect of hull forms (not yet tested in a basin). The authors conclude that the 1-DOF model can very well be employed in preliminary hull design with a damping factor μ tuned with the use of empirical data. It gives an idea of the threshold wave height for which parametric roll occurs. For the estimation of the amplitude of the actual roll angle a nonlinear model is needed for which for example the code PRETTI can be used. Of course this requires more computing time.

1.2.2 Further investigations

Most of the literature on parametric roll focusses on two approaches: the derivation and theoretical analysis of a mathematical point-ship model and/or the numerical implementation of such a model using the appropriate vessel and wave specifications.

France et al. [9] and Shin et al. [26] considered a 1 degree-of-freedom (DOF) ship model. By linearizing the moments of damping and restoring, the authors arrive at a

Mathieu-equation to describe the ship's behaviour. However, Spyrou [28] pointed out several disadvantages of linearized models: the linearization is only valid for small roll angles, whereas parametric roll is characterised by large roll amplitudes. In addition, the instability regions of Mathieu-type models reflect the behaviour given an infinite number of roll cycles – but we already mentioned that parametric roll only takes a few of them to build up.

1.3 A Model for Heave-Roll Motion(2-DOF's)

The specification of resonance conditions is one of the most important topics in the prediction of parametric roll. Running large scale models formulated by researchers and engineers may need hours or even days of computing time so that a quick estimate of parameters relevant for roll cannot be obtained from it. Therefore, research is also directed to the formulation of low dimensional models or simple schemes which may predict heavy rolling given certain operational conditions. In this section one of such models will be considered, see Figure 4.

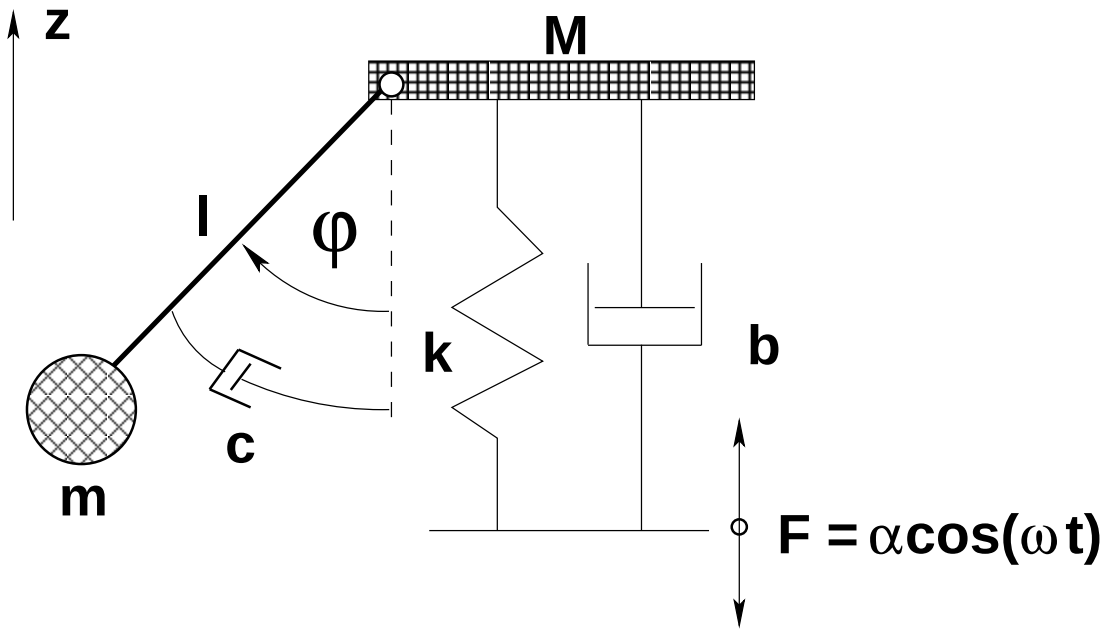

Figure 1.4: Driven spring-pendulum model for heave-roll motion.

It is a two degrees of freedom model formulated by Tondl et al. [30] consisting of a mass mounted to a periodically moving floor by a linearly damped spring. In addition a

pendulum is connected to this mass. An external periodic force of the form $\alpha\cos(\omega t)$ is applied to the floor with α and ω respectively amplitude and frequency of the external periodic force. This system satisfies the following two coupled differential equations:

$$\begin{aligned}&(M+m)(\ddot{z}-\alpha\omega^2 cos(\omega t))+b\dot{z}+kz+ml(\ddot{\phi}sin(\phi)+\dot{\phi}^2cos(\phi))=0,\\&ml^2\ddot{\phi}+c\dot{\phi}+mglsin(\phi)+ml(\ddot{z}-\alpha\omega^2cos(\omega t))sin(\phi)=0.\end{aligned} \tag{1.5}$$

After carrying out transformations of the time- and the dependent variables and making a small perturbation approximation the system is described by the following dimensionless linear differential equations:

$$\begin{aligned}&\ddot{u}+\kappa\dot{u}+q^2u=0,\\&\ddot{\psi}+\kappa_0\dot{\psi}+\psi-a\eta^2\left[(1+A)\cos(\eta\tau)+B\sin(\eta\tau)\right]\psi=0,\end{aligned} \tag{1.6}$$

where u and ψ are the vertical and the angular displacements of respectively the mass and the pendulum, $\kappa_0=c/(\omega_0 ml^2)$, $a=\alpha/l$, $\eta=\omega/\omega_0$ and

$$A=\frac{\eta^2(q^2-\eta^2)}{\Delta},\ B=\frac{\kappa\eta^3}{\Delta},$$

with $\Delta=(q^2-\eta^2)^2+(\kappa\eta)^2$, $q^2=k/(\omega_0^2(M+m))$, $\kappa=b/(\omega_0(M+m))$. Here m is the mass of the pendulum, M is the mass of the oscillator, l is the length of the pendulums weightless rod, ω_0 is the eigenfrequency of the oscillator, b and c are the damping coefficients of the linear and angular motions, k is the stiffness coefficient of the spring, and the dot denotes the derivative with respect to the time variable τ.

The value of the parameter η determines for a large part the stability of the rolling motion. For a value close to 2 an unstable motion may occur with large angular deflections of the pendulum. This only happens if the forcing amplitude a is above a threshold. In [30] it is derived that this threshold depends on η and the other parameters in the following way:

$$a_{lin}=\frac{2}{\eta^2}\left[\frac{\left(1-\frac{1}{4}\eta^2\right)^2+\frac{1}{4}\kappa_0^2\eta^2}{(1+A)^2+B^2}\right]^{1/2}, \tag{1.7}$$

Thus, for a ship-wave system Eq.(1.7) represents the value of the amplitude of regular wave with a fixed frequency above which the equilibrium gets unstable.For a vessel it means that heavy rolling may build up in a short time.

1.3.1 Computation of the threshold values from experimental data

This section deals with four experiments with physical models in the form of scaled ships towed in a basin. In addition at MARIN computations are carried out for these ships using

a potential flow code (PFC). For each case these numerical investigations yield estimates for parameters occurring in Eq.(1.7). Two of the MARIN experiments are related to a vessel X with two different input data-sets. Parametric rolling has been observed during the experiments and was confirmed by the model computations. The third experiment with vessel Y did not show parametric roll which is in agreement with the computations using PFC. In the last experiment related to the vessel Z parametric roll was observed during the tests of the physical scaled model in the basin. However, in this case it was not predicted by the computations. All experiments were carried out with head waves acting upon the vessel.

The input data-sets for the models include the following parameters: the frequency of the external force ω [rad/s], the heave and the roll damping coefficients being respectively b and c and the stiffness coefficient k. The data-sets have been computed by MARIN using PFC for a range of frequencies of the excitation force, and were given to us to analyse them using the results of the linear approximation Eq.(1.6). It is noted that the threshold amplitude Eq.(1.7) does not explicitly depend on the speed of the waves (or the vessel). Of course, the speed is implicitly present in the input data (Doppler effect).

Substituting the input data into Eq.(1.7) the threshold amplitude a_{lin} can be obtained for each of the four vessels. However, some transformations and additional calculations have to be made first. The length l of the pendulum can be found using the well-known equation of the period of small oscillations of the physical pendulum $T = 2\pi\sqrt{l/g}$, where T is the roll period of the ship, l is the length of the 'equivalent' mathematical pendulum satisfying $l = gT^2/(4\pi^2)$. Since the rolling does not depend directly on the wave excitation but on the heave motion of the vessel, a transfer function is needed. Thus the final threshold amplitude for the waves can be written as

$$a_{thr} = \frac{a_{lin} l \omega^2 M_h}{S(\omega)}, \tag{1.8}$$

where M_h is the mass of the vessel moved by the heave motion and $S(\omega)$ the transfer function taken from the 'MARIN' data.

To simplify the calculations a C-code program has been written in such a way that the threshold amplitude a_{thr} directly can be obtained from the input data-set for each value of the frequency of the excitation force. The graphs of the threshold amplitude for each set of data (for each vessel) are presented in Figures 5abcd.

Parametric resonance of the vessel for all models can most likely be expected in the range of small frequencies of the external force up to 0.8 [rad/s], with values for the

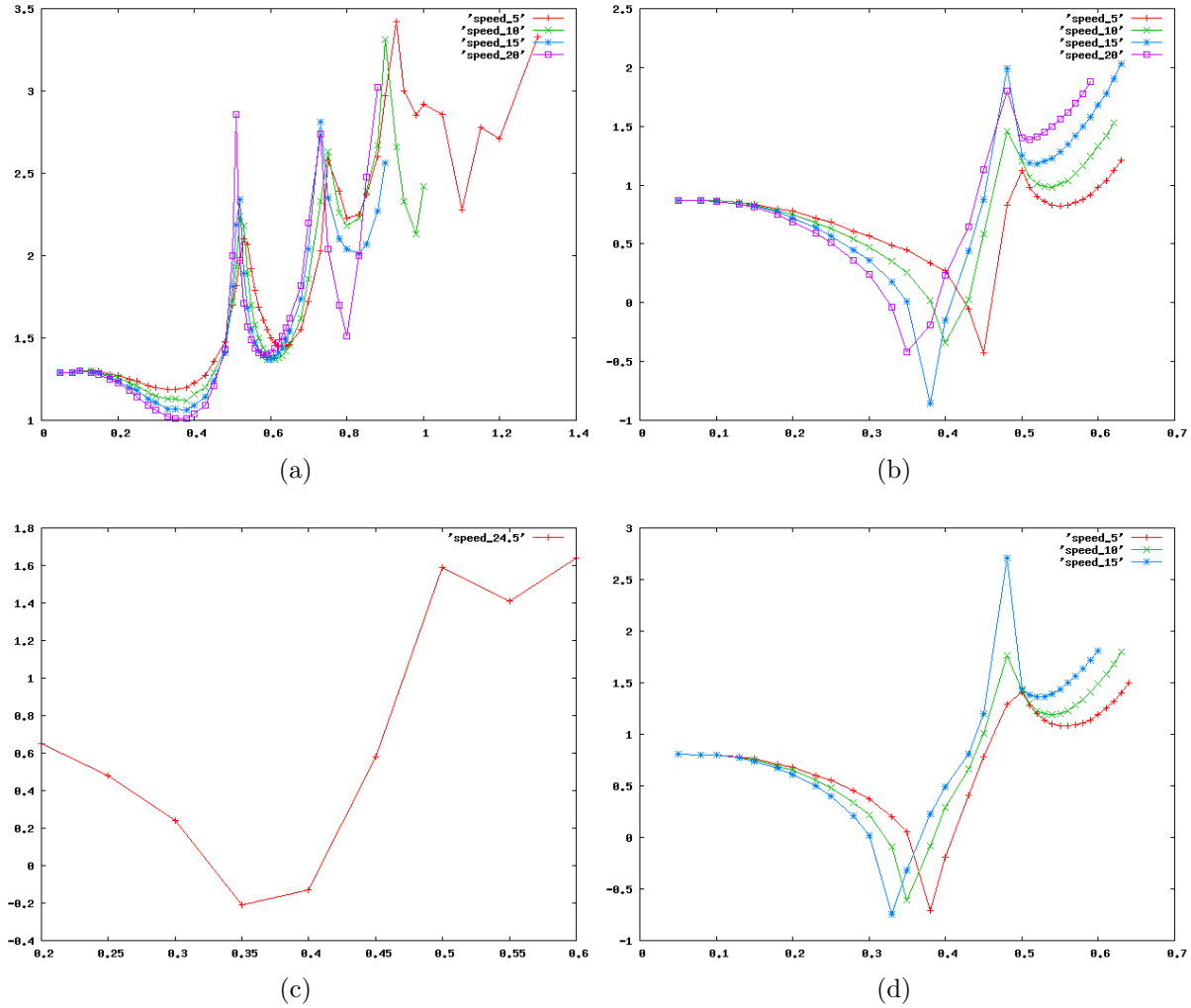

Figure 1.5: Logarithm of threshold amplitudes α of a single frequency wave forcing as a function of the wave frequency ω for different speeds of the vessel. (a) Vessel X for a large range of wave frequencies. (b) Vessel X for low frequency waves in more detail; note the low threshold value at $\omega = 0.4$. (c) For vessel Y a similar, but less pronounced, resonance is found. (d) Result for vessel Z.

amplitude of the waves from around 0.4 to around 30 meters. These threshold amplitudes can be so small because all energy goes in that wave with a single frequency. In reality waves consisting of various stochastic components will increase these values considerably. The parts of the graphs for the higher frequencies are cut out because the values of the threshold amplitude are very large.

1.4 A Spring with Two Pendulums (3-DOF's)

For further study of ships, we define our coordinate vectors as $\vec{x}$ pointing from the center of mass towards the front of the ship while $\vec{y}$ and $\vec{z}$ are pointing towards the right side of the ship and in the upwards vertical direction, respectively. We know that roll and pitch (rotation around resp. the x and the y axis) and heave (movement in the z direction) are strongly coupled. The important question is now: under what conditions is it possible that pitch and heave motion are rapidly converted into roll motion? As a first observation we should mention that the moment of inertia for pitch is much larger than the one for roll. This means that even for small pitch angles a potentially dangerous amount of angular momentum is stored in pitch.

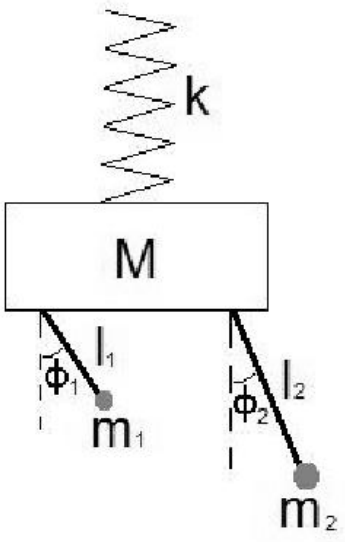

Figure 1.6: Spring and two-pendulum model for heave-roll-pitch motion. Note that in this model there is no direct interaction between roll and pitch.

In a spring - pendulum model (SPM) the pendulum accounts for roll and the spring accounts for heave, see Tondl et al.[30]. In this publication also the double spring - pendulum model (DSPM) is introduced, so that also pitch is included. The system consists of two springs connected by a rod that turns freely around its length axis. A pendulum, connected to this rod, models the roll of the ship, while the two springs capture heave and pitch. Because in this model heave and pitch are identified by separate state variables, it is not possible to see the effect of "pitch energy" being released. Instead of the DSPM, we prefer to present the spring - double pendulum model (SDPM), see Figure 6. It consists of a large mass M (related to the mass of the ship) mounted by a vertical spring to a fixed point. The spring constant k represents the restoring force component in the heave mode. The position of the fixed point may change in time denoting forcing by waves. The pendulums have masses m_i attached to the large mass by rods of length l_i, see Figure 6. It is assumed that the pendulums swing independent from each other (physically in directions perpendicular to each other) and that they only interact via the heave motion. The product $m_1(l_1)^2$ is determined by the moment of inertia around the x axis (roll), $m_2(l_2)^2$ is determined by the moment of inertia around the y axis (pitch). The three degrees of freedom of this system are z, ϕ_1 and ϕ_2 being the vertical displacement of mass M from the stationary state and the angles with the vertical for both pendulums. The kinetic energy is given by

$$T = \frac{1}{2}M\dot{z}^2 + \sum_{i=1}^{2}\left[\frac{1}{2}m_i\left(\dot{z} + l_i\dot{\phi}_i\sin\phi_i\right)^2 + \frac{1}{2}m_i\left(l_i\dot{\phi}_i\cos\phi_i\right)^2\right], \tag{1.9}$$

and the potential energy is given by

$$V = \frac{1}{2}kz^2 + \sum_{i=1}^{2} m_i g l_i(1-\cos\phi_i). \tag{1.10}$$

Since the heavy sideway roll in practice appears within a few periods, we assume that external forcing cannot account for this effect. Therefore we do not include an external force. Of course the initial heave and pitch are caused by an external pulse or by setting a far from equilibrium initial state. We investigate the conditions for transfer of heave and pitch into roll.

1.4.1 Equations of Motion

For the Lagrangian $L = T - V$, the Euler-Lagrange equation for variable u is given by

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{u}} - \frac{\partial L}{\partial u} = 0. \tag{1.11}$$

For the model under consideration, the Lagrange equations are found to be

$$(M+m_1+m_2)\ddot{z}+b\dot{z}+kz+m_1l_1(\ddot{\phi}_1\sin\phi_1+\dot{\phi}_1^2\cos\phi_1) \\ +m_2l_2(\ddot{\phi}_2\sin\phi_2+\dot{\phi}_2^2\cos\phi_2)=0, \tag{1.12a}$$

$$l_1\ddot{\phi}_1+c_1\dot{\phi}_1+g\sin\phi_1+\ddot{z}\sin\phi_1=0, \tag{1.12b}$$

$$l_2\ddot{\phi}_2+c_2\dot{\phi}_2+g\sin\phi_2+\ddot{z}\sin\phi_2=0, \tag{1.12c}$$

where b, c_1, and c_2 are the damping coefficients and g is the gravitation acceleration. A large amplitude response from a periodic forcing is expected if the damping terms are small. Therefore, in our model we neglect these terms. Eqs. (1.12) are linearized using the small angle approximation:

$$(M+m_1+m_2)\ddot{z}+kz+m_1l_1(\ddot{\phi}_1\phi_1+\dot{\phi}_1^2-\frac{1}{2}\dot{\phi}_1^2\phi_1^2) \\ +m_2l_2(\ddot{\phi}_2\phi_2+\dot{\phi}_2^2-\frac{1}{2}\dot{\phi}_2^2\phi_2^2)=0, \tag{1.13a}$$

$$l_1\ddot{\phi}_1+g\phi_1+\ddot{z}\phi_1=0, \tag{1.13b}$$

$$l_2\ddot{\phi}_2+g\phi_2+\ddot{z}\phi_2=0. \tag{1.13c}$$

It is noted that only small deflections from equilibrium are correctly approximated. Still one can study possible strong responses of the nonlinear system to periodic perturbations. At the moment the deflections grow large only qualitative information is obtained from this approach. The solutions for the uncoupled equations are given by

$$z(t)=a_0\sin(\omega_0 t)+b_0\cos(\omega_0 t), \tag{1.14a}$$

$$\phi_1(t)=a_1\sin(\omega_1 t)+b_1\cos(\omega_1 t), \tag{1.14b}$$

$$\phi_2(t)=a_2\sin(\omega_2 t)+b_2\cos(\omega_2 t), \tag{1.14c}$$

$$\omega_0=\sqrt{\frac{k}{M+m_1+m_2}},\quad \omega_1=\sqrt{\frac{g}{l_1}},\quad \omega_2=\sqrt{\frac{g}{l_2}}. \tag{1.14d}$$

1.4.2 Series Solutions based on the Mathieu equation

In this section we explore the possible application of a Galerkin type of approximation of the solution of Eq.(1.13). This approach is widely used and, in the way it applies to this problem, it is also known as the "spectral method", see [2].

Since all equations of motion are invariant under time translations, we have the freedom of taking $a_0=0$, because later $z(t)$ can be shifted in time, and all solutions of ϕ_1 and ϕ_2

with it. We insert the harmonic solution $z(t) = b_0 \cos(\omega_0 t)$ in Eqs.(1.13bc). Introduction of a new independent variable $x \equiv (\omega_0 t)/2$ transforms both equations into the Mathieu equation:

$$\frac{d^2y}{dx^2} + [a - 2k^2 \cos(2x)]y = 0, \quad k^2 = q. \tag{1.15}$$

It has periodic solutions for infinitely many a, depending on the value of q. Four series of these 'eigenvalues' a do exist, and correspond to four possible combinations $k, l \in \{0, 1\}$ in the general series of solutions (Gradshteyn and Ryzhik [14]):

$$f_{kl}(x) = (\cos(x))^k \sum_{n=0}^{\infty} f_{kln} (\sin(x))^{2n+l}. \tag{1.16}$$

The derivative $(df_{kl})/(dx)(x)$ can be expressed as $g_{(k+1)(l+1)}(x)$ with a series expansion as given by Eq.(1.16) and with indices taking values modulo 2. In a similar way we handle $(d^2 f_{kl})/(dx^2)(x) = h_{kl}(x)$. For a product $p_{k_1 l_1}(x) q_{k_2 l_2}(x)$ we can derive an expression of the form $r_{(k_1+k_2)(l_1+l_2)}(x)$ with again indices that are taken modulo 2.

Let us ignore the perturbation of the heave motion by the pendulums and replace the solution $z = b_0 \cos(2x)$ by

$$z = \sum_{n=0}^{\infty} z_n (\sin(x))^{2n}. \tag{1.17}$$

Then we can express the solutions of the pitch- and roll equations as

$$\phi_i(x) = (\cos(x))^{k_i} \sum_{n=0}^{\infty} (\phi_i)_n (\sin(x))^{2n+l_i}. \tag{1.18}$$

For z it is important to take $k = l = 0$ to have the same 'overall' k and l for all terms in Eq.(1.13a) . For ϕ_1 and ϕ_2, k and l are not fixed by Eqs.(1.13bc), mainly because their first derivatives appear quadratically in Eqs.(1.13). Since products of series do couple all coefficients, it is not expected that the nonlinear recursion relations give analytical results. However, taking only the first 5 terms in the series representations, for explicit values of the constants one can numerically solve the equations for the coefficients. A stable solution with strong pitch and roll indicates the possibility of parametric roll.

1.4.3 Effect of head waves

If we include forcing from waves coming in with an angular velocity ω_0 and if we also assume that it only acts upon heave being heavily damped, then the system (14a)-(12bc) covers separately the 1-DOF model for roll (Section 2) as well as for pitch. If wave forcing

is added to DSPM (12), then we arrive at a type of model with periodic forcing. A new element is brought in if the pitch-pendulum is considered to be directly forced by the head waves. It is known that heavy roll occurs at wave length of about the length of the ship. This means that the pitch motion will have a component with the wave frequency and an additional energy flow is expected from the pitch motion to heave and roll. Resonant forcing of the pitch motion would even enlarge this energy flow.

1.5 Stochastic aspects

Up to now we have only considered the response of a ship that encounters a single frequency wave. In reality, the sea state is a complex mixture of waves with many different frequencies, and in this section we describe possible approaches to this more difficult problem.

1.5.1 Stochastic description of ocean waves

Realistic sea states are conveniently described by their spectral properties, see Podgorski et al. [20]. Let $\zeta(t)$ denote the sea surface elevation at time t. It is assumed that this is a weakly stationary ergodic random process, which is usually satisfied for deep water waves. More specifically, $\zeta(t)$ can usually be assumed to be a Gaussian process (whose variance characterizes the sea severity) [19]. Its autocorrelation function is defined as expectation

$$R(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} (\zeta(t) - \mu)(\zeta(t+\tau) - \mu) \, \mathrm{d}t, \tag{1.19}$$

where μ is the mean surface elevation, usually chosen to be zero. By the Wiener-Khintchine theorem, the power spectral density $S(\omega)$ is given as the Fourier transform of $R(\tau)$,

$$S(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} \, \mathrm{d}\tau, \tag{1.20}$$

and represents the average wave energy (density) for a given frequency component.

The sea state is generated by the complex interaction of the local wind field with the sea surface, and therefore this stochastic description is often adequate. Moreover, the spectral density $S(\omega)$ can be calculated from first principles (see e.g., [18]). In practice, however, it is more conveniently described by phenomenological models.

The Pierson-Moskowitz spectrum describes a *fully-developed* sea, i.e., a sea state that is in equilibrium with the local wind field:

$$S(\omega) = \frac{\alpha g^2}{\omega^5} \exp\left[-0.74 \left(\frac{\omega_0}{\omega}\right)\right], \tag{1.21}$$

where $\alpha = 8.1 \times 10^{-3}$, $g = 9.81\text{m/s}^2$ is the usual gravitational acceleration, and $\omega_0 \approx g/(1.026 \cdot U_{10})$ depends on the wind speed U_{10} at 10 meter height. An example for different wind speeds is shown in Figure 7a.

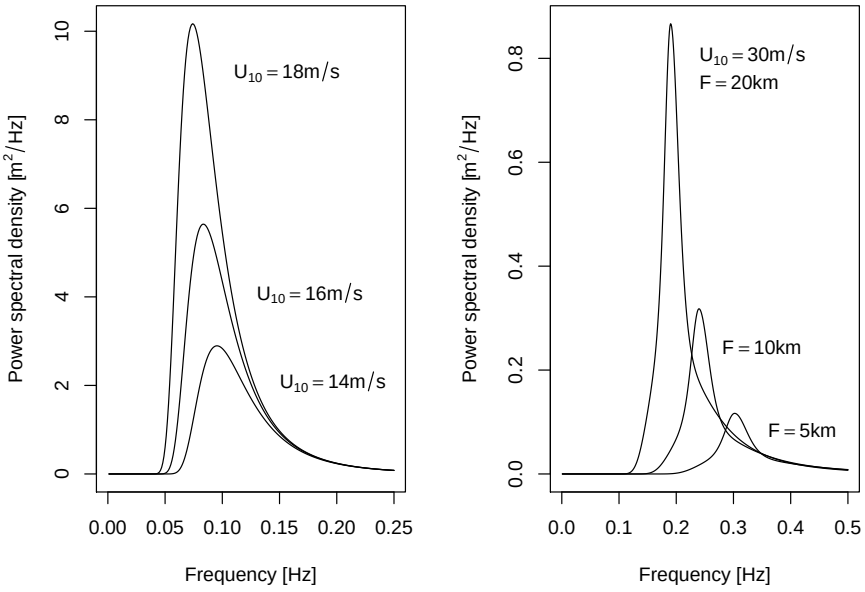

Figure 1.7: Widely used wave spectra for stochastic simulation. (a) Pierson-Moskowitz spectrum for a fully developed sea. (b) JONSWAP spectrum for the North Sea for different values of the fetch (see text). Note that the JONSWAP spectrum does not represent a fully developed sea!

The JONSWAP (Joint North Sea Wave Project) spectrum in Figure 7b takes into account that most sea states are rarely fully developed. It features the additional parameter F, the so-called *fetch*, which represents the distance over which the wind blows with constant velocity U_{10}. Explicitly, it is given by

$$S(\omega) = \frac{\alpha g^2}{\omega^5} \exp\left[-\frac{5}{4}\left(\frac{\omega_p}{\omega}\right)^4\right] \gamma^r, \quad r = \exp\left[-\frac{(\omega-\omega_p)^2}{2\sigma^2\omega_p^2}\right], \tag{1.22}$$

where now $\alpha = 0.076(U_{10}^2/(F \cdot g))^{0.22}$, $\omega_p = 22(g^2/(U_{10} \cdot F))^{1/3}$, $\gamma = 3.3$, and $\sigma = 0.07$ for $\omega \leq \omega_p$, and $\sigma = 0.09$ otherwise.

Sometimes an additional component is visible in real spectra, the so-called *swell*. This is caused by a distant wave field that has travelled into the area and is superimposed on the locally generated field.

1.5.2 Simulations

Knowledge of $S(\omega)$ allows the efficient simulation of the underlying process, by approximating it by a finite mixture of frequencies. Assuming that the process $\zeta(t)$ is bandlimited[2], i.e., $S(\omega)$ is zero outside an interval $[-\Delta, \Delta]$, let us divide the interval $[0, \Delta]$ into N subintervals of length Δ/N. Then we can write

$$\tilde{x}(t) = \sum_{k=1}^{N} A_k \cos(\omega_k t + \epsilon_k), \tag{1.23}$$

and the process $\tilde{x}(t)$ exhibits (approximately) the same stochastic properties as $\zeta(t)$. Here the phases ϵ_k are chosen uniformly from the interval $[-\pi, \pi]$, and the amplitudes A_k are given by $A_k = 2\left(S(\omega_k)\Delta_k\right)^{1/2}$. This method goes back to the seminal work of Rice [22], and has been widely used in applications. It is also the basis for the recently developed *surrogate data* methods in nonlinear time series analysis. However, mathematically it is preferable to use

$$x(t) = \sum_{k=1}^{N} R_k \cos(\omega_k t + \epsilon_k), \tag{1.24}$$

where R_k has a Rayleigh distribution with parameter $2S(\omega_k)\Delta_k$ [29].

Instead of using a regular division of the power spectral density, one can also sample frequency components randomly, according to the probability density

$$p(\omega) = \frac{S(\omega)}{\sigma^2}, \tag{1.25}$$

where $\sigma^2 = \int_0^\infty S(\omega)\,\mathrm{d}\omega$ is the variance of $\zeta(t)$ [16].

A drawback of both these methods is (i) that the spectrum of the simulated elevations is discrete, and (ii) that the process $\zeta(t)$ is only approximated well for N sufficiently large. To overcome this limitation, one can consider a *disordered periodic process*,

$$y(t) = \sum_{k=1}^{N} B_k \cos(\omega_k t + \nu_k), \tag{1.26}$$

where the ν_k are independent white-noise processes. This leads to a continuous spectrum and even allows the derivation of analytical results in a few special cases, applying the theory of stochastic differential equations [1, 16, 21]. However, there is no simple relation between the spectra of $\zeta(t)$ and $y(t)$, and the former therefore needs to be fitted nonlinearly to $S(\omega)$.

[2]In practice, the cut-off frequency Δ is usually chosen to be the Nyquist-frequency with which the data have been sampled.

Finally, for simulation purposes the most efficient representation is in terms of a time-discrete ARMA(P,Q) process,

$$z_t = \sum_{k=1}^{P} a_k z_{t-k} + \sum_{l=1}^{Q} b_l \epsilon_{t-l} + \epsilon_t, \tag{1.27}$$

where the ϵ_t are uncorrelated Gaussian white-noise errors. As before, the coefficients need to be found by nonlinearly fitting the spectrum of z_t to $S(\omega)$. These processes have been popularized by Spanos and Mignolet [27].

Multivariate generalizations of these techniques also exist. In particular, if one is interested in a specific sea state, from which time series recordings are available, the nonparametric simulation technique of DelBalzo et al.[4], going back to earlier work of Scheffner and Borgman [25], can be employed [4]. Thereby, the variables of interest (sea elevation, wave periods, wave directions, etc.) are transformed to (correlated) Gaussian variables. Realizations of these with the desired correlation structure can then be easily obtained from the eigendecomposition of the covariance matrix, and transformed back into time series.

1.5.3 Nonstationary sea states

In the above, we have still assumed that the sea state is stationary. This assumption is often warranted for timescales of up to a few hours to days, but in general it is clear that the properties of the sea change over the course of time.

The standard methods to deal with this issue are so-called sea-state *prediction graphs* (e.g., see Table 2.2 in [7]). These are basically discretized probability distributions of wave spectra that state the probability to find a given wave spectrum in a random observation interval. From these, a nonstationary time series of surface elevations $\zeta(t)$ can be achieved as a hidden Markov process, where a continuous-time Markov chain allows the process to switch from one wave spectrum to another regime.

1.5.4 Connection with ship dynamics

How do the above considerations connect with the dynamical evolution of the state of a ship? We have already seen in Eq.(6) that the equations of motions are uncoupled to first order in the parameter η with respect to roll motion. Excitation of roll motion is facilitated through a restoring moment $\gamma(\varphi, t)$, that is given by the so-called *righting lever curve* for the ship under consideration. In calm water, this moment is time-independent, but in general the righting-lever curves capture the effect of dynamical changes in stability and

depend on wave properties and the ship's speed. The slopes of these curves in the upright position correspond to the variation in metacentric height GM that the ship experiences, and the restoring moment can be modelled by

$$\gamma(\varphi, t) = (1 + \delta \cos \omega t)\varphi - \alpha\varphi^3, \tag{1.28}$$

where 2δ corresponds to this variation, and ω is the wave encounter frequency [16]. Note that the latter is related to the wave frequency ω_0 by a Doppler shift, $\omega = \omega_0 - \omega_0^2 U/g \cdot \cos\mu$, where μ is the angle between wave propagation and the ship's forward direction, and U is the ship's speed. The additional parameter α captures the nonlinearity of the restoring lever curves and can be fitted from available design data.

The effect that a sea state will have upon the motion of a ship is usually expressed in terms of the so-called *response amplitude operator*, which is the transfer function for the linear ship model and readily available for most ship designs. An external forcing by a frequency component

$$A_k \cos(\omega_k t + \epsilon_k) \tag{1.29}$$

will result in a steady-state response

$$A_k |H(\omega_k)| \cos(\omega_k t + \delta(\omega_k) + \epsilon_k), \tag{1.30}$$

where $|H(\omega_k)|$ is the frequency-dependent response per unit wave amplitude and $\delta(\omega_k)$ is a phase angle [7].

The beauty of linear equations of motion is that the response amplitude operators corresponding to distinct frequency components can be simply superposed.

1.5.5 Risk quantification

As the dynamical system of the ship is nonlinear and nonautonomous, it is almost impossible[3] to find analytical results for its stability under random sea states, even if the stochastic properties of the latter are known.

In practise one is interested in a quantification of the *risk* of occurence of parametric roll resonance. Assuming that the *loss* is total when the roll angle rises above a certain threshold (e.g., 20 degrees, as then containers are likely to fall off from a container ship), simplifies the problem enormously. Assuming that such an event is unlikely, it can be mathematically modelled as a homogeneous Poisson process with intensity $\lambda > 0$, whereby

[3]The work of Farrell et al. on generalized stability theory [8] could offer a possible way to deal with these problems.

λ represents the expected number of events per unit time. Let N be the number of such events during a fixed time of operation T of the ship, then

$$\text{pr}(\text{loss}) = \text{pr}(N \geq 1) = 1 - e^{-\lambda T} \tag{1.31}$$

is the desired risk of loss.

We can estimate λ from numerical simulations, resetting the simulation to randomly chosen initial conditions whenever the roll angle exceeds the threshold, and counting the occurences of these events. However, it is more advantageous to consider the mean waiting time τ before such an event occurs. The interarrival times in the Poission process are exponentally distributed, and the relation between τ and λ is simply $\tau = 1/\lambda$.

Since expectation is a linear operation, it is clear that one only needs to estimate λ for each stationary sea state of interest, characterized by a single wave spectrum, and can use the above-mentioned prediction graphs (or a hidden Markov model, for a more accurate estimate) to combine risks for different sea states into a global *operational risk*.

Alternatively, one could use extreme value statistics [3], e.g., by fitting the maxima of roll angles observed in a simulation to an extreme value distribution. This would be particularly useful to quantify the operational risk in conditions where resonance might be expected, but occurs too seldomly to calculate mean waiting times.

1.5.6 Numerical results

Stochastic waves generated by (24) act upon the vessel. Therefore, *effective* transfer functions have to be used with the appropriate parameters such as damping constants and added masses. The values of these constants are weighted averages over 500 frequency components of a random Pierson-Moskowitz spectrum, weighted according to spectral power. Figure 8 shows an example for wind speed $U_{10} = 21$ m/s. The wave time series has been generated by the method of Rice (Section 1.5.2) from the random spectrum shown on the right of it. For comparison, also the corresponding analytical Pierson-Moskowitz spectrum is depicted. Such a wind speed (Beaufort scale 8.3) results in a rough sea with a significant wave height of 5.9 meters. Although the ship heaves with a comparable amplitude, there is no parametric rolling. In contrast to that, increasing the wind speed to $U_{10} = 22$ m/s leads to the situation in Figure 9. For such wind conditions (Beaufort scale 8.6) the significant wave height amounts to 7.8 meters and stochastic resonance occurs.

Finally, Figure 9 shows a risk function for the occurrence of parametric roll under Pierson-Moskowitz spectra. The data is based on 100 simulations that stopped either when

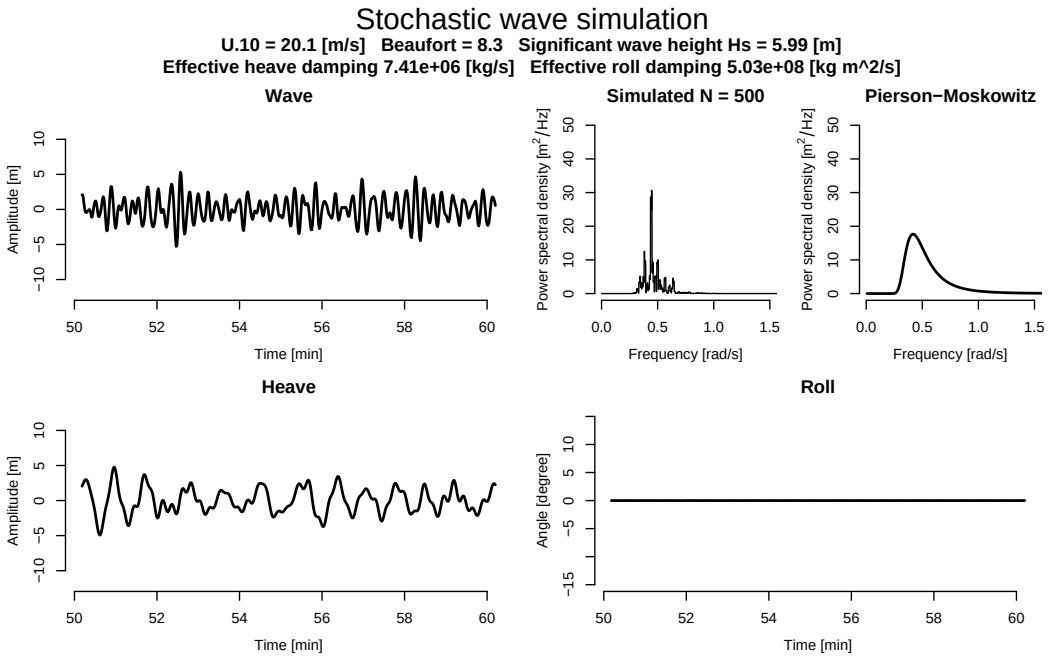

Figure 1.8: Example of stochastic simulation. Top left: Time series of wave heights generated by the method of Rice [22]. Top right: Spectrum of generated wave heights and the corresponding analytical Pierson-Moskowitz spectrum. Bottom left: Heave motion of the ship in the pendulum-spring model under this wave forcing. Bottom right: Roll angle. Results shown are for a wind speed of 8.3 Beaufort. Note that roll resonance does not occur.

the roll angle exceeded 20 degrees, or after 10 hours simulated time, if no resonance could be observed during that time. Each simulation is based on a different randomly generated spectrum, starting from small initial conditions (0.01 meter heave and 1.0 degree roll) and the mean waiting time was recorded. The finite simulation time introduces a bias, since some waiting times are *censored*meaning that only a lower bound is known. For simplicity, this effect was corrected by subtracting the minimal possible rate (one event per 10 hours) from the corresponding mean rate $\bar{\lambda} = 1/\bar{\tau}$. The minimal observed fraction of events (three events, for $U_{10} = 18$ m/s) was added to the values obtained from Eq.(1.31) and the result is the approximate operational risk shown in Figure 10.

Note that this example just illustrates the methodology. In particular, the simple pendulum model is difficult to use with stochastic waves, since the equivalent pendulum length L, and the damping and stiffness constants depend explicitly on the wave frequency. Using effective values as here is only a rough approximation and serves to illustrate the

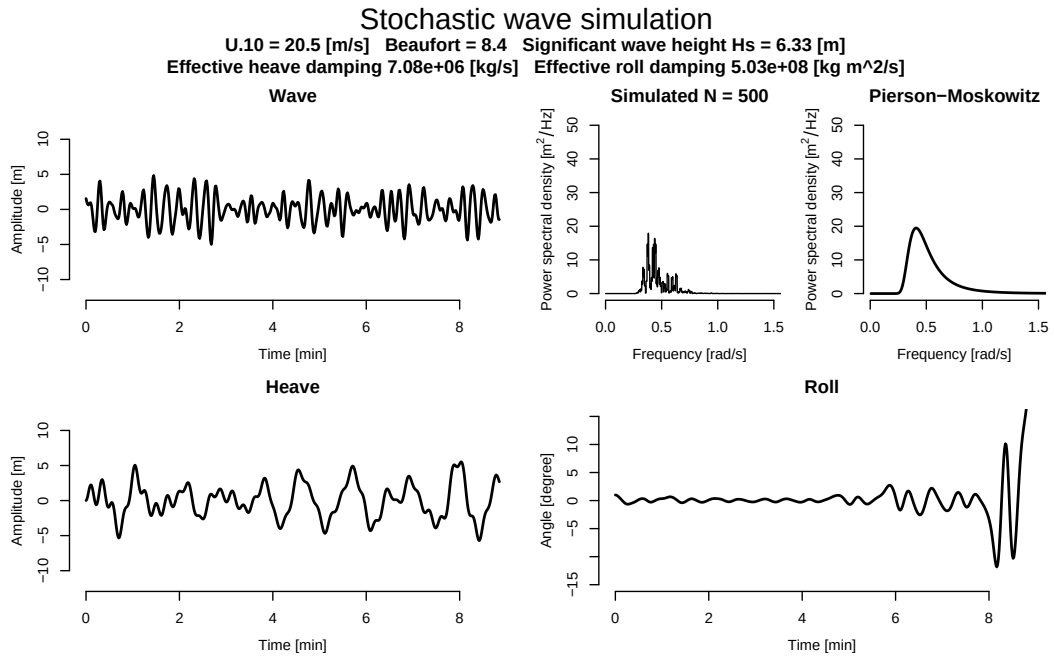

Figure 1.9: Example of a stochastic simulation as in Figure 1.7. Results shown are for a slightly larger wind speed of 8.4 Beaufort. Now roll resonance does occur.

procedure; in reality, one would need to use a different dynamical model that accounts for these frequency-dependent effects.

1.6 Recommendations for Future Investigations

In the study of the rolling motion of a vessel as it is forced by waves we discern two mayor directions:

- Analysis of the nonlinear dynamical equations stressing the interaction of different modes (heave, pitch, and roll) and the possibility of (parametric) resonance.
- The stochastic description of ocean waves and their effect upon the meta- centric height change of the vessel.

Of course also other aspects of the problem play a role such as the computation of dynamical parameters from the design and loading of a vessel. In our search for causes of extreme roll in heavy seas these studies play an ancillary role. For each of the two directions indicated above we bring up ideas for further research which possibly may lead

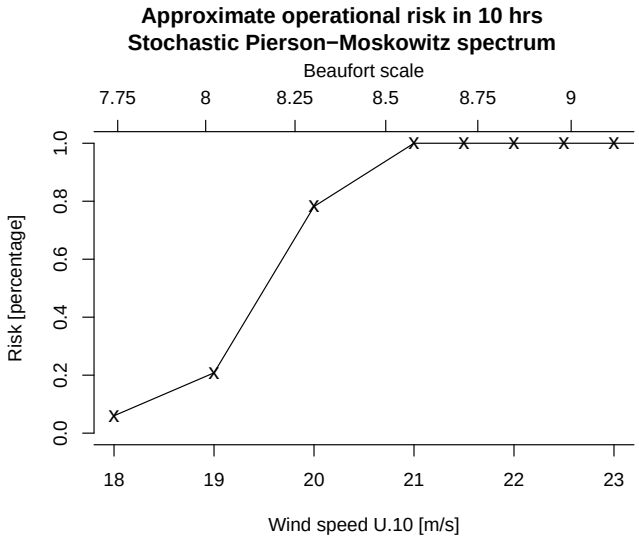

Figure 1.10: Risk of parametric roll (angle larger than 20 degrees) for waves having a Pierson-Moskowitz spectrum at different wind speeds.

to a better understanding of the problem of extreme roll resulting in solutions remedying this unwanted phenomenon.

From the point of view of modeling ship motion by nonlinear differential equations 3-DOF models fully cover the motion of a ship if we consider it as a point mass and ignore displacement in the horizontal plane. In Section 4 we discussed one such a model being an alternative to a model analyzed in [30]. Of course the number of degrees of freedom can easily be extended, see Korvin- Kroukovsky [15]. However, we have the idea that for a 3-DOF system not yet all possible causes of extreme roll have been identified. In particular the fact that this phenomenon occurs in combination with waves having a wavelength in the order of the length of the vessel suggests that forcing is not only through the heave mode of the system but also through pitch. If this component of the forcing is present, then in a short time a large amount of energy can be transferred from the waves to the ship motion, as we already pointed out in Section 4. If in addition the forcing frequency is close to the natural pitch frequency, the roll amplitude may increase even more. It is worth to have this analyzed. Furthermore in Section 4 it is pointed out that the solution of systems with small DOF's can be approximated by series expansions based on series solutions of the Mathieu equation.

In Section 5 it is suggested that the most efficient representation of stochastic waves is that of a linear ARMA-system see Eq.(1.27). This approach can also be used in the

stochastic description of roll forced by waves as presented by Dunwoody [5]. He formulates the Langevin equations for roll amplitude and phase, see formula's (5-6)of [5]. Instead of solving the equation for the corresponding Fokker-Planck equation one can formulate the exit problem for the amplitude exceeding some large value. It yields the expected value for the time needed to arrive at this value, see [11]. This exit time quantifies the risk of extreme roll better than the mean amplitude.

Bibliography

[1] Arnold, L., Chueshov, I., and Ochs, G. (2004). Stability and capsizing of ships in random sea — a survey. *Nonlinear Dynamics*, 36:135–179.

[2] Boyd, J. (2001). *Chebyshev and Fourier Spectral Methods*. Dover Publications.

[3] Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag.

[4] DelBalzo, D., Schultz, J., and Earle, M. (2003). Stochastic time-series simulation of wave parameters using ship observations. *Ocean Engineering*, 30:1417–1432.

[5] Dunwoody, A. (1989). Roll of a ship in astern seas response to gm fluctuations. *Journal of Ship Research*, 33(4):84–290.

[6] Eissa, M., El-Sera, S., El-Sheikh, M., and Sayeda, M. (2003). Stability and primary simultaneous resonance of harmonically excited non-linear spring pendulum system. *Appl. Math and Computation*, 145:421–442.

[7] Faltinsen, O. M. (1993). *Sea Loads on Ships and Offshore Structures*. Cambridge University Press.

[8] Farrell, B. and Ioannou, P. (1996). Generalized stability theory. Part II: Nonautonomous operators. *Journal of the Atmospheric Sciences*, 53:2041–2053.

[9] France, W., Levadou, M., Treakle, T., Paulling, J., Michel, R., and Moore, C. (2003). An investigation of head-sea parametric rolling and its influence on container lashing systems. *Marine Technology and SNAME News*, 40(1):1–19.

[10] Francescutto, A. and Bulian, G. (2004). Nonlinear and stochastic aspects of parametric rolling modelling. *Marine Technology*, 41(2):74–81.

[11] Grasman, J. and van Herwaarden, O. (1999). *Asymptotic Methods for the Fokker-Planck Equation and the Exit Problem in Applications*. Springer-Verlag, Berlin Heidelberg.

[12] Hochstadt, H. (1986). *The Functions of Mathematical Physics*. Dover Publications.

[13] Holden, C., Galeazzi, R., Rodrìguez, C., Perez, T., Fossen, T., Blanke, M., and de Almeida Santos Neves, M. (2007). Nonlinear container ship model for the study of parametric roll resonance. *Modeling, Identification and Control*, 28(4):87–103.

[14] Jeffrey, A. and Zwillinger, D. (2007). *Gradshteyn and Ryzhiks Table of Integrals, Series and Products*. Academic Press.

[15] Korvin-Kroukovsky, B. (1961). *Theory of Seakeeping*. Soc. Naval Architects and Marine Engineers, New York.

[16] Kreuzer, E. and Sichermann, W. (2006). The effect of sea irregularities on ship rolling. *Computing in Science & Engineering*, 8(3):26–34.

[17] Levadou, M. and van't Veer, R. (2006). Parametric roll and ship design. In *Proceedings of the 9th International Conference on Stability of Ships and Ocean Vehicles*.

[18] Massel, S. (1996). *Ocean Surface Waves: Their Physics and Prediction*. World Scientific Publishing.

[19] Ochi, M. (2005). *Ocean Waves: The Stochastic Approach*. Cambridge University Press.

[20] Podgorski, K., Rychlik, I., and Machado, U. (2000). Exact distributions for apparent waves in irregular seas. *Ocean Engineering*, 27:9791016.

[21] Poulin, F. and Flierl, G. (2008). The stochastic Mathieu's equation. *Proceedings of the Royal Society A*, 464:1885–1904.

[22] Rice, S. (1944). Mathematical analysis of random noise. *Bell System Technical Journal*, 23:282–332.

[23] Santos-Neves, M., Pérez, N., and Lorca, O. (2003). Analysis of roll motion and stability of a fishing vessel in head seas. *Ocean Engineering*, 30:921–935.

[24] Santos-Neves, M. and Rodríguez, C. (2007). Influence of non-linearities on the limits of stability of ships rolling in head seas. *Ocean Engineering*, 34:1618–1630.

[25] Scheffner, N. and Borgman, L. (1992). Stochastic time-series representation of wave data. *J. Waterway Port Coast. Ocean Eng.*, 118(4):337–351.

[26] Shin, Y., Belenkey, V., Pauling, J., Weems, K., and Lin, W. (2004). Criteria for parametric roll of large containerships in longitudinal seas. *Transactions – Society of Naval Architects and Marine Engineers*, 112:14–47.

[27] Spanos, P. and Mignolet, M. (1986). Z-transform modelling of P-M wave spectrum. *Journal of Engineering Mechanics*, 112:745–759.

[28] Spyrou, K. (2000). Designing against parametric instability in following seas. *Ocean Engineering*, 27:625–653.

[29] Sun, T. and Chaika, M. (1997). On simulation of a Gaussian stationary process. *Journal of Time Series Analysis*, 18:79–93.

[30] Tondl, A., Ruijgrok, T., Verhulst, F., and Nabergoj, R. (2000). *Autoparametric Resonance in Mechanical Systems*. Cambridge University Press, Cambridge.

Chapter 2
An objective method to associate local weather extremes with characteristic circulation stuctures

Corinna Ziemer [1] Katarzyna Marczyska [2] Anna Szczepańska [2] Joanna Zyprych [2] Jos Hageman [3] Timo Doeswijk [3]

abstract:
In this paper we give methods to find characteristic circulation patterns which are connected to local extreme temperature anomalies. Two data reduction techniques are applied: Legendre polynomial fitting and watershedding. For polynomial fitting a clear trend is found with respect to local temperatures. However, the trend is not distinctive enough to give clear answers on the type of circulation patterns belonging to local extremes. The main advantage of watershedding is that the physical properties of the circulation patterns are retained while the dimension of the data is largely reduced. Expert knowledge, however, is needed to model these main features as predictors.
KEYWORDS: *circulation pattern, extreme temperature, watershedding, Legendre polynomial.*

[1]Centre for Industrial Mathematics, University of Bremen, Germany
[2]Poznan University of Life Sciencies, Poland
[3]Biometris, Wageningen University, The Netherlands

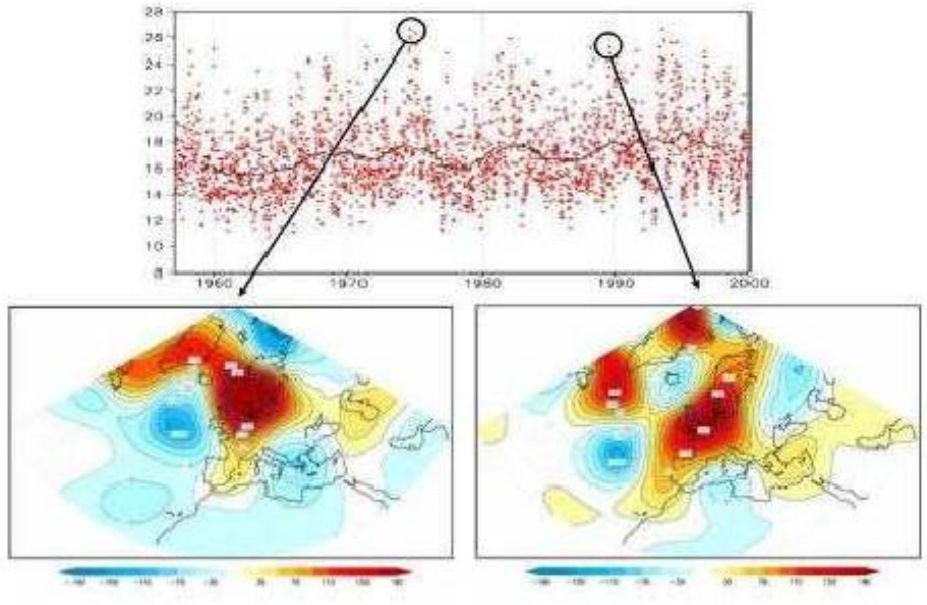

Figure 2.1: Upper panel: time series of maximum daily temperatures of the months july and august in the years 1958-2000; left panel - circulation pattern that relates to a local extreme temperature in 1975; right panel - circulation pattern that relates to a local extreme temperature in 1983

2.1 Introduction

Meteorological events such as severe storms, heavy rains, cold surges or drought which occur locally are usually connected to circulation structures of much larger scale in the atmosphere. In this paper we study the relation between local extreme temperatures and circulation patterns in the atmosphere. In meteorology it can be observed that extreme temperatures (temperature anomalies) appear for several different states of atmosphere circulation. For example, in 1975 a high pressure anomaly was located above Scandinavia leading to advection of warm, dry, continental air into the Netherlands by easterly winds and local extreme heat. Eight years later, a high pressure anomaly was located right above the Netherlands with clear skies, no wind, an abundance of sunshine and as a result, extreme high temperatures. Figure **??** shows these two different circulation patterns which caused local temperature extremes.

The concept of weather regimes was considered by Michelangeli et al. [1]. The authors compared two different definitions of weather regimes. The first definition treats weather regimes as the states of the atmosphere with the highest probability of occurrence. In the second case weather regimes are defined as the states for which large-scale motion is stationary in the statistical sense. The authors applied these methods on the same dataset and they showed that these methods give the same number of weather regimes - four over the Atlantic sector and three over the Pacific sector. They observed that the patterns differ significantly and the investigation of the tendency, or drift, of clusters shows that recurrent flows have a systematic slow evolution, explaining this difference. The patterns are in agreement with the one obtained from previous studies, but their number differ. Panja and Selten [2] presented a new method to optimally link local weather extremes to large scale atmospheric circulation structures. This method objectively identifies, in a robust manner, the different circulation patterns that favor the occurrence of local weather extremes and is based on considering linear combinations of the dominant Empirical Orthogonal Functions that maximize a suitable statistical quantity. Moreover, Salameh and Dobrinski [3] related the occurrence of extreme events (in terms of temperature, precipitation and wind speed) to weather regimes. They evaluated the uncertainty associated with North Atlantic weather regime clustering with the re-analyses data set and its impact on the relationship between weather regimes and extreme events over and around the North Atlantic.

The aim of this paper is to find a method that identifies pressure patterns which lead to extreme values of temperature in one fixed point and to work out a method to predict when local temperature extremes occur.

To analyze circulation patterns in relation to extreme temperature anomalies we use date obtained from the ERA-40 reanalysis dataset. The data of the circulation patterns contained the pressure field for 1372 grid points which are arranged on 20^0 N - 90^0 N latitude and 60^0 W - 60^0 E longitude ($2.5^0 \times 2.5^0$ latitude-longitude grid). A time series of daily circulation patterns were available for July and August of the years 1958-2000 (all together 43 years and 2666 time points in total). The local temperature was taken at the center of the Netherlands (52.5^oN, 5^oE). The 5 per cent most extreme (positive) anomalies were taken as extreme values. In this way 133 circulation patterns were connected to local extreme temperatures. One of the main issues to be dealt with is data reduction. Two methods are used and explored: 1) Legendre polynomial fitting and 2) watershedding.

2.2 Modelling approaches

The discrete pressure field contains $28 \times 49 = 1372$ data points. Using this raw data as an input, the model would have to base its decision (whether the pattern belonged to an extreme temperature) on a huge amount of data. Directly using these data causes problems such as ill-conditioned matrices because of high correlations between grid points and long computation times. Therefore, first the available data is reduced while retaining the information before processing it. That can be achieved by fitting Legendre polynomials to the pressure distribution or by the watershedding technique. As a second step, a connection between the global weather situation and temperature has to be found. In this work, the method of linking the pressure anomaly patterns with the local temperature extremes uses empirical data. In order to see whether the method gives correct results, firstly the patterns belonging to the most extreme temperatures are picked out from the empirical data. Then this set is split into reference patterns (used for calibrating the method) and validation patterns (used to check whether the method works correctly).

2.2.1 Data Reduction

We have 133 patterns with extreme temperature in one fixed point. Each circulation pattern can be described as 28×49 pressure table where rows represents latitude and columns longitude. Mathematically speaking, we are looking for a function from the set of pressure patterns $\{Z(t_i)\}$ to a set of characteristic parameters $\{C(t_i)\}$, where t_i indicates the point in time at which the pressure measurement was taken. When choosing the dimension of the space of characteristic parameters much smaller than the dimension of the patterns' space, we can store approximately the same amount of information with much less data.

Legendre Polynomials

The general idea in this approach is a known result from Linear Algebra: a function f belonging to a finite dimensional space of functions X_N (e.g. all polynomials of order N) can be represented by a linear combination of basis functions $P_l \in X_N,\ l = 1, \ldots, N,\quad N \in \mathbb{N}$:

$$f(x) = \sum_{l=1}^{N} \alpha_l P_l(x) \qquad \forall x \in \mathrm{Dom}(f) \tag{2.1}$$

where $\mathrm{Dom}(f)$ denotes the domain of f, *i.e.* all x for which f is defined. In this work, the function f is the pressure anomaly distribution along a line in latitudinal or longitudinal direction. For the purpose of data reduction, the function is represented by a linear

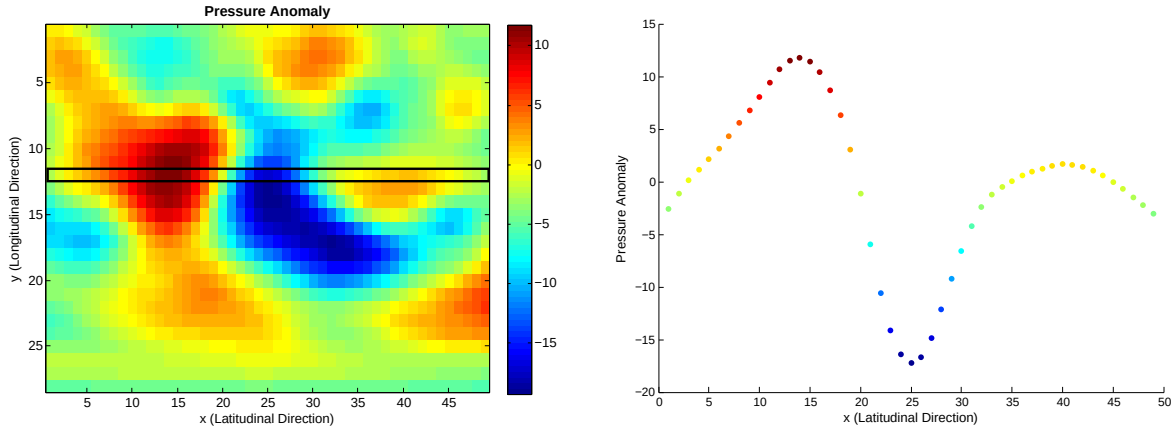

Figure 2.2: Discrete pressure map (left). The rectangle indicates the sample line in latitudinal direction (right), along which the polynomial is fitted.

combination of basis elements P_l. Then only the coefficients α_l associated with the basis elements are kept. In nature, the pressure distribution is smooth in any direction, but contrary to our assumptions above, it is not belonging to a finite dimensional space of functions. Consequently, we can only try to approximate it by a function $f \in X_N$. The order of accuracy of such an approximation increases with the number of basis functions N. In the present situation, the given data contains the pressure only at discrete points on a grid. For fitting a function along a row or column of the pressure distribution, Eq. (2.1) has to hold for each of the grid points along that line. This leads to a system of equations from which the coefficients can be computed as follows: Let $Z(t) \in \mathbb{R}^{28\times 49}$ be the discrete pressure distribution at time t and let $Z_{i,j}(t)$ indicate the evaluation of the pressure field at the grid point (x_i, y_j), with $i = 1, \ldots, 28$ and $j = 1, \ldots, 49$. Then we can solve for $\alpha, \beta \in \mathbb{R}^N$

$$\begin{pmatrix} P_1(x_1) & P_2(x_1) & \ldots & P_N(x_1) \\ P_1(x_2) & P_2(x_2) & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ P_1(x_{49}) & \ldots & \ldots & P_N(x_{49}) \end{pmatrix} \begin{pmatrix} \alpha_1^i(t) \\ \alpha_2^i(t) \\ \vdots \\ \alpha_N^i(t) \end{pmatrix} = \begin{pmatrix} Z_{i,1}(t) \\ Z_{i,2}(t) \\ \vdots \\ Z_{i,49}(t) \end{pmatrix}$$

or in shorthand

$$\begin{aligned} &(P_l(x_k))_{kl} \left(\alpha_l^i\right)_l = (Z_{i,k})_k \\ \Longleftrightarrow: \quad &P^h \vec{\alpha}^i = \vec{Z}^i \end{aligned} \tag{2.2}$$

for fitting a function to each row i. Likewise, we can set up a linear equation system for fitting a function to each column j, namely

$$\begin{aligned} &\left(P_l(y_\kappa)\right)_{\kappa l} \left(\beta_l^j\right)_l = \left(Z_{\kappa,j}\right)_\kappa \\ \Longleftrightarrow: \qquad &P^v \vec{\beta}^j = \vec{Z}^j \end{aligned} \tag{2.3}$$

where $k = 1, \ldots, 49$ and $\kappa = 1, \ldots, 28$ denote the number of columns or rows, respectively, and $l = 1, \ldots, N$ denotes the basis polynomials.

Although it is possible to compute the coefficients with Eqs. (2.2) or (2.3), resp., the question of choosing the basis polynomials still remains. As the pressure distribution along a line is continuous, it can be approximated by polynomials. Consequently, the most obvious choice would be the monomial basis $\{1, x, x^2, \ldots\}$, i.e. $P_l(x) = x^{l-1},\ l \in \mathbb{N}$. The resulting matrix would be the so called Vandermonde matrix. However, it is not suitable for numerical purposes due to its very bad condition number. Choosing Legendre polynomials as a basis avoids those difficulties. Legendre polynomials can be obtained by orthonormalization of the monomial basis on the interval $[-1, 1]$, subject to the condition that $P_l(1) = 1$ (cf. Fig. 2.3). We obtain:

$$\begin{aligned} L_1(x) &= 1, \\ L_2(x) &= x, \\ L_3(x) &= \frac{1}{2}\left(3x^2 - 1\right), \\ &\vdots \\ L_{N-1}(x) &= \frac{1}{2^N N!} \frac{d^N}{dx^N} \left[(x^2 - 1)^N\right]. \end{aligned}$$

The Legendre polynomials are orthonormal only on the interval $[-1, 1]$. Thus the grid is implicitly assumed to be transformed on $[-1, 1]^2$. For data reduction purposes the number of basis elements has to be chosen much smaller than the number of grid points. Thus (2.2) and (2.3) are overdetermined systems of equations and there does not exist an exact solution. This means that the linear combination of basis elements cannot represent the discrete pressure distribution exactly. However, we want the function to fit with an error as small as possible. The resulting coefficients can be obtained by solving (2.2) and (2.3) by linear regression:

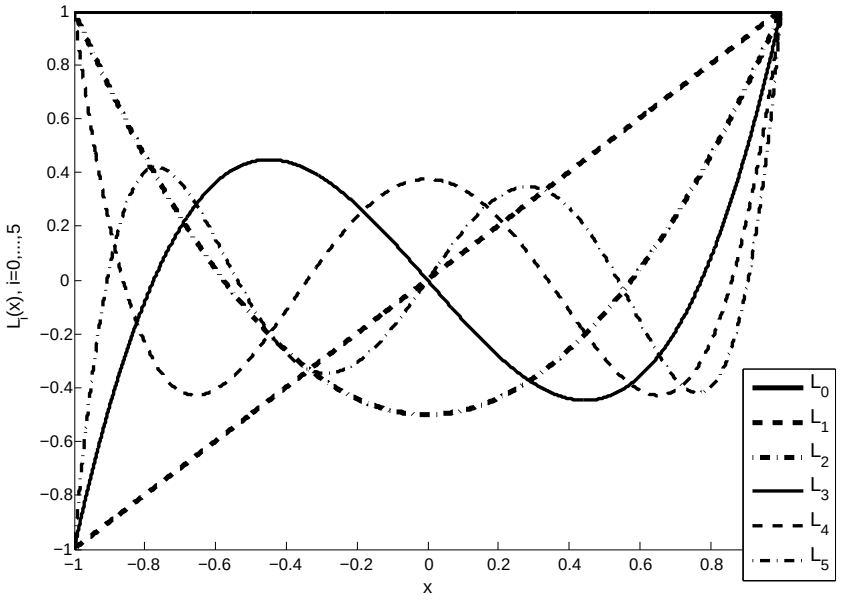

Figure 2.3: Legendre basis polynomials up to fifth order.

$$(L^h)^T L^h \vec{\alpha}^i = (L^h)^T \vec{Z}^i$$
$$(L^v)^T L^v \vec{\beta}^j = (L^v)^T \vec{Z}^j$$

Here the matrices L^h, L^v denote the analogues to the matrices P^h, P^v in Eq. (2.2) or (2.3), where the polynomials used for the entries are the Legendre polynomials specified before. In order to classify a given pattern with less data, we then collect all coefficient vectors of the pattern:

$$C = \{\vec{\alpha}^1, \ldots, \vec{\alpha}^{28}, \vec{\beta}^1, \ldots, \vec{\beta}^{49}\} \tag{2.4}$$

For a further reduction of data, a function was fitted only to each second row and column of the discrete pressure field. Moreover, the coefficients belonging to the first two Legendre polynomials were neglected. This implies that neither the bias nor the tilt of the pressure distribution are taken into consideration. The reasoning behind this is that a pattern of pressure anomaly is to a greater extent defined by its spatial oscillations than by its offset or slope.

Two-dimensional polynomials

A further reduction based on polynomials can be established by fitting a two-dimensional polynomial. Eqn. (2.1) is extended to:

$$f(x,y) = \sum_{l=0}^{N} \sum_{m=0}^{N-l} \alpha_{lm} P_{lm}(x,y) \qquad \forall x, y \in \mathrm{Dom}(f) \tag{2.5}$$

In this approach the function f is the pressure anomaly distribution of the surface in which x represents the longitudinal direction and y the latitudinal direction. Following section 2.2.1 orthonormalization on the domain [-1,1] for both x and y is strongly preferred. However, because of the restricted amount of time only the monomial basis, *i.e.* $\{1, x, y, x^2, xy, y^2, x^3, \dots\}$ was implemented. In analogy to eqns. (2.3) and (2.2) we can set up the system:

$$\begin{pmatrix} P_{00}(x_1,y_1) & P_{01}(x_1,y_1) & \dots & P_{0N}(x_1,y_1) & P_{10}(x_1,y_1) & \dots & P_{N0}(x_1,y_1) \\ P_{00}(x_2,y_1) & P_{01}(x_2,y_1) & & & & & \vdots \\ \vdots & \vdots & \ddots & & & & \vdots \\ P_{00}(x_{49},y_1) & P_{01}(x_{49},y_1) & & \ddots & & & \vdots \\ P_{00}(x_2,y_2) & P_{01}(x_2,y_2) & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & \vdots \\ P_{00}(x_{49},y_{28}) & P_{01}(x_{49},y_{28}) & \dots & P_{0N}(x_{49},y_{28}) & P_{10}(x_{49},y_{28}) & \dots & P_{N0}(x_{49},y_{28}) \end{pmatrix} \begin{pmatrix} \alpha_{00}(t) \\ \alpha_{01}(t) \\ \vdots \\ \alpha_{0N}(t) \\ \alpha_{10}(t) \\ \vdots \\ \alpha_{NN}(t) \end{pmatrix} = \begin{pmatrix} Z_{1,1}(t) \\ Z_{2,1}(t) \\ \vdots \\ Z_{49,1}(t) \\ Z_{1,2}(t) \\ \vdots \\ Z_{49,28}(t) \end{pmatrix}$$

Generally, a seventh order polynomial for the two-dimensional case gives a good reconstruction of the surface. The number of parameters to be estimated are in this case $1+2+3+4+5+6+7 = 28$. Figure 2.4 shows the approximation for a 9th order polynomial function.

Watershedding

Watersheds were first used in topography. The main idea consists of geographical regions that are divided in so-called catchment basins and the division between two regions is called the watershed line. Suppose a droplet of water falls down on a surface. This droplet would run down to the lowest point of the region. Adding more droplets would immerse the surface to a lake. The lake will continue to fill until this lake start to flood into a neighbor valley. The line where two valleys, the so-called catchment basins, come together are called the watershed line. Apart from topography the watershedding transform is frequently used in image processing. The pressure anomalies considered in this report can also be viewed

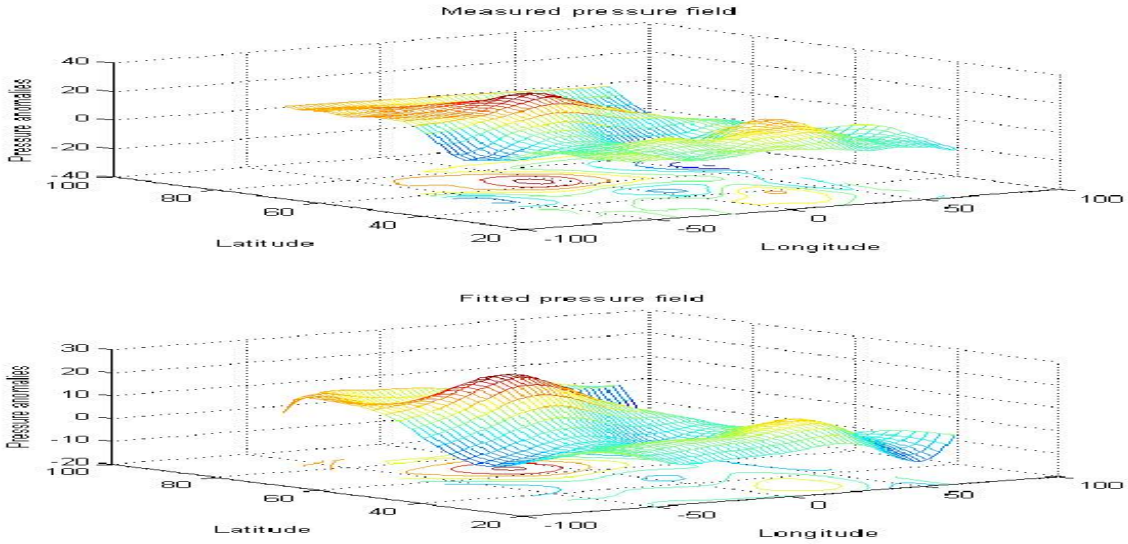

Figure 2.4: Pressure field approximation with an ninth order two-dimensional polynomial function

as a topographic surface. The watershedding approach is used as a tool for data reduction by obtaining areas of high and low anomalies.

Many algorithms exist that are based on the watershedding principle. The Matlab® implementation which is used in this study is based upon the paper by Vincent and Soille [4]. The algorithm consist of two steps: sorting and flooding. Let $Z(t) \in \mathbb{R}^{28\times 49}$ be the discrete pressure distribution at time t and let $Z_{i,j}(t)$ indicate the evaluation of the pressure field at the grid point (x_i, y_j), with $i = 1, \ldots, 28$ and $j = 1, \ldots, 49$. Because it is assumed that high pressure anomalies are equally important as low pressure anomalies, pressure distribution is transformed such that $\tilde{Z}(t) = -|Z(t)|$. In this way, high pressure anomalies are regarded as catchment basins. For each time point the watershed transform is applied to $\tilde{Z}(t)$ and result in the watershed matrix $W(t)$. Further details and an exact description of the algorithm can be found in [4]. An example of a watershed transform is given in figure 2.5. From the watershed transform $W(t)$ information from the catchment basins is extracted such as the center and total area. By selecting the p most important basins, *i.e.* those with the lowest watershed index or in other words with the largest pressure anomaly, the number of variables is reduced dramatically.

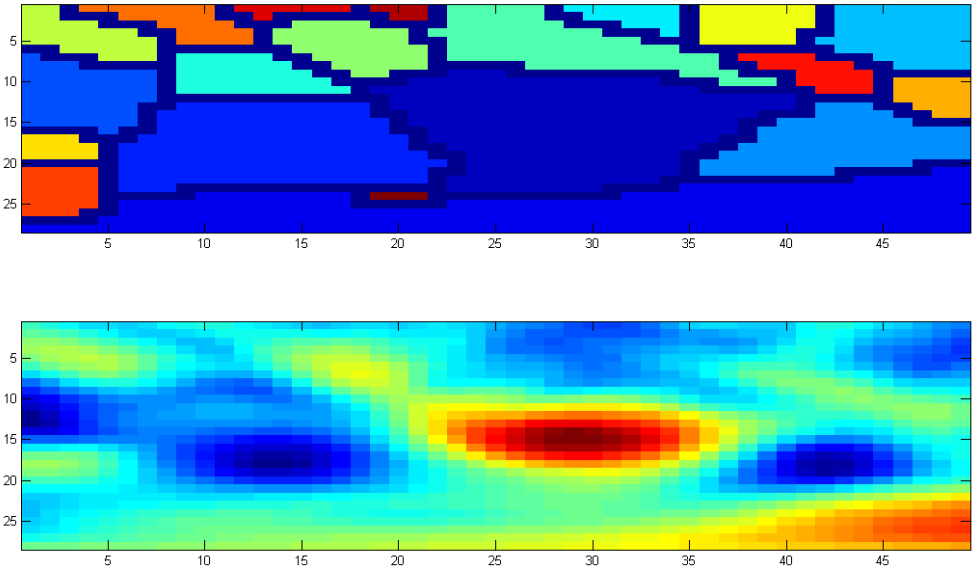

Figure 2.5: A pressure anomaly field and its watershed transform. The x and y axis are the indices are the i and j indices representing the lattitude/longitude of the anomaly field

2.2.2 Data Processing and Evaluation

Legendre Polynomials

First of all, a temperature threshold T_{ext} is defined above which a temperature shall be regarded as being extreme. Throughout the simulations, the topmost five per cent of temperature observations were regarded as extreme. As mentioned above, the set of patterns associated to an 'extreme' temperature is then arbitrarily divided into two disjoint subsets: one for calibrating and one for validating the model:

$$I = \{t \mid T(t) \geq T_{\text{ext}}\} := I_{\text{cal}} \dot{\cup} I_{\text{val}}.$$

That is to say: the model is set up with the patterns associated to I_{cal}, and with the patterns belonging to I_{val}, the correctness of the results is checked. Subsequently, for a specific time point t_0 an error $\epsilon(t_0)$ is assigned to each pattern $Z(t_0)$. This error indicates how 'far' that pattern is away from the patterns $\{Z(t) \mid t \in I_{\text{cal}}\}$:

$$\epsilon(t_0) := \min_{t \in I_{\text{cal}}} \|C(t_0) - C(t)\| \tag{2.6}$$

where $C(s)$ denotes the set of characteristic parameters for a pattern $Z(s)$. The error

can be measured in any suitable norm, for example the L^2 vector norm.

Two-dimensional polynomials

Although the used data reduction reduction technique is based on the same idea as the Legendre polynomials, a different evaluation technique was performed for the two-dimensional polynomial. Here, the evaluation is based on linear regression.

As in section 2.2.2 the extremes are divided in a calibration and validation subset. In addition, the non-extreme patterns are also divided in a calibration and validation subset. Therefore, the calibration and validation data sets need to contain an equal amount of the interesting extremes compared to the much larger amount of non-extreme data.

$$J = \{t \mid T(t) < T_{\text{ext}}\} := J_{\text{cal}} \dot{\cup} J_{\text{val}}.$$

Hence, the calibration and validation data sets are formulated as:

$$\Gamma_{\text{cal}} = I_{\text{cal}} \cup J_{\text{cal}}$$
$$\Gamma_{\text{val}} = I_{\text{val}} \cup J_{\text{val}}$$

A linear model to predict the local temperature based on the polynomial parameters is proposed:

$$T_{\text{local}}(t) = \alpha(t)\gamma + e(t)$$

with α the time dependent polynomial parameters that define the circulation pattern, γ the model parameters and $e(t)$ the error term. Based on the calibration data set the model parameters γ are estimated by linear regression.

$$\gamma = A^{+} T_{\text{local}}$$

with A^{+} the pseudo-inverse of $[\alpha(t_0), \alpha(t_1), \ldots, \alpha(t_n]^T$. Both calibration and validation data sets are used to estimate the local temperature by:

$$\hat{T}_{\text{local}}(t) = \alpha(t)\hat{\gamma}$$

Watershedding

Before the catchment basins are projected onto local extremes meteorological information must be incorporated. The local extreme is a nonlinear function of the catchment basin

and perhaps of the interaction between catchment basins. A first approach was done by constructing a function based on distance, anomaly and area of the catchment basin. Let $\vec{B}_i(t)$ a vector with the variables of catchment basin i at time t extracted from the watershed transform $W(t)$. The local temperature anomaly can be modeled by:

$$T_a(t) = \sum_{i=1}^{p} \delta_i f(\vec{B}_i) + e(t) \tag{2.7}$$

where $f(\vec{B}_i)$ is a nonlinear function with variables from catchment basin i, δ the parameter vector and $e(t)$ the error term. The model is linear in its parameters and, hence, these parameters δ are estimated with an ordinary least squares approach, *i.e.* $\min_\delta ||e(t)||_2 \; \forall t$. Because the system now is largely overdetermined, division into a calibration and validation is not needed. Results are evaluated by comparing the estimated and measured local temperatures of the total data set. Local extremes are part of the data and verification is possible.

2.3 Results

2.3.1 Legendre Polynomials

For testing the L^2-norm comparison method, the set of patterns belonging to an extreme temperature was divided into calibration and validation patterns, analogously to the explanations above. The set of extreme patterns is divided in several fashions:

half: every second pattern is used for calibration, the other half used for testing/validation.

thirds: every third pattern is used for validations, so that 66% of patterns are used for calibration.

rand: each extreme pattern gets assigned a random number from a uniform distribution on the interval [0,1]. For validation, only the patterns with a random number higher than 2/3 are taken.

Concerning the order of the fitted polynomial, an integer value around 4 was chosen. This is sensible, as the pressure distribution along one direction usually contains $n = 2$ to $n = 4$ major high or low pressure areas. As these areas should correspond to the extremes of the fitted polynomial, we need a polynomial degree of $n - 1$. The error was measured in the standard L^2 vector norm.

As can be seen from any of the plots in Fig. 2.6, there is no clear visible distinction between the two groups 'errors of non-extreme patterns' (plotted in red) and 'errors of validation patterns' (plotted in green). What can be seen, however, is the (anticipated) tendency of the 'green mean' to be below the 'red mean', i.e. that the patterns belonging to an extreme temperature have on average a lower error than those arbitrary, non-extreme patterns.

Choosing a higher order of polynomial does improve the distinction between the two groups. Nevertheless, taking a too high order of polynomial increases the danger of unnatural oscillatory behavior when fitting the polynomial. Concerning the change of the fashion of partitioning the validation set, the following can be observed: The two groups are the more distinct the more patterns for calibration are used. This is a result one would also expect by common sense.

2.3.2 Two-dimensional polynomial

In figure 2.7 a linear trend is clearly observable. However, it can be clearly seen that the linear trend bends off in the top right corner, just before the crossing horizontal and vertical lines. These are the lines that distinguish the ordinary values from the extremes. This is exactly the part in which we are interested. When we look at the histogram (see figure 2.8 of how the local temperatures are distributed) it is clear that it is right tailed. In this tail the extremes are defined. Unfortunately, this tail seems hard to predict. Several procedures have been tried to improve the results, such as a log-transformation of the local temperatures and a neural network approach to account for nonlinearities. These procedures did not improve the results visibly (results not shown).

2.3.3 Watershedding

In this specific example five watersheds are taken for each circulation pattern. The function $f(B)$ from eqn. (2.7) is given by:

$$f(B_i) = \frac{\sin(\alpha) \cdot P \cdot A}{dist} + \frac{\cos\alpha \cdot P \cdot A}{dist}$$

with α the angle between the location of the local measurement and the location of the center of the watershed; $dist$ the distance between the center of the watershed and the local measurement; P the maximum pressure anomaly in the watershed; and A the total area

of the watershed. In figure 2.9 a slight linear trend is visible but it is clearly not sufficient. Local extremes cannot be predicted in the current framework.

2.4 Discussion

In the Legendre polynomial the groups of 'errors of non-extreme patterns' and 'errors of validation patterns' are not clearly distinctive, but we can observe the tendency that the error-mean of the patterns belonging to the validation set is less than the error-mean of the patterns belonging to the set of non-extremes. Consequently, the method of L^2-norm comparison can only state whether a (new) pattern is similar to a pattern for which an extreme temperature has been recorded. Our tests show that it is not possible to state from just the L^2-error of that pattern whether there will be an extreme temperature or not. The method of comparing the norms can only give a tendency, but not lead to a decision.

Although the evaluation method differed a similar statement can be made about the 2-dimensional polynomial approach. A linear trend is clearly visible in figure 2.7. Only twelve out of 133 extremes (from the total data set), however, are predicted as extremes which is less then 10%. A decision cannot be given based on this relationship.

These results indicate that the local temperature is determined by more than just the pressure distribution on that particular day. A suggestion for further work might be to take other factors into account like moisture or cloudiness. An initial idea that has not been worked out is to use the dynamic changes of the patterns, *i.e.* subtracting pressure fields of two successive days.

The watershed procedure showed poor results. The possibilities of using the basins, however, are very large. In the presented approach only the basins with the largest absolute anomaly were taken into account. Improvements are likely when additional information is included. For instance, it is likely that the pressure anomaly above the local temperature is an important feature. Furthermore, interactions between pressure systems may be good predictors for local extremes. In conclusion, the watershed approach is interesting because of its simplicity and by retaining physical interpretability. Due to this physical interpretability expert knowledge is required to implement a sensible relationship from the basins to the local temperature.

2.5 Concluding remarks

Three methods for data reduction have been presented in order to predict local extremes from large scale circulation patterns. Although the results show trends that relate prediction of local extremes with measurements, these trends are not sufficient to reliably predict typical circulation patterns that cause local extremes. The methods, though, were not fully explored in this report. Further development of the methods with contribution of export knowledge of the application area is needed to improve the results.

Bibliography

[1] Michelangeli, P. A., Vautard, R., and Legras, B. (1995). Weather regimes - recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, 52(8):1237–1256. ISI Document Delivery No.: QV955 Times Cited: 105 Cited Reference Count: 30.

[2] Panja, D. and Selten, F. M. (2007). Extreme associated functions: optimally linking local extremes to large-scale atmospheric circulation structures. *Atmospheric Chemistry and Physics Discussions*, 7(5):14433–14460.

[3] Salameh, T. and Dobrinski, P. (2008). Extreme climatic events and north atlantic weather regimes: uncertainty assessment using era-40 and ncep re-analyses. *Geophysical Research Abstracts*, 10:EGU2008–A–8266.

[4] Vincent, L. and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(6):583–598.

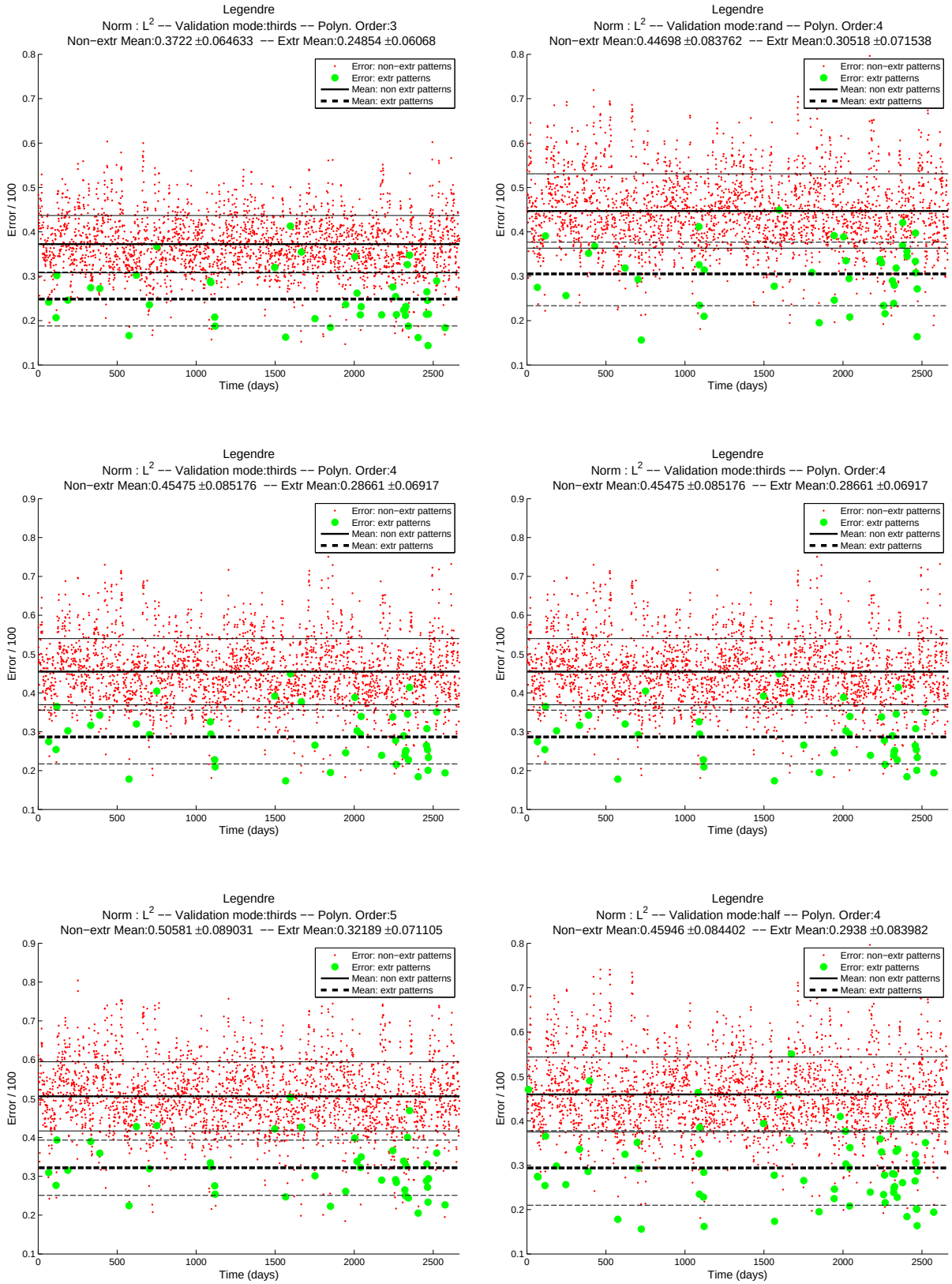

Figure 2.6: Comparison of different polynomial orders (left column – with validation mode `thirds`) and comparison of different validation modes (right column – with polynomial order 4).

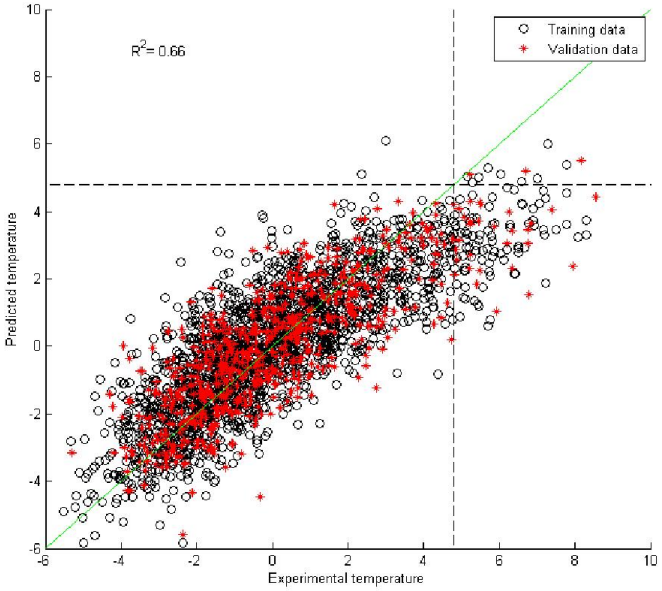

Figure 2.7: Measured vs. Predicted temperature anomalies from the 2-dimensional polynomial with 'o' calibration data and '*' validation data

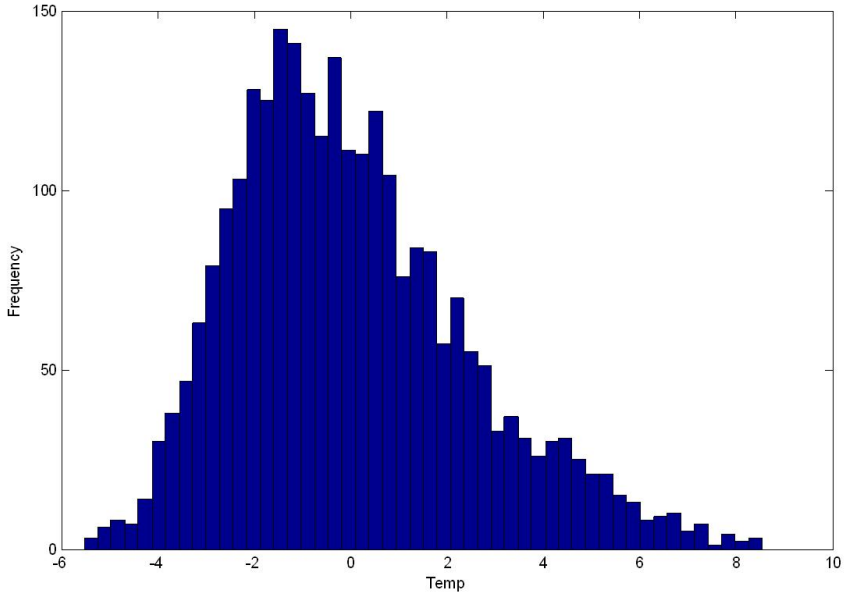

Figure 2.8: Histogram of the measured local temperature anomalies

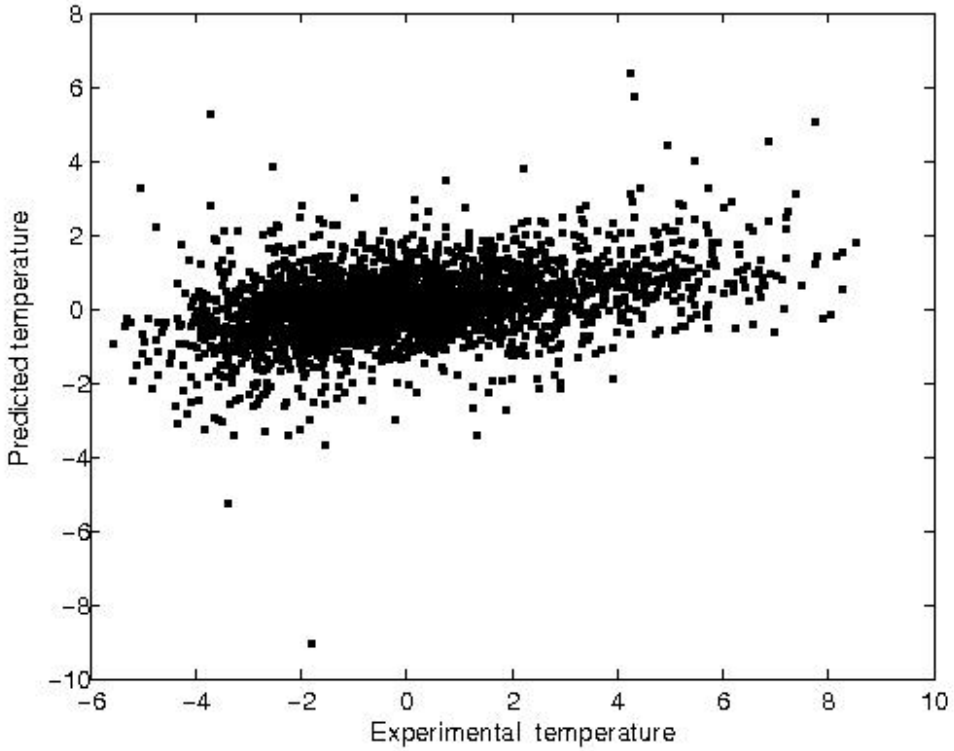

Figure 2.9: Measured vs. Predicted temperature anomalies from the watershed transform

Chapter 3
How to Mix Molecules with Mathematics

Bas van't Hof[1] Jaap Molenaar [2] Lennart Ros [1] Martijn Zaal [3]

abstract:
In this paper we develop two methods to calculate thermodynamic properties of mixtures. Starting point are the basic assumptions that also form the basis for the COSMO-RS model. In this approach, the individual molecules are represented by their geometrical shape with an electrical charge density on their surfaces. Next, the surface is split up into surface segments each with its own charge. In COSMO-RS a strong reduction is introduced by treating the segments as if they are completely independent. In the present study we take into account that the coupling between two patches is essentially dependent on the charge distribution on neighboring segments and on the local geometrical structure of the surface. Two approaches are followed. The first one points out how the model equations, which comprise the optimization of the entropy and conservation of internal energy, can efficiently be solved in general, thus also if the dependency between segments and the local geometry is included in the expression for the coupling energy between segments. In the second method the configuration with maximal entropy and prescribed energy is sought via simulation. Successive molecular configurations of the mixture are simulated and updated via a genetic algorithm to optimize the entropy. The second method is more time consuming but very general.
KEYWORDS: *Mixture properties, Entropy, Optimization, COSMO-RS*

[1]Vortech, Delft, The Netherlands
[2]Wageningen University, Wageningen, The Netherlands
[3]Free University, Amsterdam, The Netherlands

3.1 Introduction

Thermodynamic properties of a mixture, such as the miscibility of the components and partial vapor pressures, could in principle be calculated by accounting for all the interactions between the constituting molecules. In practice, however, a rigorous approach along these lines is only tractable for a highly restricted number of molecules. In view of the huge number of molecules in a fluid, one has to rely on methods from statistical physics, in which averaging procedures are applied over possible configurations. Even then one has to introduce severe assumptions in order to make calculations for realistic mixtures possible.

In 1995, a promising idea to solve this longstanding problem was worked out by Andreas Klamt [1, 2, 3]. His approach is referred to as COSMO-RS (COnductor like Screening MOdel for Realistic Solvents) and has proven to be quite powerful in some cases. One of the strong points is that the computation times are very modest. The method has its limitations, since it is based on rules that completely ignore the geometry of the molecules. The aim of the present project is to reconsider the problem of mixing anew preferably including the geometrical effects.

We decided to maintain a basic principle of COSMO-RS, namely to represent a molecule via a rigid shell with an electric charge distribution. This will be explained in §3.2. This approach assures that long-range interactions and screening effects are taken into account, but in an averaged manner, and will not lead to unacceptably long computing times.

We followed two lines of research. One line, presented in §3.3 can be looked upon as a natural extension of COSMO-RS with now the geometrical features of the molecules taken into account. In this approach, the optimization the entropy of the mixture under the condition of conserved energy is appropriately done via a fast numerical scheme.

In the second research line, presented in §3.4, the configuration with maximal entropy and prescribed energy is sought via simulation. A molecular configuration is represented in the computer by specifying the positions and orientations of a big number of molecules. An initial configuration is randomly chosen and gradually updated via a genetic optimization algorithm to optimize the entropy.

In §3.5 the results and recommendations are summarized.

3.2 The COSMO-RS model

Basic ingredients

For a clear understanding of the present project it is necessary to first explain the essential ingredients of the COSMO-RS model. Lots of details can also be found in [6, 7].

The first step in this model is taking into account long range interactions and screening effects in an averaged way. To that end the molecule is thought to be embedded in a cavity located in a perfect conductor, that is a material with an infinitely large dielectric constant. Since the molecule will in general have a charge distribution and therefore possess an electric field, it will polarize the embedding medium. That will result in an electric field that can be thought to stem from a charge distribution on the surface of the molecular cavity. In the method the molecule is replaced by the surface of the cavity together with the induced electrical charge distribution. In Figure 3.1 a sketch of such a surface and its charge distribution is given for a water molecule. Such a charge distribution is the result of a quantum mechanical calculation and is throughout this project assumed to be given for each type of molecule in the mixture.

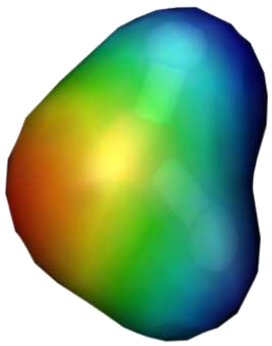

Figure 3.1: *The surface of the cavity of a water molecule with its charge distribution.*

The next step is to divide the surface up into small segments, each with a fixed amount of charge. This segmental charge is obtained by integrating the local charge distribution over the segment. So each molecule is now represented by a number of charged segments on the surface of its cavity. To keep this approach realistic, the size of these segments should be large enough to make the concept of individual pairing of segments meaningful. In practice the segment area is chosen in the range 3–25 $(\text{Angstrom})^2$.

The following step is to realize that in a fluid the molecules are nearly space filling. Each molecule is thus in touch with a number of neighboring molecules. The consequence

is that most of the time a segment of one molecule is in touch with a segment of another molecule. This contact implies a certain amount of energy, depending on the signs and the values of the segment charges. Segments with opposite charge signs attract each other and segments with equal charge signs will repel each other. The total amount of internal energy U is the sum of all the local contributions.

If the mixture would have vanishing temperature, all positions and orientations of the molecules would be fixed. The system would be *frozen in* and have maximal order. In reality we are interested in mixtures at positive temperature. In such a system the molecules move around and perform so-called Brownian motions and the overall molecular configuration is varying all the time. Macro properties of the system are then calculated by averaging either over time or over all possible microstates with appropriate weighting functions. From statistical mechanics we know that the system most frequently attains those configurations in which the *entropy* is maximal. In fact, the preference for these microstates is so high that we may ignore all the other microstates in the averaging procedure. That's why in the following we will concentrate on the calculation of maximum entropy configurations.

Entropy

Since the number of molecules is in the order of the number of Avogadro (in the order of 10^{26}), it is completely intractable to compute the time evolution of all individual molecules, the so-called *microstate*. Instead, COSMO-RS follows a different approach. To explain this, we first discuss the labeling of segments. For simplicity, let us assume that the mixture consists of two components X and Y: a molecule X has N_X segments and a molecule Y has N_Y segments. Since the molecules of type X are mutually indiscernible and the same holds for type Y, we meet in this system with $N = N_X + N_Y$ essentially different segments. In a particular microstate one could count the frequency that a segment n is coupled to a segment m, and use the frequencies to compute probabilities. However, in the present approach we prefer an alternative scaling based on surface areas involved, which will be explained underneath. We shall denote the scaled frequencies, that do not longer correspond to integers, by $p_{n,m}$. A *macrostate* of the system is now characterized by the values $p_{n,m}, n = 1 \dots N, m = n \dots N$. It is clear that one macrostate may be realized by very many different microstates, which in statistical mechanics all together are referred to as an *ensemble*. Shannon proved that the appropriate expression for the entropy S, i.e. of

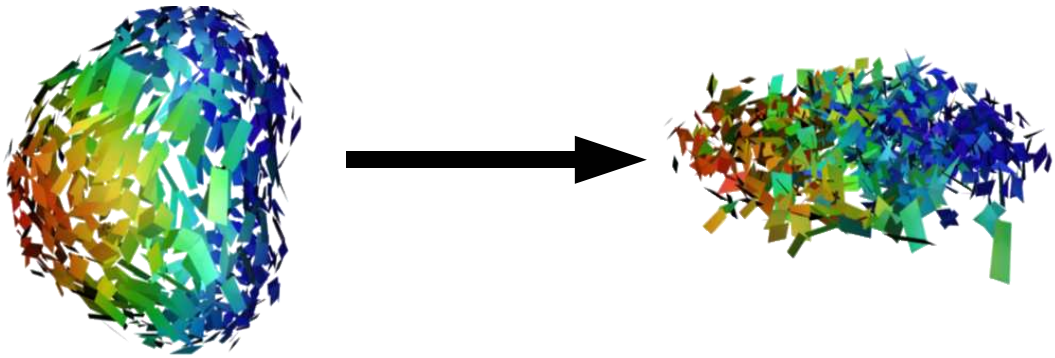

Figure 3.2: *Impression of the surface segments being treated as independent.*

the disorder of such a macrostate, reads as [5]

$$S = -k \sum_{i=1}^{N} \sum_{j=i}^{N} p_{i,j} \log p_{i,j} \tag{3.1}$$

Here, k is the Boltzmann constant ($\sim 1.3806504\,\mathrm{J\,K^{-1}}$).

Model equations and modeling assumptions

In this subsection we state the model equations and discuss the assumptions introduced by COSMO-RS.

In a microstate, two segments are considered to be coupled if they are located next to each other. A highly restrictive assumption of COSMO-RS is that the spatial embedding of a segment between its neighboring segments is completely ignored. In fact, all segments are cut free from their molecules and treated as if they are independent. In this view the mixture consists of a set of segments that move around independently, as illustrated in Figure 3.2.

As a consequence of this approximation, the energy involved in coupling segments n and m is take to be dependent on the charges of these segments only. Denoting the charge of segment n by σ_n, the coupling energy $E_{n,m}$ is assumed to be of the form

$$E_{n,m} = \alpha \left(\sigma_n + \sigma_m\right)^2 \tag{3.2}$$

for some positive coefficient α. Note that segments with equal but opposite charges have zero coupling energy, and segments with equal charges have high coupling energy. Steric hindering and the multipolar nature of the electric field of a molecule are thus not taken into

account. Obviously, coupling segment n to segment m is equivalent to coupling segment m to segment n, therefore, both $E_{n,m}$ and $p_{n,m}$ are symmetric: $E_{m,n} = E_{n,m}$ and $p_{m,n} = p_{n,m}$.

The normalization of the $p_{n,m}$ is chosen to follow from considering the relative area that is involved in such a coupling. For this normalization we take

$$\forall n : \sum_{j=1}^{N} p_{n,j} + p_{n,n} = [X_n]\gamma_n, \tag{3.3}$$

where $[X_n]$ is the molar fraction of the molecule type segment n belongs to, and γ_n is the surface area of segment n. The extra term $p_{n,n}$ stems from the fact that coupling of segment n with itself requires *two* segments n.

Given these normalizations, the internal energy of the mixture U is easily expressed in terms of the frequencies $p_{n,m}$ and the energies $E_{n,m}$:

$$\sum_{i} \sum_{j \geq i} p_{i,j} E_{i,j} = U. \tag{3.4}$$

The COSMO-RS model formally involves the optimization of the entropy as a function of the variables $p_{n,m}, n = 1 \ldots N, m = n \ldots N$ under the condition that the $p_{n,m}$ are normalized and that the internal energy equals some prescribed value U. In formulae, the required macrostate will be the solution of the following constrained optimization problem:

$$\begin{cases} \max & S(\{p_{i,j}\}) = -k \displaystyle\sum_{i=1}^{N} \sum_{j \geq i}^{N} p_{i,j} \log p_{i,j} \\ \text{under the conditions that} & \forall n : \displaystyle\sum_{j} p_{n,j} + p_{n,n} = [X_n]\gamma_n \\ \text{and the condition} & \displaystyle\sum_{i} \sum_{j \geq i} p_{i,j} E_{i,j} = U \end{cases} \tag{3.5}$$

Formally, only $p_{n,m}$ with $m \geq n$ are part of the problem. If in the following $m < n$, $p_{n,m}$ is considered to be shorthand notation for $p_{m,n}$. Although this might seem artificial at first, it makes formulas involving sums easier to read and understand.

The value of U is determined by the external conditions of the system. In practice, one often fixes the temperature T of the mixture. As discussed later on, the value of U is then an outcome, rather than an input of the system. The roles of U and T are in fact interchangeable in the procedure.

3.3 Extended COSMO-RS model

The assumption of independency of segments allows for an explicit solution of this problem along combinatorial lines using the notion of partition function. For this derivation, see

Appendix I in [4]. This reduction is a great advantage from a computational point of view. However, this assumption forms a weak point, since it makes the model quite unrealistic, e.g., to deal with irregularly shaped molecules that give rise to steric hindering. In the present approach we want to get rid of this assumption. The consequence is that we have to face the original optimization problem (3.5). It also implies that (3.2) is no longer applicable. The energy involved in coupling two segments should be made to depend on the neighboring segments, too. In the next subsection this point will be touched. For the present procedure we propose for solving (3.5) it is only relevant that some (nonnegative) expression for the coupling energy $E_{n,m}$ is available.

A general method to solve the constrained maximization problem (3.5) is to make use of Lagrange multipliers. For that purpose we form the Lagrangian

$$\begin{aligned} L(\{p_{n,m}\},\{\lambda_n\},\mu) = &- k\sum_{i=1}^{N}\sum_{j\geq i}^{N} p_{i,j}\log p_{i,j} + \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{N} p_{i,j} + p_{i,i} - [X_i]\gamma_i\right) \\ &+ \mu\left(\sum_{i=1}^{N}\sum_{j\geq i} p_{i,j}E_{i,j} - U\right) \\ = &-k\sum_{i=1}^{N}\sum_{j=i}^{N} p_{i,j}\log p_{i,j} + \sum_{i=1}^{N}\sum_{j=i}^{N}(\lambda_i+\lambda_j)p_{i,j} - \sum_{i=1}^{N}\lambda_i[X_i]\gamma_i \\ &+ \mu\left(\sum_{i=1}^{N}\sum_{j\geq i} p_{i,j}E_{i,j} - U\right) \end{aligned} \tag{3.6}$$

This Lagrangian has as variables the frequencies $p_{n,m}$, $n = 1\ldots N$, $m = n\ldots N$ and the Lagrange multipliers λ_n, $i = n\ldots N$ and μ. For the second identity, the convention $p_{m,n} = p_{n,m}$ has been used in order to eliminate any $p_{n,m}$ with $m < n$. All other quantities such as the internal energy U and the coupling energies $E_{m,n}$ act as parameters. The term containing $\lambda_i + \lambda_j$ follows by replacing $p_{i,j}$ with $p_{j,i}$ whenever $i < j$, and rearranging the double sum:

$$\sum_{i=1}^{N}\lambda_i\sum_{j=1}^{i} p_{i,j} = \sum_{j=1}^{N}\sum_{i=j}^{N}\lambda_i p_{i,j} = \sum_{i=1}^{N}\sum_{j=i}^{N}\lambda_j p_{j,i} = \sum_{i=1}^{N}\sum_{j=i}^{N}\lambda_j p_{i,j} \tag{3.7}$$

Note that the Lagrangian does not include the kinetic energy, since in a fluid the molecules motions are quite slow, so that the total energy is completely dominated by the potential (internal) energy.

Standard theory tells us that the solution of (3.5) is also the solution of the set of equations obtained by setting the derivatives of the Lagrangian with respect to each of its

variables equal to zero. So, (3.5) is equivalent to solving the system

$$\begin{cases} -k(\log p_{n,m} + 1) + (\lambda_n + \lambda_m) + \mu E_{n,m} = 0, & \forall n, \forall m \geq n \\ \sum_j p_{n,j} + p_{n,n} = [X_n]\gamma_n, & \forall n \\ \sum_i \sum_{j \geq i} p_{i,j} E_{i,j} = U. \end{cases} \tag{3.8}$$

The term $(\lambda_n + \lambda_m)$ follows from the second equality in (3.6).

A result from thermodynamics states that the Lagrange multiplier μ is related to the absolute temperature via

$$\mu = -\frac{1}{T}.$$

Since the temperature of the mixture can be controlled, μ will from now on be considered as a parameter. This implies that we only need to solve the equations in the first two lines of (3.8) for the variables $p_{n,m}$, $n = 1 \dots N$, $m = n \dots N$ and λ_n, $n = 1 \dots N$. The equation in the third line will be used afterwards to calculate the internal energy U.

Solving the first equation in (3.8) for $p_{n,m}$ and substituting in the second one, we obtain the following set of equations:

$$\begin{cases} p_{n,m} = e^{-1 + \frac{\lambda_n + \lambda_m + \mu E_{n,m}}{k}} & \forall n, \forall m \geq n \\ \sum_j e^{-1 + \frac{\lambda_n + \lambda_j + \mu E_{n,j}}{k}} + e^{-1 + \frac{2\lambda_n + \mu E_{n,n}}{k}} = [X_n]\gamma_n & \forall n \end{cases} \tag{3.9}$$

To rewrite these equations in a more tractable form we introduce the vector

$$\Lambda_n := e^{\lambda_n / k}, n = 1 \dots N$$

and the matrix

$$F_{n,m} := e^{\mu E_{n,m}/k} + \delta_{n,m} e^{\mu E_{n,n}/k}$$

with the Kronecker delta as is usual defined as $\delta_{n,m} = 1$ if $n = m$ and $\delta_{n,m} = 0$ if $n \neq m$. The last equation of (3.9) can then be written as

$$\forall n : \Lambda_n \sum_j F_{n,j} \Lambda_j = e[X_n]\gamma_n =: \alpha_n \tag{3.10}$$

The right hand sides and the matrix $F_{n,m}$ are known. So, we arrive upon a set of N quadratic equations for the unknowns $\Lambda_n, n = 1 \dots N$. This system is not simple to solve explicitly, but it has a pretty nice form for numerical evaluation. The Jacobian matrix of the set of equations (3.10) is easy to obtain explicitly. So, we resort to a numerical, and thus iterative approach and need therefore an initial guess for the Λ_n. To that end we observe that the exponentials in $F_{n,m}$ are expected to be close to one, since the coupling

energies $E_{n,m}$ are small. Setting $F_{n,m} = 1$ for all $n \neq m$ and $F_{n,n} = 2$ for all n, we obtain the approximating equation

$$\Lambda_n^2 + \Lambda_n \sum_j \Lambda_j = \alpha_n.$$

Neglecting the first term Λ_n^2 since it is expected to be small compared to the sum in the second term, we find as initial guess

$$\Lambda_n^0 := \frac{\alpha_n}{\sqrt{\sum_j \alpha_j}}.$$

Once the Λ_n are known, the values of the variables $p_{n,m}$ follow from

$$p_{n,m} = e^{-1+\frac{\lambda_n+\lambda_m+\mu E_{n,m}}{k}} = \Lambda_n \Lambda_m e^{-1+\frac{\mu E_{n,m}}{k}} \tag{3.11}$$

Example

To solve Λ_n from (3.10), we choose as iterative scheme the Newton-Raphson method. As a toy model we consider a fluid with only one molecule type with $N = 4$ segments of equal size. Furthermore, we use $\gamma_n = 1$ for all n. Taking for the $E_{n,m}$ matrix

$$E_{n,m} = \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix},$$

representing charges of equal size, but opposite sign, we found for the $p_{n,m}$ matrix

$$p_{n,m} = \begin{pmatrix} 0.0945 & 0.3583 & 0.0945 & 0.3583 \\ 0.3583 & 0.0945 & 0.3583 & 0.0945 \\ 0.0945 & 0.3583 & 0.0945 & 0.3583 \\ 0.3583 & 0.0945 & 0.3583 & 0.0945 \end{pmatrix}$$

at $T = 300$ K. This clearly shows that segments with opposite charges tend to attract each other, whereas segments with charges of equal signs repel each other. As expected, the lower the temperature, the stronger the influence of the energy. The convergence appeared to be very fast, thanks to the system being quadratic.

In Figure 3.3 it is illustrated that some couplings are geometrically impossible. In a second example we illustrate how to deal with such a situation. In the example we consider again the fluid in the example above, but now we assume that segments one and two cannot

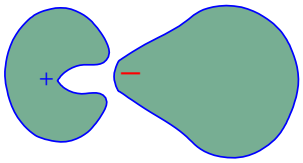

Figure 3.3: *Sketch of a situation in which a coupling is geometrically impossible, although the involved charges would favor it.*

touch each other. two. This can be taken into account by a very high entry in the energy matrix, say $E_{1,2} = 20$:

$$E_{n,m} = \begin{pmatrix} 4 & 20 & 4 & 0 \\ 20 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix},$$

The coupling frequencies now become

$$p_{n,m} = \begin{pmatrix} 0.2073 & 0.0010 & 0.1219 & 0.4625 \\ 0.0010 & 0.2073 & 0.4625 & 0.1219 \\ 0.1219 & 0.4625 & 0.0717 & 0.2721 \\ 0.4625 & 0.1219 & 0.2721 & 0.0717 \end{pmatrix}$$

As expected, the coupling frequency between segments one and two dropped to almost zero. Note that also the other entries have changed. The highest frequency is now found between one and four, as was to be expected, since this is energetically speaking the most favorable coupling.

Choice of coupling energies

Using the above model, the macrostate with the highest entropy can be easily calculated, provided that the coupling energies $E_{n,m}$ are given. It remains to specify them such that the geometrical effects are accounted for. In the present project we developed some ideas, which are worth to be worked out. out.

- Include neighboring effects. If two segments couple, also the neighbors come close together. It depends on the charges on the neighboring segments and their distances what the effect will be on the energy. A possibility to take this into account is to choose

$$E_{n,m} = \alpha \left(\sigma_n + \sigma_m\right)^2 + \beta \sum_{i_n, j_m} d_{i,j} \left(\sigma_i + \sigma_j\right)^2,$$

where i_n runs over all neighbors of segment n and j_m runs over all neighbors of segment m and $d_{i,j}$ is some appropriate distance function. The factor β has to be finetuned in order to get the correct balance between the two terms. In this way we introduce a penalty if a coupling involves neighbors that repel each other. So the second term acts as a penalty function. Including higher order neighbor effects might also be an option.

- An alternative would be to include the local curvatures into $E_{n,m}$, for instance a term proportional to
$$(H_n + H_m)^2 ,$$
where H_n is the (average) mean curvature of the molecule surface around the position of segment n. The advantage of this criterion is that it is much less subjective than defining penalties for individual couplings.

- Forbidden couplings. If illustrated in the example above, if some coupling is physically infeasible due to the shape of molecules, it can be forbidden simply by assigning to it a very high energy cost. It is to be expected that this will somewhat reduce the quality of the initial guess discussed above, which means that the numerical method will need more time to find the solution.

3.4 Entropy optimization via simulation

In this section we follow an approach that is considerably different from the one presented in the preceding section. The aim is the same: to find a configuration with maximum entropy and prescribed energy. The idea is to do perform this via simulation. We focus on a part of the fluid, a so-called parcel, with a tractable number of molecules. The rest of the fluid is represented by periodic boundary conditions, as explained below. The molecular configuration in this fluid parcel is represented in the computer by specifying the positions and orientations of all molecules in it. An initial configuration is randomly chosen and gradually updated via a genetic optimization algorithm to optimize the entropy, meanwhile keeping the energy at or close to the prescribed value. This approach has the complication that randomly placed molecules will in general overlap. So, this leads to an extra optimization goal: minimization of the overlap.

The present approach has the following features:

- As we already did above in the (extended) COSMO-RS model, we ignore the kinetic energy. So, our search space is the set of static configurations in the fluid parcel.

- The surface of the molecule is approximated by segments, each with its own charge. The geometry of the surface is taken into account, so the segments are connected.
- The state of a molecule consists of is 6 parameters per molecule: 3 coordinates for the location and 3 angles for the orientation. From these the position of each segment directly follows.
- In the coupling energy between segments we incorporate the geometry, in the way discussed in §3.3.

Periodic boundary conditions

In the simulation approach we calculate the properties in a small fluid parcel. this is based on the assumption that on average the parcel represents the fluid as a whole quite well. To avoid boundary effects, periodic boundary conditions are applied. This results in a periodic domain, as illustrated in Figure 3.4. Now, we deal with an infinitely large domain, but represented with only a finite amount of information because of the repeating patterns.

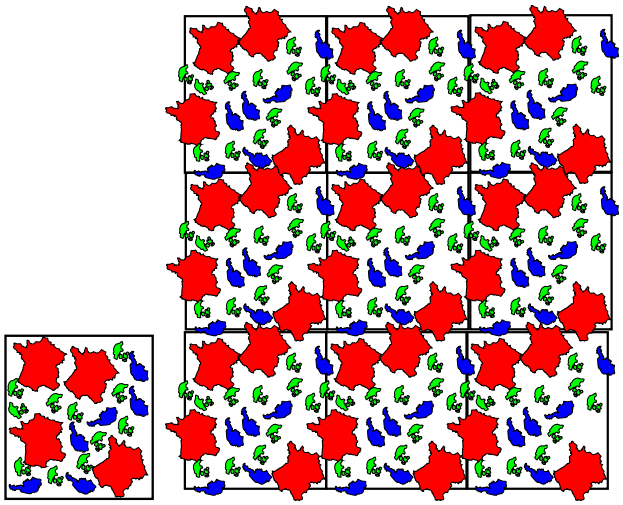

Figure 3.4: *Left: A small fluid parcel. Right: A periodic domain. A periodic domain has no boundaries.*

Optimization procedure

Let us consider n molecules (maybe of different species) in the fluid parcel. We use the following noattions:

state The state $\boldsymbol{x} \in \mathbb{R}^{6n}$ of the configuration consists of the locations and orientations of all n molecules;

Energy function The energy function $E : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the binding energy for the given state;

Entropy function The entropy function $S : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the entropy for the given state;

Overlap function The function $V : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the amount of space occupied by two or more molecules at the same time.

For a given *target energy* E_t we have to solve the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & S(\boldsymbol{x}) \\ \text{under the restrictions that} \quad & V(\boldsymbol{x}) = 0, \\ \text{and} \quad & (E(\boldsymbol{x}) - E_t)^2 = 0. \end{aligned} \tag{3.12}$$

3.4.1 Technical details

The optimization problem (3.12) has many local optima. By the way, it is good to realize that it also has many global optima. For example, if we have an optimal solution and we shift the whole solution a little bit (and/or rotate it) we again have an optimal solution. In general, it is typical for many-particles systems that one and the same macro state may correspond to a huge amount of micro states, all having the same entropy and energy. In the present approach we need to find only one of the global optima. Since the system has so many degrees of freedom, optimization may lead to unacceptably long computation times. The success of the method will therefore heavily depend on how efficiently the functions E, S and V and their gradients are evaluated. In this section we discuss several related technical details.

Efficient evaluation of overlap V

Each molecule may be described as a set of tetrahedra. The overlap in a configuration can therefore be determined by comparing every one of these tetrahedra to every other tetrahedron, calculating the volume they share and adding all these overlap volumes. Such a process is quadratic in the number of tetrahedra in the configuration and would become prohibitive very quickly when many molecules are to be modelled, or when detailed shapes are to be used to model them.

The calculation of the overlap can be sped up considerably by keeping track of the *circumscribed spheres* of the molecules, as illustrated in Figure 3.5. This is very simple to do, because the circumscribed sphere of the molecules does not change when the molecule is rotated and because its radius only depends on the molecule species. If the circumscribed spheres do not intersect, the molecules do not intersect and their tetrahedrons need not be compared. In this way, every molecule is only seriously compared to the molecules near it. A similar speed-up may be achieved by comparing the circumscribed spheres of the individual tetrahedra before calculating their overlap.

A further reduction in the calculation can be achieved using a *grid*, as illustrated in Figure 3.6. The domain is split up into grid cells. For every grid cell, a list is made of all molecules in or near it (i.e. whose center of gravity is in the shaded area). Molecules near a grid cell boundary may be in more than one list.

In this case the calculation of the overlap consists of the following steps:

```
1: for all molecules do
2:    place it in a list of all grid cells in or near which it is located
3: end for
4:
5: for all molecules M_1 do
6:    for all molecule M_2 in or near the grid cell where molecule M_1 is located  do
7:       compare circumscribed spheres:
8:       if spheres do not intersect then
9:          Overlap V is zero.
10:      else
11:         compare all tetrahedra of M_1 to all tetrahedra of M_2:
12:         if there is no intersection then
13:            Overlap V is zero.
14:         else
15:            a detailed calculation is needed
16:         end if
17:      end if
18:   end for
19: end for
```

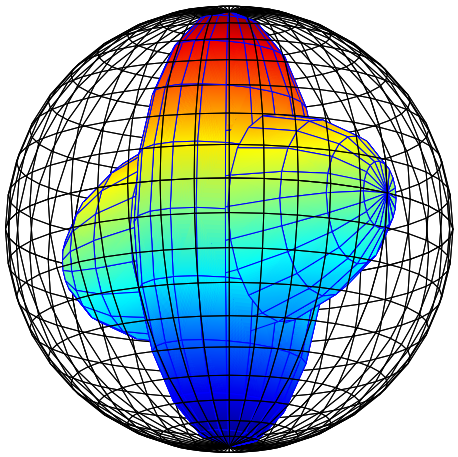

Figure 3.5: *A molecule and its circumscribed sphere: molecules do not overlap if their circumscribed spheres do not do.*

Efficient evaluation of coupling frequencies

For the evaluation of the energy and the entropy, it is necessary to determine for every segment of the molecule shell to which segment(s) it is 'coupled'. A simple way to determine these couplings is by the overlap calculation of slightly enlarged molecules. This idea is illustrated in Figure 3.7. The molecules $M1$ and $M2$ (dark colors) do not overlap. The enlarged molecules (lighter colors), however, have some overlap. Segment A1, or rather the tetrahedron that it is a face of, overlaps with B2 and a little bit with A2. Hence, we say that A1 is coupled mostly to B2 and a bit to A2 and we let both couplings contribute to the entropy, but in a weighted fashion.

Smoothing the functions

The overlap-function V and the coupling frequencies (and hence the energy E and entropy S) are continuous and differentiable functions of the state $\boldsymbol{x}$. Their derivatives, however, are not continuous, so the Hessian matrices of the functions V, E and S do not exist. Since many optimization techniques need Hessian matrices, it is useful to smooth these functions. A simple way to do this is to 'soften' the tetrahedra. When doing so, the original overlap

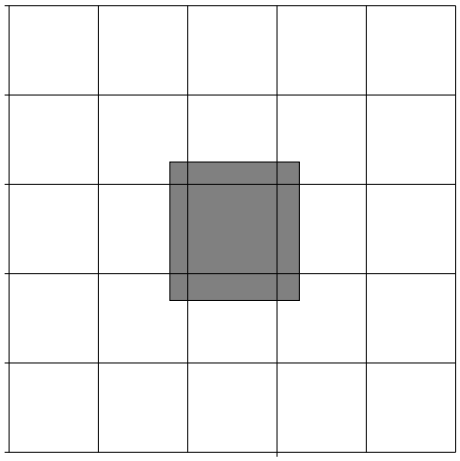

Figure 3.6: *The grid used to speed up the calculation of the overlap.*

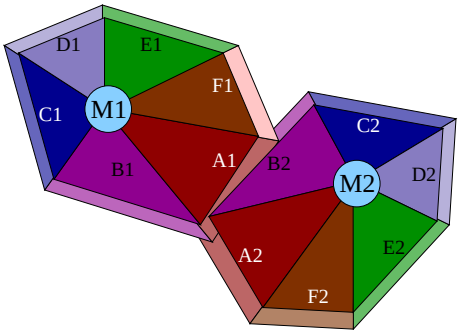

Figure 3.7: *Example for the calculation of the couplings: segment A1 is mostly coupled to segment B2, and also a little bit to A2.*

V_{ij} between two tetrahedra i and j is modified to V'_{ij} according to

$$V'_{ij} := \frac{V_{ij}^2}{\epsilon \min(V_i, V_j) + V_{ij}}, \tag{3.13}$$

where V_i and V_j are the volumes of tetrahedra i and j, and ϵ is a 'small' parameter. Larger values for ϵ make 'softer' overlap functions.

3.4.2 Efficient optimization of the configuration

The original optimization problem (3.12) involves a target function and constraints. The constraints can be incorporated in the target function by giving a penalty for constraint violation. The modified optimization method is then

$$\text{maximize} \qquad S(\boldsymbol{x}) - c_V V(\boldsymbol{x}) - c_E (E(\boldsymbol{x}) - E_t)^2. \tag{3.14}$$

with c_V and c_E weighting factors that determine the relative contributions of the two penalty functions. This optimization problem is standard problem and may be solved using steepest descent or variations of Newton's method. In the present context some problems might be expected:

- Local optimization methods are very likely to find local optima which are not global optima.
- Local search techniques may also converge very slowly. This may happen for instance in configurations with regions that are too crowded and regions which are too empty. A lot of molecules have to move in order to even this out. They will moreover have to move in complicated patterns because the target function is not allowed to increase on the way.

To find a global optimum, additional techniques may be needed. When a local optimum is found or when convergence slows down, the solution has to be 'shaken up' in order to move away from a local optimum. Sudden changes which may help are for example

- Some (randomly chosen) molecules may be taken from the most crowded regions and placed in the emptiest regions;
- Some (randomly chosen) molecules are moved and rotated to a random place and orientation in the domain.

3.4.3 Preliminary results

The simulation approach requires a lot of programming. Due to time limitations it was not possible to produce a working molecular simulation model in only a few days. A modest start in 2D was made, which provides us with some understanding of what is involved in the calculations. The evaluation of overlap turned out to be not too complicated. The couplings were evaluated only in a simple way: every segment was considered to couple to the nearest segment of another molecule. Local search was not yet applied. For purpose of demonstration, optimization was studied via a simple random search algorithm. In that approach, a configuration $\boldsymbol{x}$ is chosen entirely randomly, after which the target function (3.14)is evaluated. The first configuration is saved and a new configuration is randomly produced. If this configuration turns out to have a higher value of the target function, then the latter replaces the former. This can be repeated many times. Obviously, this method has very slow convergence. The results of this procedure are shown in Table 3.1 and Figure 3.8. Two types of molecules are mixed: 18 of one type and 7 of another type.

Iteration	Overlap	Energy	Entropy
1	40.7	0.30	4.53
2	38.4	0.36	4.50
4	37.5	0.35	4.44
5	30.2	0.34	4.54
10	27.6	0.39	4.46
31	20.2	0.32	4.52
593	18.3	0.39	4.43
939	17.2	0.41	4.41

Table 3.1: *Results when maximizing the target function (3.14)during a random search approach. The overlap indeed reduces in the course of the time*

The dimensions of the molecules, the domain and charges were not realistically chosen, that's why no units are shown in the results. The coefficients c_V and c_E were set at one and for the target energy we use $E_t = 40$. A thousand configurations were produced, and 8 times a new 'best sofar' configuration was encountered. Table 3.1 shows that in this instance the overlap is indeed minimized, but the entropy and energy are still varying much. The initial and final (after 8 improvement steps) configurations are shown in Figure 3.8

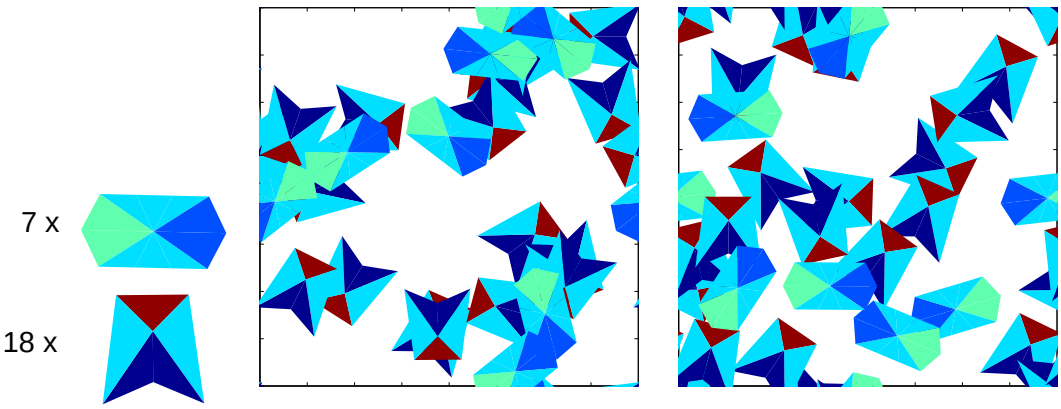

Figure 3.8: *First (left) and final (right) configurations in the a simple random search summarized in Table 3.1.*

3.5 Conclusions and Recommendations

We have shown that the COSMO-RS procedure to calculate the properties of mixtures can be extended to incorporate the geometrical effect of constraints that may drastically influence the chance that two surface segments of the constituting molecules couple. The general problem concerns the optimization of the entropy under the condition that the energy has a prescribed value. To perform this task while accounting for the geometrical effects, we followed two lines.

In the first approach, we show that the optimization problem can be very efficiently solved, by putting it in a form that is appropriate for numerical optimization methods. The geometrical constraints are included via specification of the energy involved in coupling two segments. We discuss suggestions for the effective choice of these coupling energies, such that the effect of the local geometry and the local charge distribution is taken into account.

In the second approach, we tackle the optimization problem via simulation. We focus on a part of the fluid, a so-called parcel, with a tractable number of molecules. The rest of the fluid is represented by periodic boundary conditions. The molecular configuration in this fluid parcel is represented in the computer by specifying the positions and orientations of all molecules in it. The idea is to start from a randomly chosen configuration, that is gradually updated via a genetic optimization algorithm. The object function consists of the entropy together with penalty functions that have to assure that the procedure converges to a configuration with the correct energy and without overlapping molecules. A fairly complete image of the computational aspects was obtained from developing a simple piece of software, that is restricted to 2D.

Our conclusion is that the first approach answers the original specific question quite efficiently, while the second approach is highly general and could also be applied to answer many other questions concerning mixtures.

3.6 Acknowledgements

The authors wish to thank Jaap Louwen and Peter Daudeij from Albemarle Catalysts Company BV, Amsterdam, for bringing the problem to their attention and inspiring and pleasant discussions on this topic.

Bibliography

[1] A. Klamt, *Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena*, J. Phys. Chem. 99 (1995) 2224..

[2] A. Klamt, V. Jonas, T. Brger and J.C. Lohrenz, *Refinement and Parametrization of COSMO-RS*, J. Phys. Chem. A 102 (1998) 5074.

[3] A. Klamt, *COSMO-RS From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, Elsevier, Amsterdam (2005), ISBN 0-444-51994-7.

[4] S.T. Lin and S.I. Sandler, *A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model*, Ind. Eng. Chem. Res. 41 (2002), pp. 899 - 913

[5] E.T. Jaynes, *Information Theory and Statistical Mechanics*, The Physical Review, Vol. 106, No. 4, (1957), pp 620 - 630.

[6] C.C. Pye, T. Ziegler, E. van Lenthe, J.N. Louwen, *An implementation of the conductor-like screening model of solvation within the Amsterdam density functional package. Part II. COSMO for real solvents* accepted for publication in Can. J. Chem. (2009).

[7] See www.scm.com

Chapter 4
Approximate solution to a hybrid model with stochastic volatility: a singular-perturbation strategy

Lech Grzelak[1] Tasnim Fatima[2] Harrie Hendriks[3] Simone Munao[4] Adrian Muntean[2]
Martin van der Schans[5]

abstract:
We study a hybrid model of Schöbel-Zhu-Hull-White-type from a singular-perturbation-analysis perspective. The merit of the paper is twofold: On one hand, we find boundary conditions for the deterministic non-linear degenerate parabolic partial differential equation for the evolution of the stock price. On the other hand, we combine two-scales regular- and singular-perturbation techniques to find an approximate solution to the pricing PDE. The aim is to produce an expression that can be evaluated numerically very fast.
KEYWORDS: *Stochastic volatility, European options, singular-perturbation analysis*

[1]Delft University, of Technology, Delft, The Netherlands
[2]Technical University of Eindhoven, Eindhoven, The Netherlands
[3]Radboud University of Nijmegen, Nijmegen, The Netherlands
[4]Free University of Amsterdam, Amsterdam, The Netherlands
[5]Leiden University, Leiden, The Netherlands

4.1 Introduction

Although the famous Black-Scholes model has been widely applied to price plain vanilla options, comparisons with data analysis of real markets show that some of the assumptions beyond the Black-Scholes equations are unrealistic. It seems that one of the major reasons why this inconsistency happens is the use of the constant volatility modeling assumption. Recently, a lot of attention is paid to more general volatility models - in particular for cases where the volatility is governed by a stochastic differential equation; compare [8] for a brief discussion of these aspects. Very popular in this class of models is the Schöbel-Zhu scenario, where the volatility is driven by a mean-reverting Ornstein-Uhlenbeck process [9, 10]. We refer the reader to [17] for an accessible introduction to the topic of options pricing and to [4, 14], e.g., for a presentation of concepts related to the involved stochastic differential equations.

The problem posed by Rabobank to the 64th *European Study Group Mathematics With Industry* was the following:

(A) Assuming non-zero-correlation between the processes, develop a hybrid model that can handle the stochastic behavior of both the volatility for the equity product and the interest rates.

(B) Use singular-perturbation methods, construct an approximate solution to the non-linear degenerate partial-differential equation arising in the context of pricing European-style options when the governing asset process is defined by a Schöbel-Zhu-Hull-White hybrid model, which satisfies the requirements mentioned in (A).

This paper is organized in the following fashion: In Section 4.2 we concisely describe the so-called Schöbel-Zhu-Hull-White hybrid model and indicate the form of the partial differential equation (PDE) for pricing an European option. We also mention at this point some of the main theoretical difficulties that this PDE involves. The derivation of the PDE is reported in Section 4.3. Partly based on our "physical" intuition and partly based on the Black-Scholes methodology, we propose boundary conditions for the pricing PDE. The bulk of the paper, that is Section 4.4, contains our singular-perturbation solution strategy. Section 4.5 contains our main result, i.e the approximate expression for the price given by (4.21). We discuss here a few aspects that we consider as relevant when using perturbation approaches to pricing plain-vanilla claims under multi-asset models.

4.2 Problem description

In this note we study the following Schöbel-Zhu-Hull-White hybrid model, viz.

$$\begin{cases} \mathrm{d}S_t = & r_t S_t \mathrm{d}t + & \sigma_t S_t \mathrm{d}W_t^S \\ \mathrm{d}\sigma_t = & \kappa(\overline{\sigma} - \sigma_t)\mathrm{d}t + & \eta \mathrm{d}W_t^\sigma \\ \mathrm{d}r_t = & \lambda(\bar{r} - r_t)\mathrm{d}t + & \gamma \mathrm{d}W_t^r \end{cases} \tag{4.1}$$

Here W_t^S, W_t^σ and W_t^r denote standard Brownian motions with quadratic covariation processes $\mathrm{d}W_t^S \mathrm{d}W_t^\sigma = \rho_{S\sigma}\mathrm{d}t$ and likewise for ρ_{Sr} and $\rho_{\sigma r}$. Furthermore, $\rho_{SS} = \rho_{\sigma\sigma} = \rho_{rr} = 1$. We mention that W_t^S, W_t^σ and W_t^r are standard Brownian motions under the risk neutral measure $\mathbb{Q}$. Note that the model given by the first two equations and with constant interest rate, is investigated in [16]. In what follows, we refer to (4.1) as SZHW.

A European call option is a contract that gives the buyer of the contract the right to buy a number of shares from the writer of the contract at a specified time T in the future, the expiry date, for a fixed price K, the strike price of the option. Because, the writer possibly has to sell shares to the option holder for a price less than their value on the stock market the buyer pays a premium to the writer, this is the price of the option at $t = 0$. At expiry the value of the option is $\max(S(T) - K, 0)$ where S is the price of the underlying stock at expiry. The central question in pricing of derivatives is: What is the price of the option at time $t = 0$, which is calculated by determining its price at all times between $t = 0$ and expiry?

In Section 4.3, we derive the pricing PDE for an European option

$$\begin{aligned} 0 \;=\; & \frac{\partial V}{\partial t} + \frac{\partial V}{\partial S} rS + \frac{\partial V}{\partial \sigma}\kappa(\overline{\sigma} - \sigma) + \frac{\partial V}{\partial r}\lambda(\bar{r} - r) + \\ & \frac{1}{2}\frac{\partial^2 V}{\partial S^2}\sigma^2 S^2 + \frac{1}{2}\frac{\partial^2 V}{\partial \sigma^2}\eta^2 + \frac{1}{2}\frac{\partial^2 V}{\partial r^2}\gamma^2 + \\ & \frac{\partial^2 V}{\partial S \partial \sigma}\sigma S \eta \rho_{S\sigma} + \frac{\partial^2 V}{\partial S \partial r}\sigma S \gamma \rho_{Sr} + \frac{\partial^2 V}{\partial \sigma \partial r}\eta\gamma\rho_{\sigma r} - rV. \end{aligned} \tag{4.2}$$

The SZHW model allows σ and r to become negative. When σ is negative, it should be noted that the correlation between changes in time of S and changes in σ reverses in sign. We remark that this causes degeneracies at several places in the pricing PDE. To be more precise, for $\sigma = 0$ the determinant of the diffusion matrix vanishes. We do not treat these difficulties here (see also Remark 4.1), but we suggest three possible solutions:

The first one is the introduction of a positive function $f(\sigma)$. The stochastic differential equation for S is then replaced by $\mathrm{d}S_t = r_t S_t \mathrm{d}t + f(\sigma_t) S_t \mathrm{d}W_t^S$. This approach has been adopted in [5], e.g.

The second solution is the Heston-Cox-Ingersoll-Ross model, see for example [8]. In the SZHW S_t cannot become negative because of the $S\mathrm{d}W$ in the equation. In the Heston-Cox-Ingersoll-Ross model the potential negativity of σ is removed in a similar way.

A third solution is to take κ large. If κ is large then if σ_t becomes negative it is pushed back very fast towards the value $\bar{\sigma}$. Thus, we might still produce realistic results if we only allow for positive σ in the pricing PDE. We adopt here the third approach.

4.3 Derivation of a deterministic PDE

Consider the SZHW, see (4.1). We define

$$V(t, S_t, \sigma_t, r_t) = B(t)\mathbb{E}^{\mathbb{Q}}\left(\frac{\max(S_T - K, 0)}{B(T)} \,\middle|\, \mathcal{F}_t\right) = \mathbb{E}^{\mathbb{Q}}\left(\frac{\max(S_T - K, 0)}{B(T)/B(t)} \,\middle|\, \mathcal{F}_t\right)$$

Here $B(t) = \exp\left(\int_0^t r_s \mathrm{d}s\right)$ and $\mathcal{F}_t = \sigma(S_s, \sigma_s, r_s; s \leq t$. In particular $B(t)$ satisfies the "ordinary" differential equation

$$\mathrm{d}B(t) = r_t B(t)\mathrm{d}t.$$

We are very well aware of the fact that the coefficients in (4.1), in particular the coefficient $\sigma_t S_t$, do not satisfy the usual Lipschitz condition for an Itô diffusion. This might cause difficulties, for example in ensuring the existence of solutions of SZHW model in the precise time interval of interest for the financial situation, cf. [13], for a solution see [8, 9]. In this paper, we waive these complications and *assume* that there exists a differentiable function $\Pi = \Pi(t, S, \sigma, r)$ such that

$$\mathbb{E}^{\mathbb{Q}}\left(\frac{\max(S_T - K, 0)}{B(T)} \,\middle|\, \mathcal{F}_t\right) = \frac{V(t, S_t, \sigma_t, r_t)}{B(t)} = \Pi(t, S_t, \sigma_t, r_t)$$

We postpone the investigation of the existence of Π for a later stage. It is clear from the definition that $\Pi_t = \Pi(t, S_t, \sigma_t, r_t)$ is a martingale. Since $B(t)$ is such a simple process, Itô formula leads to

$$\mathrm{d}\Pi_t = \mathrm{d}\left(\frac{V_t}{B(t)}\right) = \frac{1}{B(t)}\mathrm{d}V_t - r_t\frac{V_t}{B(t)}\mathrm{d}t. \tag{4.3}$$

Now, we derive the Itô differential equation for V. Using Itô formula Theorem 4.2.1 from [14], we obtain

$$\begin{aligned}
\mathrm{d}V_t &= \frac{\partial V}{\partial t}\mathrm{d}t + \frac{\partial V}{\partial S}\mathrm{d}S_t + \frac{\partial V}{\partial \sigma}\mathrm{d}\sigma_t + \frac{\partial V}{\partial r}\mathrm{d}r_t + \\
&\quad \frac{1}{2}\frac{\partial^2 V}{\partial S^2}\mathrm{d}S_t\mathrm{d}S_t + \frac{1}{2}\frac{\partial^2 V}{\partial \sigma^2}\mathrm{d}\sigma_t\mathrm{d}\sigma_t + \frac{1}{2}\frac{\partial^2 V}{\partial r^2}\mathrm{d}r_t\mathrm{d}r_t + \\
&\quad \frac{\partial^2 V}{\partial S\partial \sigma}\mathrm{d}S_t\mathrm{d}\sigma_t + \frac{\partial^2 V}{\partial S\partial r}\mathrm{d}S_t\mathrm{d}r_t + \frac{\partial^2 V}{\partial \sigma\partial r}\mathrm{d}\sigma_t\mathrm{d}r_t \\
&= \frac{\partial V}{\partial t}\mathrm{d}t + \frac{\partial V}{\partial S}\mathrm{d}S_t + \frac{\partial V}{\partial \sigma}\mathrm{d}\sigma_t + \frac{\partial V}{\partial r}\mathrm{d}r_t + \\
&\quad \frac{1}{2}\frac{\partial^2 V}{\partial S^2}\sigma_t^2 S_t^2\mathrm{d}t + \frac{1}{2}\frac{\partial^2 V}{\partial \sigma^2}\eta^2\mathrm{d}t + \frac{1}{2}\frac{\partial^2 V}{\partial r^2}\gamma^2\mathrm{d}t + \\
&\quad \frac{\partial^2 V}{\partial S\partial \sigma}\sigma_t S_t\eta\rho_{S\sigma}\mathrm{d}t + \frac{\partial^2 V}{\partial S\partial r}\sigma_t S_t\gamma\rho_{Sr}\mathrm{d}t + \frac{\partial^2 V}{\partial \sigma\partial r}\eta\gamma\rho_{\sigma r}\mathrm{d}t
\end{aligned}$$

Eventually, by the martingale representation theorem Theorem 4.3.4 of [14], the $\mathrm{d}t$ term in the full expansion of Eqn. (4.3) in $\mathrm{d}t$, $\mathrm{d}W_t^S$, $\mathrm{d}W_t^\sigma$ and $\mathrm{d}W_t^r$ has to vanish. After multiplication with $B(t)$ it leads to pricing PDE Eqn. (4.2)

$$\begin{aligned}
0 &= \frac{\partial V}{\partial t} + \frac{\partial V}{\partial S}rS + \frac{\partial V}{\partial \sigma}\kappa(\overline{\sigma} - \sigma) + \frac{\partial V}{\partial r}\lambda(\bar{r} - r) + \\
&\quad \frac{1}{2}\frac{\partial^2 V}{\partial S^2}\sigma^2 S^2 + \frac{1}{2}\frac{\partial^2 V}{\partial \sigma^2}\eta^2 + \frac{1}{2}\frac{\partial^2 V}{\partial r^2}\gamma^2 + \\
&\quad \frac{\partial^2 V}{\partial S\partial \sigma}\sigma S\eta\rho_{S\sigma} + \frac{\partial^2 V}{\partial S\partial r}\sigma S\gamma\rho_{Sr} + \frac{\partial^2 V}{\partial \sigma\partial r}\eta\gamma\rho_{\sigma r} - rV.
\end{aligned}$$

We look for a solution V which is bounded by a polynomial in (S, σ, r). The final condition, given at $t = T$, is

$$V(T, S, \sigma, r) = B(T)\frac{\max(S - K, 0)}{B(T)} = \max(S - K, 0), \tag{4.4}$$

where K is the strike price of the call option. It is worth noting that above procedure provides a deterministic PDE for the price evolution but does not specify the boundary conditions needed to close the formulation of the problem. The solution being bounded by a polynomial in its variables may be enough as boundary condition. Based upon the solution and boundary conditions typically used for the Black-Scholes equation as well as by the "physics" of the problem, we suggest the following boundary conditions:

$$\begin{aligned}
V &\to 0 \text{ as } r \to -\infty, \\
V &\sim S \text{ as } S \to \infty,\ \sigma \to \infty \text{ or } r \to \infty, \\
V &\to 0 \text{ as } S \to 0 \\
V &\sim S - Ke^{-r(T-t)} \text{ as } \sigma \to -\infty.
\end{aligned} \tag{4.5}$$

This is one of the important results of this paper. Note that depending on the financial scenario in question, other boundary conditions might be employed. The fundamental question which needs to be addressed is: To which extent such choices of boundary conditions lead to well-posed PDEs? We refer the reader to [17] Section 3.7 for a nice and inspiring discussion of the boundary conditions to the Black-Scholes equation.

4.4 Our solution strategy

Our basic idea is to combine regular and singular perturbation techniques to analyze the parabolic PDE for V (arising when pricing the options in the presence of stochastic volatility) for a non-degenerate scenario in the presence of couple of characteristic time scales. The forthcoming sections have the following structure. In Section 4.4.1 we discuss a slightly different model and a reference in which perturbation methods are applied to this model. We believe these results can be extended to the SZHW model. Unfortunately, a full extension of these results is not feasible within the scope of the study group. In the remaining sections we make a step towards extending these results to the SZHW model.

4.4.1 Perturbation methods applied to a slightly different model

In [5] the authors discuss the following model

$$\begin{cases} \mathrm{d}X_t = & \mu X_t \mathrm{d}t + & f\left(Y_t, Z_t\right) X_t \mathrm{d}W_t^X \\ \mathrm{d}Y_t = & \frac{1}{\epsilon}(m - Y_t)\mathrm{d}t + & \frac{\nu\sqrt{2}}{\sqrt{\epsilon}}\mathrm{d}W_t^Y \\ \mathrm{d}Z_t = & \delta c\left(Z_t\right)\mathrm{d}t + & \sqrt{\delta} g\left(Z_t\right)\mathrm{d}W_t^r, \end{cases} \tag{4.6}$$

where both $\epsilon, \delta \ll 1$ and the three stochastic processes are correlated. In this model the stochastic processes for Y and Z should be interpreted as a *fast* and a *slow* volatility. This model differs from the SZHW model in the first equation. In this model the first equation depends on Z (the third equation) through the function f in front of the stochastic term $\mathrm{d}W_t^X$. In the SZHW model the dependence on the third equation appears in front of the deterministic term $\mathrm{d}t$. Apart from only suggesting an asymptotic expansion, the authors of [5] also discuss the error analysis making use of higher order terms in their expansion. Additionally, they also perform a calibration of their solution to existing data. Here we concentrate on finding the asymptotic expansion. To this end, we apply a perturbation method involving two scales to approximate SZHW model in some limiting situations. In Section 4.4.2 we describe the basic setup, in Section 4.4.3 we discuss the limit $\epsilon \to 0$, while in Section 4.4.4 we discuss the second limit $\delta \to 0$. In Section 4.5 we list our expansion.

Note that Section 2.6.2 of the PhD thesis [18] contains a summary of the multiscale expansion developed in [5]. Both [11] and [12] report on a detailed perturbation analysis for the fast mean reverting model (consisting of only the first two equations).

4.4.2 Set-up

Consider the SZHW model (4.1). Analogously to the approach in [5], we look to the scales

$$\kappa = \frac{\bar{\kappa}}{\epsilon},\ \eta = \frac{\bar{\eta}}{\sqrt{\epsilon}},\ \lambda = \delta\bar{\lambda},\ \gamma = \sqrt{\delta}\bar{\gamma}. \tag{4.7}$$

In terms of these scales, the SZHW model becomes

$$\begin{cases} \mathrm{d}S_t = & r_t S_t \mathrm{d}t + & \sigma_t S_t \mathrm{d}W_t^S \\ \mathrm{d}\sigma_t = & \frac{\bar{\kappa}}{\epsilon}(\overline{\sigma} - \sigma_t)\mathrm{d}t + & \frac{\bar{\eta}}{\sqrt{\epsilon}}\mathrm{d}W_t^\sigma \\ \mathrm{d}r_t = & \delta\bar{\lambda}(\bar{r} - r_t)\mathrm{d}t + & \sqrt{\delta}\bar{\gamma}\mathrm{d}W_t^r. \end{cases} \tag{4.8}$$

We note that the second equation can be obtained from the second equation in (4.1) by scaling time with a factor $\frac{1}{\epsilon}$ and that the third can be obtained from the third equation in (4.1) by scaling time with a factor δ. Intuitively the choice of these scales implies that the volatility σ is pushed very fast towards the average value $\bar{\sigma}$. Furthermore, we expect that the interest rate r evolves very slowly in time, and thus is approximately constant on short time scales.

If we set $S = e^x$ and choose only one of the correlations $\rho_{\sigma r}$ to vanish, then according to the derivation in Section 4.3 the corresponding PDE becomes

$$\begin{aligned} V_t + \frac{\sigma^2}{2}V_{xx} + \frac{\bar{\eta}^2}{2\epsilon}V_{\sigma\sigma} + \frac{\bar{\gamma}^2\delta}{2}V_{rr} + \sigma\frac{\bar{\eta}}{\sqrt{\epsilon}}\rho_{S\sigma}V_{x\sigma} + \sigma\bar{\gamma}\sqrt{\delta}\rho_{Sr}V_{xr} \\ + \frac{\bar{\kappa}}{\epsilon}(\bar{\sigma} - \sigma)V_\sigma + \bar{\lambda}\delta(\bar{r} - r)V_r + \left(r - \frac{\sigma^2}{2}\right)V_x - rV = 0. \end{aligned} \tag{4.9}$$

The correlation $\rho_{\sigma r}$ is the instantaneous correlation between the short rate process r_t and the volatility process σ_t. In practice this additional parameter could be used as an additional degree of freedom in the calibration. However, for simplicity we set this correlation equal to zero while assuming non-zero correlation between: the stock process S_t and the interest rate process r_t, ρ_{Sr}, and the stock process S_t and the volatility process σ_t, $\rho_{S\sigma}$.

Remark 4.1. Note that if σ vanishes, then some of the "diffusivities" vanish as well, and hence, (4.9) becomes a *degenerate* parabolic equation. Trusting the analysis work by Achdou et al. (see, for instance, [1, 2]) we expect that a variational analysis involving

weighted Sobolev spaces and the theory of semigroups may enable us to prove the existence and uniqueness of weak solutions as well as a maximum principle. From a practical point of view, the role of such an analysis is to yield a unique positive and polynomially bounded price V. It is worth noting that the PDE (4.9) might be also viewed as a diffusion equation for infinite fissured media (somehow in the spirit of [3]). As far as we know, this perspective is rich in new ideas and we think that it deserves further analytical investigation.

To solve this PDE we are going to use both singular and regular perturbation methods for two different small parameters, namely ϵ and δ. We take for granted that the price V can be approximated by an asymptotic expansion in terms of ϵ and δ as

$$V = V_0 + \sqrt{\epsilon}V_1 + \sqrt{\delta}V_2 + O(\delta, \epsilon).$$

In the next two sections we look at the limits $\epsilon \to 0$ and $\delta \to 0$ separately.

4.4.3 The limit $\epsilon \to 0$

We wish now to treat the case ϵ small and compute the terms V_0 and V_2 of the formal expansion of V. In this case the volatility is fluctuating very fast with a fixed variance, and we deduce from [5] Definition 3.3 and [12] equation (22) that the effect of this for the PDE is that we can take constant volatility $\bar{\sigma}$. Thus, using these references we obtain that in the limit $\epsilon \to 0$ the PDE simplifies and takes the form

$$V_t + \frac{\bar{\sigma}^2}{2}V_{xx} + \frac{\bar{\gamma}^2\delta}{2}V_{rr} + \bar{\sigma}\bar{\gamma}\sqrt{\delta}\rho_{Sr}V_{xr} + \bar{\lambda}\delta\left(\bar{r} - r\right)V_r + \left(r - \frac{\bar{\sigma}^2}{2}\right)V_x - rV = 0. \quad (4.10)$$

Note that V_0 does not depend on σ but only on $\bar{\sigma}$. In this way it is intuitively clear that that $O\left(\epsilon^{-1}\right)$ terms in (4.9) vanish, see [12] equation (22) for a detailed discussion of this argument.

Since in the PDE the coefficients in front of the second order derivatives are constant, we can apply the transformation

$$v(x, r, t) = e^{Ax+Br+Ct}V(x, r, t, \epsilon = 0)$$

where A, B, C are functions of (x, r). By means of an appropriate choice of A, B, and C

we obtain an equation without first-order terms. Choosing

$$\begin{aligned}
A &= -\frac{1}{2}\frac{2\delta^{3/2}\rho_{Sr}\bar{\lambda}r\bar{\sigma} - 2\delta^{3/2}\rho_{Sr}\bar{\lambda}\bar{r}\bar{\sigma} + 2\gamma r - \gamma\bar{\sigma}^2}{\bar{\sigma}^2\gamma(\delta\rho_{Sr}^2 - 1)},\\
B &= \frac{1}{2}\frac{2\lambda\delta r\bar{\sigma} + 2\gamma\sqrt{\delta}\rho_{Sr}r - \gamma\sqrt{\delta}\rho_{Sr}\bar{\sigma}^2 - 2\lambda\delta\bar{r}\bar{\sigma}}{\gamma^2(\delta\rho_{Sr}^2 - 1)\bar{\sigma}},\\
C &= -\frac{1}{4}\frac{1}{\bar{\sigma}^2\gamma^2(\delta\rho_{Sr}^2-1)^2}(4\gamma^2 r\bar{\sigma}^2\delta^2\rho_{Sr}^4 - 8\delta\rho_{Sr}^2\gamma^2 r\bar{\sigma}^2 - 12\delta^{3/2}\rho_{Sr}\gamma r\lambda\bar{r}\bar{\sigma} + 4\lambda\delta^{5/2}\bar{r}\bar{\sigma}\gamma\rho_{Sr}^3 r\\
&- 6\gamma\bar{\sigma}^3\delta^{3/2}\rho_{Sr}\lambda r + 6\gamma\bar{\sigma}^3\delta^(3/2)\rho_{Sr}\lambda\bar{r} + 4\gamma^2 r^2 + \gamma^2\bar{\sigma}^4 + 2\gamma\bar{\sigma}^3\delta^{5/2}\rho_{Sr}^3\lambda r - 4\gamma r^2\delta^{5/2}\rho_{Sr}^3\lambda\bar{\sigma}\\
&- 2\lambda\delta^{5/2}\bar{r}\bar{\sigma}^3\gamma\rho_{Sr}^3 + 12\delta^{3/2}\rho_{Sr}\gamma r^2\lambda\bar{\sigma} + 4\lambda^2\delta^2 r^2\bar{\sigma}^2 + 4\lambda^2\delta^2\bar{r}^2\bar{\sigma}^2 - 8\lambda^2\delta^2\bar{r}\bar{\sigma}^2 r)
\end{aligned}$$

we obtain

$$v_t + \frac{1}{2}\sigma^2 v_{xx} + \frac{1}{2}\gamma^2 v_{rr} + \sigma\gamma\rho_{Sr}v_{xr} = 0. \tag{4.11}$$

We eliminate the cross terms with a rotation of the axes given by the transformation

$$\begin{cases}
X &= \dfrac{\frac{1}{2}\sigma\rho_{Sr}\gamma}{\sqrt{\left(\frac{1}{2}\sigma\rho_{Sr}\gamma\right)^2 + \left(\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sqrt{(\gamma^2+\sigma^2)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)\right)^2}}x\\
&\quad - \dfrac{\frac{1}{2}\sigma\rho_{Sr}\gamma}{\sqrt{\left(\frac{1}{2}\sigma\rho_{Sr}\gamma\right)^2 + \left(\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 - \frac{1}{4}\sqrt{(\gamma^2+\sigma^2) + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)\right)^2}}r\\
R &= -\dfrac{\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sqrt{(\gamma^2+\sigma^2)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)}{\sqrt{\left(\frac{1}{2}\sigma\rho_{Sr}\gamma\right)^2 + \left(\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 + \frac{1}{4}\sqrt{(\gamma^2+\sigma^2)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)\right)^2}}x\\
&\quad + \dfrac{\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 - \frac{1}{4}\sqrt{(\gamma^2+\sigma^2)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)}{\sqrt{\left(\frac{1}{2}\sigma\rho_{Sr}\gamma\right)^2 + \left(\frac{1}{2}\sigma^2 - \left(\frac{1}{4}\gamma^2 + \frac{1}{4}\sigma^2 - \frac{1}{4}\sqrt{(\gamma^2+\sigma^2)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)\right)^2}}r.
\end{cases} \tag{4.12}$$

Thus we arrive at an equation of the form

$$\begin{aligned}
v_t &+ \frac{1}{2}\left(\frac{1}{2}\gamma^2 + \frac{1}{2}\sigma^2 + \frac{1}{2}\sqrt{\left(\gamma^2 - \sigma^2\right)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)v_{XX}\\
&+ \frac{1}{2}\left(\frac{1}{2}\gamma^2 + \frac{1}{2}\sigma^2 - \frac{1}{2}\sqrt{\left(\gamma^2 - \sigma^2\right)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}\right)v_{RR} = 0,
\end{aligned} \tag{4.13}$$

that is

$$v_t + \frac{1}{2}\alpha^2 v_{XX} + \frac{1}{2}\beta^2 v_{RR} = 0, \tag{4.14}$$

where

$$\begin{aligned}
\alpha &= \sqrt{\tfrac{1}{2}\gamma^2 + \tfrac{1}{2}\sigma^2 + \tfrac{1}{2}\sqrt{\left(\gamma^2 - \sigma^2\right)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}}\\
\beta &= \sqrt{\tfrac{1}{2}\gamma^2 + \tfrac{1}{2}\sigma^2 - \tfrac{1}{2}\sqrt{\left(\gamma^2 - \sigma^2\right)^2 + 4\sigma^2\gamma^2\rho_{Sr}^2}}
\end{aligned}$$

After performing all these transformations we derived the backward heat equation from equation (4.10). By introducing a new change of variables

$$\tau = T - t,\ \hat{x} = \frac{X}{\alpha},\ \hat{r} = \frac{R}{\beta} \tag{4.15}$$

we finally obtain

$$\begin{cases} v_\tau = \dfrac{1}{2}\left(v_{\hat{x}\hat{x}} + v_{\hat{r}\hat{r}}\right) \\ \\ v(\hat{x},\hat{r},0) = v_0(\hat{x},\hat{r}) = e^{-AF_1(\hat{x},\hat{r})-BF_2(\hat{x},\hat{r})}(e^{F_1(\hat{x},\hat{r})} - K)^+, \end{cases} \tag{4.16}$$

where the function F_1 is such that $x = F_1(\hat{x},\hat{r})$. Furthermore, let F_2 be such that $r = F_2(\hat{x},\hat{r})$. The solution of (4.16) is given by

$$v(\hat{x},\hat{r},\tau) = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{\frac{(\hat{x}-x_1)^2+(\hat{r}-r_1)^2}{-2\tau}} v_0(\hat{x},\hat{r})\ dx_1\ dr_1.$$

This allows us to compute

$$V(x,r,t,\epsilon=0) = e^{Ax+Br+Ct} v\left(F_1^{-1}(x,r), F_2^{-1}(x,r), T-t\right).$$

The 0^{th} and the 2^{nd} term of the asymptotic expansion are given by

$$V_0 = V(x,r,t,\epsilon=0)|_{\delta=0} \tag{4.17}$$

and

$$V_2 = \lim_{\delta\to 0} \frac{V(\epsilon=0) - V_0}{\sqrt{\delta}}. \tag{4.18}$$

We do not derive more explicit formulae for V_0 and V_2. We only mention that V_0 satisfies the normal Black-Scholes equation with volatility $\sigma = \bar{\sigma}$ and interest rate equal to the initial interest rate $r(t=0) = r_0$.

4.4.4 The limit $\delta \to 0$

This section deals with the case $0 < \delta \ll \epsilon \ll 1$. We first let δ tend to 0 in (4.9) and then analyse the resulting PDE for small ϵ via singular perturbation techniques. As δ tends to 0, (4.9) reduces to

$$V_t + \frac{\sigma^2}{2}V_{xx} + \frac{\bar{\eta}^2}{2\epsilon}V_{\sigma\sigma} + \sigma\frac{\bar{\eta}}{\sqrt{\epsilon}}\rho_{S\sigma}V_{x\sigma} + \frac{\bar{\kappa}}{\epsilon}\left(\bar{\sigma}-\sigma\right)V_\sigma + \left(r_0 - \frac{\sigma^2}{2}\right)V_x - r_0 V = 0, \tag{4.19}$$

where $r_0 = r(t=0)$ is the initial condition of the interest rate. As mentioned before, $\delta = 0$ means that the interest rate is constant at leading order on short timescales. Therefore, we take r equal to its initial value r_0.

We can now use known results that can be found, for instance, in [5], Section 5 of [11] and Section 4.4.2 of [12]. The authors apply singular perturbation techniques to a PDE nearly identical to (4.19). It is worth mentioning that the analysis in Section 5 of [11] is very clear and a brief summary of the general perturbation procedure can be found in Section 2.6.2 of [18]. For simplicity, we assume that there is no market price of volatility risk. Hence, we conclude that

$$V_1 = -(T-t)\left(\frac{\bar{\eta}\rho_{S\sigma}}{2}\left\langle\sigma\partial_\sigma\phi\right\rangle S\partial_S\left(S^2\partial_S^2\right)\right)V_0, \tag{4.20}$$

where ϕ solves

$$\left(\frac{\bar{\eta}^2}{2}\partial_\sigma^2 + (\bar{\sigma}-\sigma)\partial_\sigma\right)\phi = \sigma^2 - \bar{\sigma}^2$$

and is chosen in such a way that V_1 satisfies the boundary conditions. Notice that $< . >$ is defined by

$$< f >= \int_{-\infty}^{\infty} f \frac{1}{\sqrt{\pi}\bar{\eta}} e^{-(\bar{\sigma}-\sigma)^2/\bar{\eta}^2}\mathrm{d}\sigma.$$

In (4.20), V_0 is the solution to the normal Black-Scholes equation with average volatility $\bar{\sigma}$ and interest rate $r = r_0$. This results from arguments similar to those mentioned in the previous section.

4.5 Main result. Discussion

The main result of our paper is the expansion given by

$$V = V_0 + \sqrt{\epsilon}V_1 + \sqrt{\delta}V_2 + \mathcal{O}\left(\delta,\epsilon\right), \tag{4.21}$$

where V_0 solves the normal Black-Scholes equation with average volatility $\bar{\sigma}$ and the interest rate $r = r(t=0) = r_0$, V_2 is given by (4.18) and V_1 is given by (4.20).

We have set a first step in applying existing perturbation methods to equation (4.2). Clearly more work has to be done especially concerning the calibration of the approximate solution (4.21) to real market data. If the approximation turns out to be not accurate enough, the one can look at some of the higher order terms (hoping to come closer to what happens in reality). We expect that the analysis of [5] can be extended in this direction. It is expected that evaluation of the approximate solution is much faster than solving the PDE, however there is a tradeoff between speed and accuracy. Once calibration with market data has been performed more can be said about improvements in the speed of computation.

In [5], the authors interpret the corrections to the leading order Black-Scholes approximation in terms of the Greeks (sensitivities). We expect that an intuitive interpretation of the correction factors can give further insight.

Using two small parameters instead of a single one offers *flexibility*. Instead of having two small parameters δ and ϵ one may be tempted to deal with a single one, i.e. $\delta = \mathcal{O}(\epsilon)$. However, we expect this later choice to essentially complicate the perturbation analysis.

We want to stress the fact that the validity of the formal perturbation approach is restricted by the conditions under which the pricing PDE with the imposed initial and boundary conditions is well-posed. It would be particularly interesting to study the effect of the degeneracy in the coefficients of the 2nd order derivatives on the solution of the PDE. Another open question is: What happens with the well-posedness of the model, and hence, with the approximate solution (4.21) if other boundary conditions are chosen instead of (4.5).

A completely different modeling approach is the so called random field approach. Let us sketch a very simple version of this idea. Consider the SDE $\mathrm{d}S_t = rS_t\mathrm{d}t + \sigma S_t \mathrm{d}W_t^S$ and, for the moment, let σ and r be given constants. The Fokker-Planck equation for the probability distribution p of variables S and t is given by

$$\frac{\partial p}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 Sp}{\partial S^2} - \mu\frac{\partial Sp}{\partial S}.$$

If we now take μ and σ random in the above Fokker-Planck equation, then we are immediately led to *random fields*. Perturbation methods can also be applied to the resulting PDE; see, for instance, [6, 7, 15] and references therein.

We have been surprised that the seemingly straightforward problem that we addressed happened to be a box of Pandora, leaving open a lot of relevant mathematical problems of which this project is not the right framework to elaborate on. Particularly, we would like to stress that the proposed methods have not been tested at all and large deviations from reality may have been neglected.

Acknowledgments

We kindly acknowledge reviewer's comments which helped us to shape the final version of this paper and thank Rabobank for posing this interesting problem to SWI 2009. We hope that our contribution will be helpful in making a step forward towards understanding the role of the stochastic volatility when pricing European options.

Bibliography

[1] Y. Achdou, B. Franchi, and N. Tchou. A partial differential equation connected to option pricing with stochastic volatility: regularity results and discretization. *Mathematics of Computation*, 74(251):1291–1322, 2004.

[2] Y. Achdou and N. Tchou. Variational analysis for the stochastic Black and Scholes equation with stochastic volatility. *Mathematical Modelling and Numerical Analysis*, 36(3):373–395, 2002.

[3] M. Böhm and R. E. Showalter. Diffusion in fissured media. *SIAM J. Math. Anal.*, 16(3):500–509, 1985.

[4] L. C. Evans. An Introduction to Stochastic Differential Equations. Department of Mathematics, UC Berkley, 2008.

[5] J.-P. Fouque, G. Papanicolaou, R. Sircar, and K. Solna. Multiscale stochastic volatility asymptotics. *Siam J. Multiscale Model. Simul.*, 2(1):22–42 (electronic), 2003.

[6] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004.

[7] J. Grasman and O. A. van Herwaarden. *Asymptotic Methods for the Fokker-Planck Equation and the Exit Problem in Applications*. Springer Series in Synergetics. Springer-Verlag, Berlin, 1999.

[8] L. A. Grzelak and C. W. Oosterlee. On the Heston model with stochastic interest rates. *SSRN: http://ssrn.com/abstract=1382902*, 2009.

[9] L.A. Grzelak, Oosterlee C.W., and S. van Weeren. Extension of Stochastic Volatility Equity Models with Hull-White Interest Rate Process. *Technical Report 2008 TUDelft. Available at: http://ssrn.com/abstract=1344959*.

[10] A. Haastrecht, R. Lord, A. Pelsser, and D. Schrager. Pricing Long-Maturity Equity and FX Derivatives with Stochastic Interest and Stochastic Volatility. *SSRN: http://ssrn.com/abstract=1125590*, 2008.

[11] S. Howison. Matched asymptotic expansions in financial engineering. *J. Engrg. Math.*, 53(3-4):385–406, 2005.

[12] S. Howison, A. Rafailidis, and H. Rasmussen. On the pricing and hedging of volatility derivatives. *Applied Mathematical Finance*, 11(4):317–346, 2004.

[13] P.-L. Lions and M. Musiela. Correlations and bounds for stochastic volatility models. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 24(1):1–16, 2007.

[14] B. Øksendal. *Stochastic Differential Equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An Introduction with Applications.

[15] D. Repplinger. *Pricing of Bond Options*, volume 615 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 2008. Unspanned Stochastic Volatility and Random Field Models.

[16] L. O. Scott. Option pricing when the variance changes randomly: theory, estimation, and an application. *The Journal of Financial and Quantitative Analysis*, 22(4):419–438, 1987.

[17] P. Wilmott, S. Howison, and J. Dewynne. *The Mathematics of Financial Derivatives*. Cambridge University Press, Cambridge, 1995. A Student Introduction.

[18] X. Zhou. *Application of Perturbation Methods to Modeling Correlated Defaults in Financial Markets*. PhD thesis, North Carolina State University, 2006.

Chapter 5
Stiffening while drying

Frits van Beckum [1] Jan Bouwe van den Berg [1] Sören Boettcher [2] Maarten de Gee [3]
Kundan Kumar [4] Joost van Opheusden [3]

abstract:
We present two models for the drying of waterborne paints, which consist of non-volatile latex particles suspended in water. One model considers the water and latex density in a layer as a function of time. Water evaporation at the surface represents the drying. This model results in a one-dimensional free boundary problem, which is solved numerically. Extensions to the model are given that describe the stiffening of the paint. A second model is a particle based dynamical simulation where latex particles form a network through which water particles move. A thin slab of the suspension in a three-dimensional box is studied. Water particles escaping the slab at the surface represent the drying, progressing network formation the stiffening of the paint. Both models allow for validation with material properties as determined experimentally on real coatings.

KEYWORDS: *mathematical modelling, free boundary, liquid coating, evaporation*

[1]VU University Amsterdam, The Netherlands
[2]Center for Industrial Mathematics, University of Bremen, Germany
[3]Wageningen University and Research Centre, The Netherlands
[4]Eindhoven University of Technology, The Netherlands

5.1 Introduction

Waterborne coatings (WBC's) are increasingly replacing traditional organic solvent-borne coatings (SBC's) due to stricter legislation originating from an ever-growing awareness in society about environmental issues and safety at work. One of the few disadvantages of WBC's is that they generally require longer drying times as compared to SBC's. For that reason the drying process of WBC's has received considerable attention from researchers in both industry and academia. From an industrial and applicational point of view, most interest lies in the understanding and control of the simultaneous drying and stiffening of the coating or film. If a coating needs to be handled or post-processed it should have adequate mechanical integrity, such as a sufficient shear stiffness (resistance to shearing), to prevent it from incurring damage. The formation of film at the surface, that is relatively solid, while the paint layer below is still relatively liquid is a common way to realize this in practice. On the other hand such a film can prevent further drying of the liquid paint layer below, which could compromise the required integrity again. Insight in how the detailed balance can be found is of prime importance. Modern waterborne paints can be considered

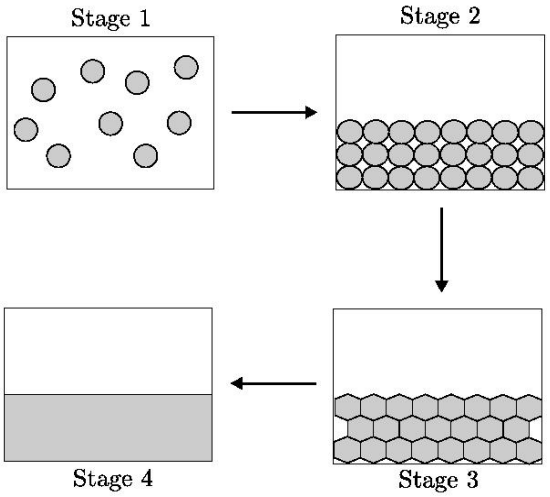

Figure 5.1: Four stages of latex film formation (cf. [4])

as a stabilized suspension of (latex) polymer particles in water. It is usually applied as homogeneous thin layers on a hard substrates. The drying process can be seen to consist of four stages (see figure 5.1). When the paint layer is applied, it is still the original suspension (stage 1). Due to disappearance of the water as a result of evaporation a concentrated latex

mass is formed, in which the polymer particles come into close contact (stage 2). Polymer particles are then subsequently deformed by the contact forces, while further water is removed by capillary forces, until most water has gone and particles start coalescing (stage 3). In the final stage particle boundaries disappear when they coalesce further to form a continuous polymer melt that further develops its mechanical integrity (stage 4). All four stages (in particular the second) have already been addressed in the literature (cf. [4]), but the development of the mechanical integrity of the layer (transition from liquid to solid) has hardly been touched upon.
The goal of this study is to develop mathematical models for the drying of waterborne coatings.

5.2 Derivation of the model

Waterborne paints consist of a stabilized suspension of particles in water. It is usually applied as a homogeneous film on a hard substrate. We will first develop a model in which we describe the drying and stiffening of this paint layer in terms of concentrations of water and latex.

5.2.1 One-dimensional model

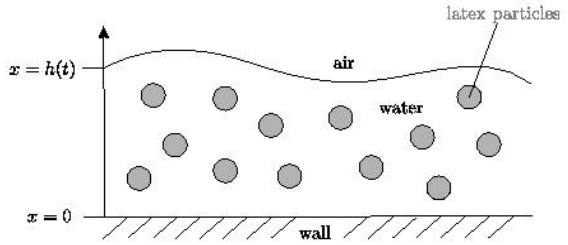

Figure 5.2: Sketch of the two-dimensional paint domain (cf. [6])

Our first model describes the shrinkage of the wet paint layer due to water evaporation as a one-dimensional process. All variation parallel to layer is ignored, we consider only variations in the perpendicular (x) direction. The impermeable substrate is at $x = 0$, the layer surface is at $x = h(t)$ (see figure 5.2). Evaporation causes the layer to shrink, hence we have a moving surface. The paint consists of latex particles and water, with volume

fractions $p(x,t)$ and $w(x,t)$, respectively. We assume there are no air bubbles in the paint, so the relation

$$w(x,t) + p(x,t) = 1 \tag{5.1}$$

holds for all time and position. This will allow us to formulate the problem in terms of h and w only, eliminating p from the equations. The particles move by diffusion in the water, with diffusion constant $D > 0$:

$$p_t(x,t) = (D\, p_x(x,t))_x \quad \text{for} \quad 0 < x < h(t),\, t > 0. \tag{5.2}$$

In this model the diffusion coefficient D can be state-dependent, i.e. , D can be a function of the particle fraction p (or equivalently, through (5.1), of w). We come back to this later. By taking into account equation (5.1) within our model the water phase satisfies a similar diffusion equation :

$$w_t(x,t) = (D\, w_x(x,t))_x \quad \text{for} \quad 0 < x < h(t),\, t > 0. \tag{5.3}$$

Note that this does not imply we assume the water itself moves diffusively, which would be physically incorrect. As long as the motion of the latex particles is mainly diffusional, our model applies. At the substrate there is no flux of water or latex:

$$w_x(0,t) = 0, p_x(0,t) = 0. \tag{5.4}$$

The free and moving upper surface takes into account the water evaporating from the layer, while the total amount of liquid remains the same. Analogous to Newtons law of cooling, we assume the evaporation rate is proportional to $w(h(t),t) - Hw_{amb}$, where w_{amb} is some ambient water concentration and H is a Henry coefficient. As the total volume of water is given by $\int_0^{h(t)} w(x,t)\,dx$, we use that the change in this volume is given by the evaporation

$$\begin{aligned}
-\alpha(w(h(t),t) - Hw_{amb}) &= \frac{d}{dt}\int_0^{h(t)} w(x,t)\,dx && (5.5)\\
&= h'(t)w(h(t),t) + \int_0^{h(t)} w_t(x,t)\,dx && (5.6)\\
&= h'(t)w(h(t),t) + \int_0^{h(t)} (Dw_x(x,t))_x\,dx && (5.7)\\
&= h'(t)w(h(t),t) + [Dw_x(x,t)]_{x=0}^{x=h(t)} && (5.8)\\
&= h'(t)w(h(t),t) + Dw_x(h(t),t), && (5.9)
\end{aligned}$$

where α is a positive constant. From the assumption that the volume of the latex fraction in the drying layer is conserved, similarly, one finds for the polymer particles:

$$0 = \frac{d}{dt}\int_0^{h(t)} p(x,t)\,dx \tag{5.10}$$

$$= h'(t)(1 - w(h(t),t)) - \int_0^{h(t)} w_t(x,t)\,dx \tag{5.11}$$

$$= h'(t)(1 - w(h(t),t)) - Dw_x(h(t),t). \tag{5.12}$$

Combining these equations, one finds

$$h'(t) = -\alpha(w(h(t),t) - Hw_{amb}), \tag{5.13}$$

which establishes a constituting equation for the thickness h of the paint layer, and

$$-\alpha(w(h(t),t) - Hw_{amb})(1 - w(h(t),t)) = Dw_x(h(t),t), \tag{5.14}$$

which establishes a boundary condition for w at the moving surface. Finally we choose for our one-dimensional model an initial thickness of the paint layer, and an initial water (and latex particle) distribution, which will typically be uniform. Thus, the system of equations for the volume fraction w with layer thickness h is given by:

$$w_t(x,t) = (Dw_x(x,t))_x, \tag{5.15}$$

$$w_x(0,t) = 0, \tag{5.16}$$

$$w_x(h(t),t) = -\frac{\alpha}{D}(w(h(t),t) - Hw_{amb})(1 - w(h(t),t)), \tag{5.17}$$

$$w(x,0) = w_0(x), \tag{5.18}$$

$$h'(t) = -\alpha(w(h(t),t) - Hw_{amb}), \tag{5.19}$$

$$h(0) = h_0 \tag{5.20}$$

for $0 < x < h(t)$ and $t > 0$.

Note that equation (5.1) allows us to find the associated volume fraction profile of the latex, which is in practice the more relevant physical property. So far we have not specified the diffusion coefficient D of the latex particles. In the dilute, and possibly also the semi-dilute regime it could well be taken constant, but at higher densities that would not be a very realistic approximation. One possible choice is a Heaviside function to account for the transition from a liquid to a solid phase. In the computations later the approximation $D = D_s + d_w w$ with $D_s, d_w > 0$ was used, to avoid problems with the discontinuity.

Typically this choice describes a low diffusivity at low water content, and a high one for more dilute situations. Note that w is a volume fraction, so its value can not be larger than unity. Note that we have left the equations in a dimensional form, which hopefully makes them easier to interpret for non-mathematically oriented readers. A simple dimension analysis shows that there two important time scales involved in the process; α/h_0 is the rate at which the thickness of the layer is decreasing (initially) due to the evaporation of the water. A second time scale is given by the D/h_0^2, the rate at which the diffusion is able to transport the water over the full layer. Important is their ratio $\epsilon = D/\alpha h_0$. If ϵ is small, the diffusive rate is not able to compensate the water loss at the surface quickly enough to keep the layer homogeneous. Close to the surface the water content w will drop, and we have a dry surface layer. The drying is then predominantly governed by the rate at which the water from the lower part of the layer permeates this dry film. If on the other hand ϵ is large, the diffusion will keep the water concentration throughout the layer the same, and the evaporation is the limiting process.
We note that this derivation of the mathematical model for drying of a paint layer is similar to [2, 3, 4, 6, 9]. Furthermore, in [8] stress-driven diffusion was incorporated in such a model. We also refer the interested reader to the existence and uniqueness results obtained in [7].

5.2.2 Clustering and stiffening

The above model can describe the drying of the paint layer, but neglects all detail of the latex phase. Already at moderate volume fractions one may expect particles to cluster and possibly even coalesce, thus adding to the mechanical stability of the material. In this section we discuss an extension to the basic model, in which cluster formation is taken into account by not considering a single particle volume fraction, but a series of volume fractions, one for each cluster size. The silent assumption here is that there is something like a primary latex particle, a monomer, that can be identified as such. In reality the particles in the original suspension, already before application of the paint and the initiation of the drying process, have a range of sizes (polydispersity), and there will be clusters, maybe small and reversible.
In the next stage of model development, we incorporate the effect that latex particles may form clusters. Let n be the cluster size, i.e. the number of particles in the cluster. Then $P(n, x, t)$ is defined as the number of clusters of size n, multiplied by the volume of one particle, divided by unit volume. Thus $nP(n, x, t)$ is the joint volume fraction of the clusters of size n, and since the total volume fraction of all components, including the water, is

unity, we have

$$1 - w(x,t) = \sum_{n=1}^{N} nP(n,x,t),\ 0 < x < h(t),\ t > 0, \tag{5.21}$$

where N denotes the upper bound to the cluster size. In principle this N can be infinite, in practice we must choose some finite value of course; the model does not include gel formation. The diffusion rate $D = D_n$ now depends also on the cluster size. The larger clusters tend to have a smaller diffusion coefficient than the smaller ones. Coagulation takes place with a certain probability when two smaller clusters meet, resulting in a reaction-diffusion equation for each separate cluster size

$$P_t(n,x,t) = (D_n P_x(n,x,t))_x - A(n,x,t) + B(n,x,t). \tag{5.22}$$

Here

$$A(n,x,t) = \sum_{m=1}^{N-n} C_{n,m} P(n,x,t) P(m,x,t) \tag{5.23}$$

denotes the loss of clusters of size n (dissipation rate) due to further aggregation, while

$$B(n,x,t) = \frac{1}{2} \sum_{m=1}^{n-1} C_{n,n-m} P(m,x,t) P(n-m,x,t) \tag{5.24}$$

stands for the gain in clusters of that size by aggregation of smaller ones. The factor $\frac{1}{2}$ is present to take care of "double counting". It is reasonable to assume that the coefficients $C_{n,m}$ increase as function of m and n since the chance of hitting a large particle is larger than the chance of hitting a small particle. When the form of a cluster is a chain, its surface is about proportional to n. The same applies to planar clusters. However, for a spherical cluster, the surface is proportional to $n^{\frac{2}{3}}$. Thus, a sensible model for probability would be an exponential law $C_{n,m} \cong (nm)^b$ where b is between $\frac{2}{3}$ and 1.
Similar choices have to be made in modelling the diffusion rate. Among others, the diffusion rate is affected by the size of a cluster and its affinity to water. It seems natural to assume that the diffusion rate decreases with the cluster size. In the numerical computations the

model $D_n = n^c D$ with $c = -1$ is used. The full set of equations now reads

$$P_t(n,x,t) = (D_n P_x(n,x,t))_x - A(n,x,t) + B(n,x,t) \tag{5.25}$$

$$P_x(n,0,t) = 0 \tag{5.26}$$

$$D_n P_x(n,h(t),t) = -h'(t) P(n,h(t),t) \tag{5.27}$$

$$w(x,t) = 1 - \sum_{n=1}^{N} nP(n,x,t) \tag{5.28}$$

$$h'(t) = -\alpha(w(h(t),t) - Hw_{amb}) \tag{5.29}$$

for $0 \leq x \leq h(t)$ and $t > 0$.

Stiffness is modelled using the assumption that it increases with cluster size. More precisely, the full sample stiffness is determined as a harmonic mean of local stiffnesses, that in turn are determined by the local cluster size distributions:

$$\frac{1}{S(t)} = \frac{1}{h(t)} \int_0^{h(t)} \frac{1}{S_{local}(x,t)}\, dx,\ t > 0 \tag{5.30}$$

$$S_{local}(x,t) = \sum_{n=1}^{N} a(n) P(n,x,t),\ 0 < x < h(t),\ t > 0 \tag{5.31}$$

where a should be a convex function of n satisfying $a(n+m) > a(n) + a(m)$, otherwise clustering would not enhance stiffness. Without detailed physio-chemical information on the composition of the paint and the nature of the aggregation process, this function cannot be specified further. In the numerical computations $a(n) = n^2$ is chosen.

5.3 Numerical implementation

A transformation to a fixed domain makes the problem numerically treatable. Using the method of lines (spatial semi-discretization), the system of partial differential equations (PDE's) is approximated by a system of ordinary differential equations. A numerical solution can be obtained with standard procedures in MATLAB$^{\text{TM}}$(cf. [1, 5]). The moving boundary is transformed into a fixed one by introducing the new variables $\xi = \frac{x}{h(t)}$ and (somewhat formally) $\tau = t$. Hence

$$\frac{\partial}{\partial x} = \frac{1}{h(t)} \frac{\partial}{\partial \xi}, \tag{5.32}$$

$$\frac{\partial}{\partial t} = -\frac{\xi h'(t)}{h(t)} \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \tau}. \tag{5.33}$$

Let us make all this explicit for the non-clustering model. A similar,more elaborate system of equations was used to study the development of the cluster distribution in the layer. Writing $t = \tau$, equation (5.15) reads:

$$w_t(\xi,t) - \frac{\xi h'(t)}{h(t)} w_\xi(\xi,t) = \frac{1}{h(t)} \left(\frac{D}{h(t)} w_\xi(\xi,t) \right)_\xi . \tag{5.34}$$

This is a diffusion equation with a pseudo convection term:

$$w_t(\xi,t) = \frac{\xi h'(t)}{h(t)} w_\xi(\xi,t) + \left(\frac{D}{h^2(t)} w_\xi(\xi,t) \right)_\xi , \; 0 < \xi < 1, \; t > 0 \tag{5.35}$$

Thus we have as physical transport model

$$w_t(\xi,t) = \frac{\xi h'(t)}{h(t)} w_\xi(\xi,t) + \left(\frac{D}{h^2(t)} w_\xi(\xi,t) \right)_\xi , \; 0 < \xi < 1, \; t > 0 \tag{5.36}$$

$$w_\xi(0,t) = 0 \tag{5.37}$$

$$w_\xi(1,t) = -\frac{\alpha h(t)}{D} (w(1,t) - H w_{amb})(1 - w(1,t)) \tag{5.38}$$

$$w(\xi,0) = w_0(\xi) \tag{5.39}$$

$$h'(t) = -\alpha (w(1,t) - H w_{amb}) \tag{5.40}$$

$$h(0) = h_0 \tag{5.41}$$

for $0 < \xi < 1$ and $t > 0$.

Figures 5.3 and 5.5 show the joint volume fraction $P(n,x,t)$ of the clusters of size n, divided by the number of latex particles in the cluster. In the upper left plot, the volume fraction of the latex particles is presented as a function of the space variable for different time steps. Initially ($t = 0$), the volume fractions of the latex particles and the water are taken the same $p_0 = w_0 = 0.5$, no clusters are present (above size one). The volume fraction of single latex particles decreases with time because of cluster formation. This process is faster at the upper boundary ($x = 1$) because the evaporation takes place at the surface of the paint layer, hence the drying process there is more rapid than at the lower boundary ($x = 0$). The water near the substrate must first be transported through the layer before it can escape to the air. According to our choice of parameter values, the rate of the diffusional transport of water is rather small compared to that of the evaporation rate, so we should expect the formation of a film. The other plots show similar results for the clusters of size 2, 3 and 4.
Figures 5.4 and 5.6 show the volume fraction of water $w(x,t)$ (left) as a function of x for

certain values of t and the thickness of the paint layer $h(t)$ (mid) as well as the stiffness $S(t)$ as a function of time. As the initial thickness is $h_0 = 1$ and the initial water fraction is chosen $w_0 = 0.5$, the paint layer shrinks to a thickness $h = 0.5$ with the evaporation of the water. The stiffness increases with time because of progressing cluster growth.

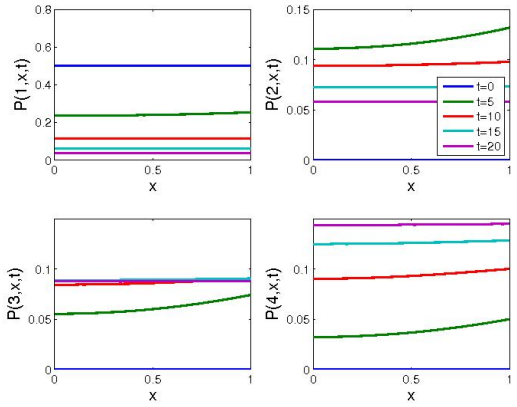

Figure 5.3: The joint volume fraction $P(n, x, t)$ of the clusters of size n, divided by the number of particles with initial data $\alpha = 0.5$, $Hw_{amb} = 0$, $w_0 = 0.5$, $h_0 = 1$, $C_1 = 0.5$, $D_s = d_w = 0.1$.

5.4 Particle simulation

A completely different approach is chosen in the model we call the particle simulation. Here both the water and the latex phase are described as soft particles, moving diffusively due

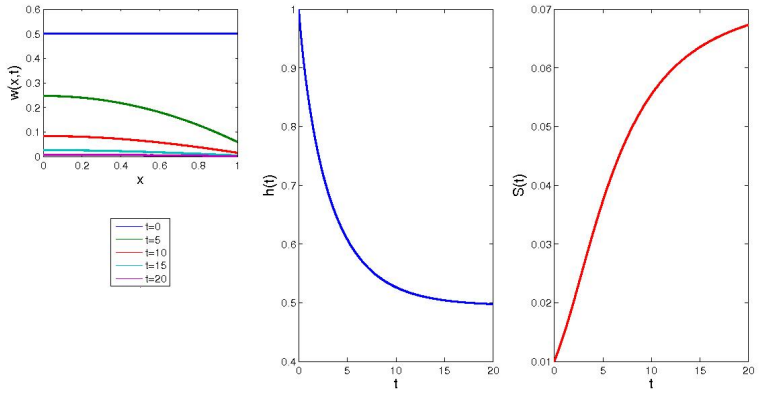

Figure 5.4: The volume fraction of water $w(x,t)$ (left), the thickness of the paint layer $h(t)$ (mid) and the stiffness $S(t)$ (right) as a function of time with parameter values $\alpha = 0.5$, $Hw_{amb} = 0$, $w_0 = 0.5$, $h_0 = 1$, $C_1 = 0.5$, $D_s = d_w = 0.1$.

to the effect of thermal fluctuations. The interaction between the particles belonging to the two different phases is described by a potential force. Moreover the latex particles can form bonds when they come into close contact and form clusters. When the bonds are stretched, for instance by cluster reorganization due to stresses in other parts of the cluster, external forcing, or collisions with other clusters or water particles, these bonds may break. The forces as generated by the potentials can be directly calculated, averaged over the sample and related to material properties of the sample as a whole. The thermal fluctuations are represented by a random force. Mathematically the motion of the individual particles is described by a Langevin equation of motion, that is integrated numerically. The generic term for these type of simulation is Brownian Dynamics (BD), with the random force generating the Brownian (diffusive) motion of the particles. The latex particles in the model are indeed supposed to be like those in the actual system, and the water particles stand for relatively large volumes of water, about the same size as the latex particles (the order of micrometers), many length scales above the size of water molecules (less than a nanometer). Hence the appropriate term of the particle model scale would be mesoscopic. By calculating densities of particles, a relation could be made with the previously describe continuous model. Note that the particle simulation is fully three-dimensional. Most of the techniques we use are quite common, and are described in detail in many standard textbooks on molecular and mesoscopic simulation techniques.

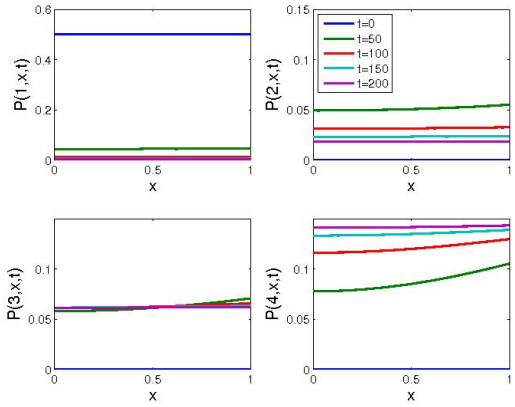

Figure 5.5: The joint volume fraction $P(n, x, t)$ of the clusters of size n, divided by the number of particles with parameter values $\alpha = 0.2$, $Hw_{amb} = 0.2$, $w_0 = 0.5$, $h_0 = 1$, $C_1 = 0.2$, $D_s = d_w = 0.01$.

5.4.1 Model description

In the BD model we have latex particles (yellow) and water particles (cyan) (cf. figure 5.7). The latex particles form clusters, and the largest cluster in the system is shown in orange. The interactions between the particles are a constant force with a finite range, attractive for the latex particles and repulsive for the water particles. Also the interaction between latex and water particles is repulsive. When latex particles come close, a bond may be formed. When that occurs the length of the bond is subject to a simple Hookean potential, a linear spring, that breaks again above a certain extension. Overlap between particles is removed by a repulsive force with the same Hookean potential. The whole system is contained in an image box, and we use periodical boundary conditions in all directions. The equations of motion for the particles are integrated numerically in time using an Euler

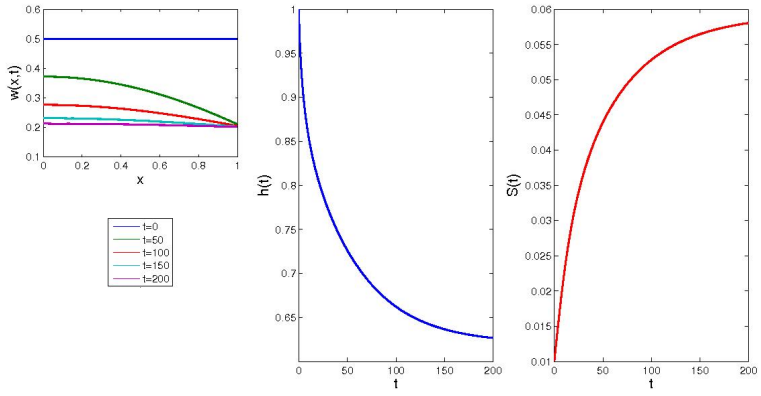

Figure 5.6: The volume fraction of water $w(x,t)$ (left), the thickness of the paint layer $h(t)$ (mid) and the stiffness $S(t)$ (right) as a function of time with parameter values $\alpha = 0.2$, $Hw_{amb} = 0.2$, $w_0 = 0.5$, $h_0 = 1$, $C_1 = 0.2$, $D_s = d_w = 0.01$.

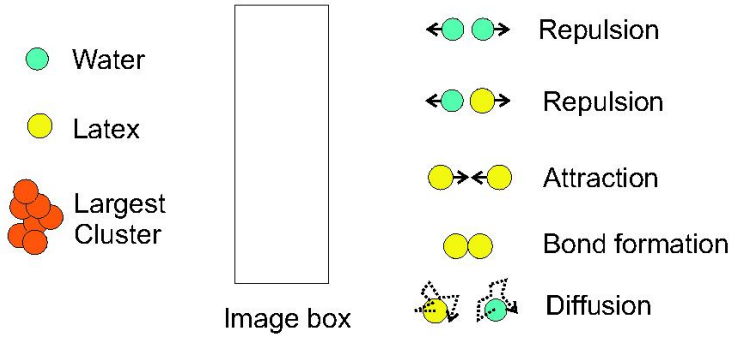

Figure 5.7: The elements of the particle simulation model.

Forward method. The time step is physically restricted by the oscillation period of the linear springs, higher order methods would not allow for substantially larger time steps.

5.4.2 Results

As the implementation of such a model is quite elaborate, we used existing private code that we modified for our purpose of the drying of a paint layer. The initial configuration consists of 200 latex and 100 water particles randomly positioned in a thin slab in the xz-plane of the image box. As the box is elongated in the vertical y-direction, and periodic in the other two directions, the system actually describes an infinite paint layer freely

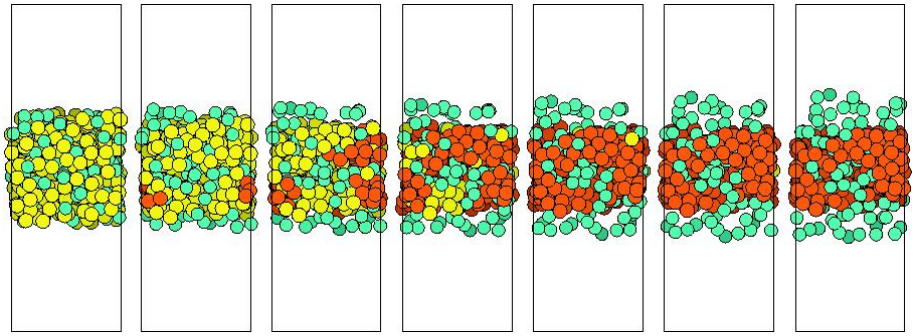

Figure 5.8: Series of snapshots from the BD simulation, as time increases water evaporates from the layer and the latex particles form a single cluster.

floating in space. The water particles may escape from the layer, describing the drying, the latex particles will form bonds and stay within the layer, describing the stiffening. There is no actual substrate present, though the model as such would allow it, unfortunately the available code does not.

Figure 5.8 depicts what happens as a function of time by providing a series of snapshots of the model system. One observes indeed the blue particles escaping the layer, be it slowly. In the current realization they still stay close to the layer, indicating that diffusion even outside the layer is slow. A relatively large number of water particles is still inside the layer, even when all the latex particles have joined into a single cluster, forming a solid slab. Some water will be trapped inside holes in that solid, other may still escape later through channels as the slab continues contracting. Apart from evaporation and coagulation, the particle simulation also describes the compactification of the partially dried paint layer.

Since the detailed motions of all individual particles and all the forces in the system that influence that motion, are readily available, a host of numerical data is available in this model system, even for relatively small system size and short simulation times. In practice one would calculate from these data observables that can be compared with actual experimental data on real samples. For the current feasibility study we present as an example the largest cluster size in the system. In fact such a parameter is not at all readily accessible in real systems, but the pronounced S-shape of the curve depicted in figure 5.9 agrees very well with results from particle gelation models. At early times there are many small clusters, which move relatively fast. When larger clusters form, which diffuse more slowly, the cluster-cluster aggregation leads to a rapid growth of large clusters when they come into contact not because of their diffusion, but rather of their growth. Once a fraction of the particles has formed the gel, that gel grows by the addition of smaller clusters joining

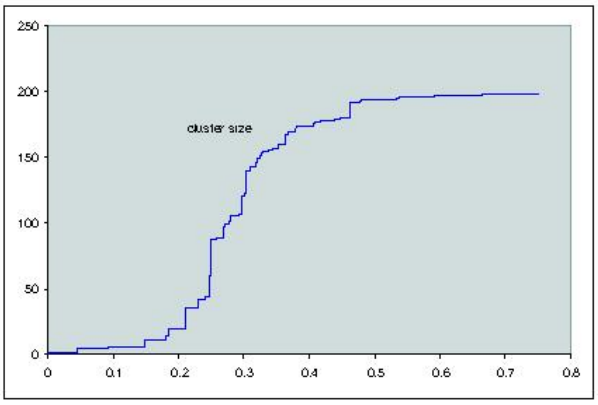

Figure 5.9: Growth of the largest cluster of latex particles in the system.

it. Slowing down of the diffusion of larger clusters is incorporated in the model directly through the random force acting on the individual particles. The larger the cluster, the more these random forces will average out.

To calculate material properties of the sample, like surface tension of the drying layer, it

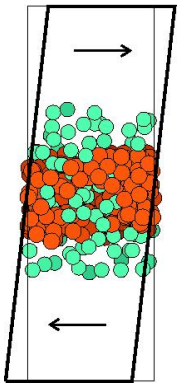

Figure 5.10: Affine deformation of the sample with the image box under shear.

suffices to sample the data from the simulation as described above. Such parameters are said to be calculated in equilibrium, using the fluctuations generated by the random force, and relations from equilibrium statistical physics, like the virial theorem, to relate those to macroscopic observables. Because the random forces are small, and consequently so are the fluctuations, only material properties for small deformations can be determined. To investigate how such a sample would behave under larger deformations, non-equilibrium techniques are used. One example that is depicted in figure 5.10 is shear deformation,

in this case shearing of the whole layer in the parallel direction. All particles are moved affinely under this deformation, while the image box is changed accordingly using the so called Lees-Edwards boundary conditions. This induced shearing motion leads to increasing stresses in the material that can only partially be relaxed by internal reorganization.

In figure 5.11 we give the sample stress as a function of shear strain parallel to the layer.

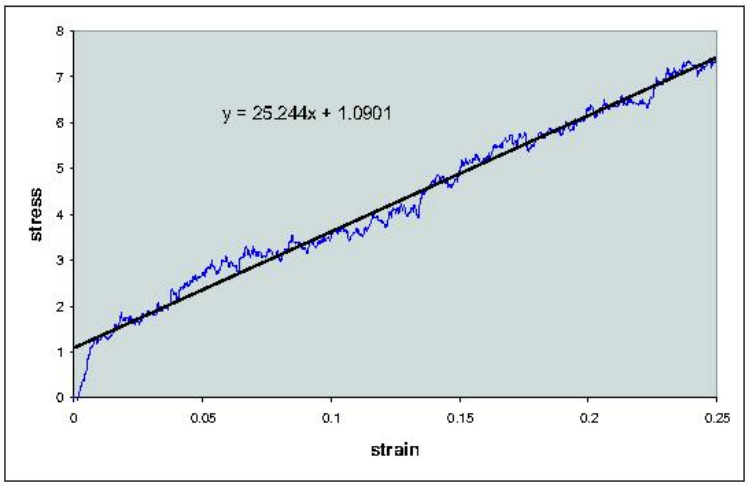

Figure 5.11: Stress in the sample as a function of imposed shear strain. Bonds do not break here and the response is quite linear.

The simulation was performed on a sample that had fully gelated, all latex particles form a single cluster. Since the deformation is rather small the system response is quite linear, apart from a small effect at very low deformation. Bonds are not stretched to the point where they start breaking, and the sample reacts as an elastic solid. A soft solid for that matter, since a strain of 0.25 for a stiff solid would be far in the non-linear regime. Stiffness is readily calculated within this model simulation, directly from the forces in the cluster network.

5.5 Conclusions and discussion

We have presented two different modelling approaches for the understanding of the drying and stiffening process of a paint layer. The first approach was a PDE model where we combined a moving boundary value problem, with a model for the stiffening of the paint. For the latter a coagulation process of the latex particles has been incorporated. The numerical studies of this PDE model show that it leads to reasonable results, which at

this stage can not be be compared with those of actual samples. One reason is that we did not have access to actual data, more important is that the model contains a large number of parameters, such as diffusion constants and aggregation probabilities. Moreover we can only expect the model to be applicable in the range where the motion of the latex particles is largely diffusional (stage 1 of the process as described in the introduction). The aggregation model finally does not take into account a sol and a gel phase. In fact that relates directly to the proviso we mention about the diffusion, a gel is actually not much more than a very large cluster that has ceased diffusing, but within an open gel still diffusion of the sol phase might take place. If, as again suggested by the intuitive model from the introduction, large rearrangements in the clusters and the gel, and actual coagulation and deformation of the latex particles plays a large role in the development of the system, a gelation model would not provide much additional insight. In principle it would be possible within the model we present to include ageing effects of the clusters by making the parameters depend on the time passed since the formation of the cluster, like in structured population models developed in biology. Whether the addition of yet another set of model parameters will add to the predictability of the model for actual paint samples is obviously quite questionable. Possible extensions of this work could be the investigation of stress-driven water flow, the cracking of the paint film, two- and three-dimensional flow and the physics of the film formation process. One main shortcoming of the model as presented is that it does not allow for external disturbances. Forced drying could be taken into account by changing the parameter α, but mechanical properties are restricted to the rather phenomenological description of the stiffness. The numerical calculations did show the feasibility of the application of the model, with modest requirements as to calculational equipment, and using quite straightforward simple implementation techniques.

Direct particle simulations provide an alternative approach for modelling. The advantage is that most real system observables can be directly compared to model results, no additional modelling is needed once the potential force parameters are specified. The disadvantage is that such potentials often present a too much simplified physical picture of reality. The model we used does not allow for deformation of the latex particles, though the effect is taken into account somewhat by the flexible bonds between the particles. The diffusional motion of the water particles in the early stage may adequately describe the dynamics of such samples, once the water has evaporated it certainly is an awkward caricature at best. Maybe it would better if such particles, when they have actually escaped from the layer, are removed in full. The rate at which that happens could then for instance stand for a form of forced drying by ventilation. Still the large size of the water particles limits the applicability of this model. Smaller sizes can be included straightforwardly, but at the

price of an considerable increase in computational effort. Much more promising seems a hybrid approach, where the latex particles are described by a Brownian Dynamics type of model, while for the water phase for instance a coupled Lattice Boltzmann type of model is used. Similar hybrid models are used extensively in CFD techniques to describe the hydrodynamics of particle laden flows. Without significant material flow, as is the case for the drying paint layer, simplified versions of such models may well apply.

Bibliography

[1] R. Ashino, M. Nagase, and R. Vaillancourt. Behind and beyond the MATLAB ODE suite. *Computers and Mathematics with Applications*, 40(4-5):491–512, 2000.

[2] S. D. Howison, J.A. Moriarty, J. R. Ockendon, E. L. Terrill, and S. K. Wilson. A mathematical model for drying paint layers. *Journal of Engineering Mathematics*, 32:377–394, 1997.

[3] A. F. Routh and W. B. Russel. Horizontal Drying Fronts During Solvent Evaporation from Latex Films. *Journal of the American Institute of Chemical Engineers*, 44(9):2088–2098, 1998.

[4] A. F. Routh and W. B. Russel. Deformation Mechanisms during Latex Film Formation: Experimental Evidence. *Industrial and Engineering Chemistry Research*, 40:4302–4308, 2001.

[5] L. F. Shampine and M. W. Reichelt. The MATLAB ODE suite. *SIAM Journal on Scientific Computing*, 18(1):1–22, 1997.

[6] B. W. van de Fliert. A free boundary problem for evaporating layers. *Nonlinear Analysis*, 47:1785–1796, 2001.

[7] B. W. van de Fliert and R. van der Hout. A nonlinear Stefan problem with negative latent heat, arising in a diffusion model. Technical Report W98-11, Leiden University, 1998.

[8] B. W. van de Fliert and R. van der Hout. Stress-driven diffusion in a drying liquid paint layer. *European Journal of Applied Mathematics*, 9(5):447–461, 1998.

[9] B. W. van de Fliert and R. van der Hout. A generalized Stefan problem in a diffusion model with evaporation. *SIAM Journal on Applied Mathematics*, 60(4):1128–1136, 2000.

Chapter 6
DHV water pumping optimization

Simon van Mourik[1] Joris Bierkens[2] Hans Stigter[1] Martijn Dirkse[3] Karel Keesman[4] and Vivi Rottschäfer[5]

abstract:
This contribution investigates the possibilities for optimizing a drinking water network over a horizon of 48 hours, given variable water demands, energy prices and constraints on the pumping strategy and water levels in the reservoirs. Both the dynamic model and goal function are non-linear in the control inputs, the pump flow rates. Since each pump can be switched on or off every 15 minutes and since there are 15 pumps in the system, for a horizon of 48 hours there are $2^{(4*48*15)}$ switching possibilities. Obviously, this problem is too big to solve it in real-time by enumeration. Hence, a decomposition of the problem is needed. Relaxing the constraints and assuming a continuous-time flow rate, allows a (semi)-analytical solution using Lagrangian theory. Furthermore, a numerical solution of the constrained optimization problem is found by using the TomLab PROMPT toolbox. The conversion from a continuous-time pump flow rate to a strategy with on/off switching is also investigated, as well as the possibility of linear feedback control. The resulting trajectories of the pump flow rates and water levels in the reservoirs are realistic and can be physically interpreted.
KEYWORDS: *Dynamic optimization, modeling, feedback control, drinking water network.*

[1]Biometris, Wageningen University, The Netherlands
[2]Leiden University, The Netherlands
[3]Ecofys Netherlands BV, Utrecht, The Netherlands
[4]Systems & Control group, Wageningen University, The Netherlands

6.1 Introduction

For the 67th Studygroup Mathematics with Industry held at the University of Wageningen, we worked on a question posed by DHV which is an international group of consulting engineers located in Amersfoort. The question concerned water pump optimization. We were asked to optimize the distribution of drinking water in a region with towns (which require drinking water), reservoirs and pumps (which pump drinking water from one part of the region to another part).

The specific setting we studied, the Grimsby drinking water supply region in Canada, is shown in Figure 6.1. It consists of three towns Smithville, Beamsville and Grimsby. The drinking water demand of each town has a typical pattern that is more or less known in advance. Typical demand curves are given in Figure 6.2 where the demand is known per 15 minute sections per day.

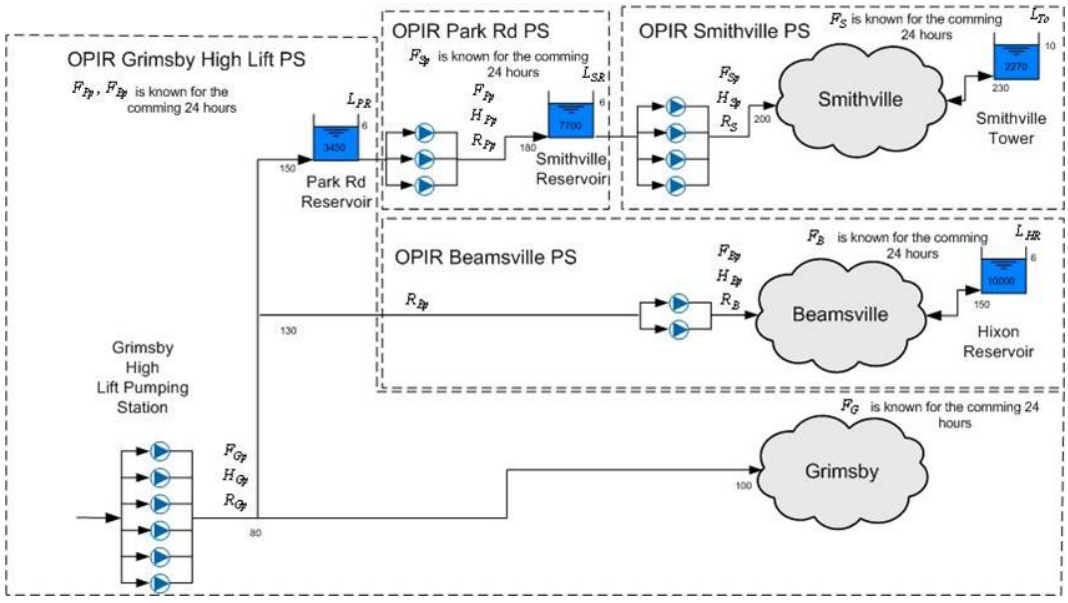

Figure 6.1: A sketch of the Grimsby drinking water supply region with the three towns Smithville, Beamsville and Grimsby, and the pump stations and the reservoirs.

Drinking water is pumped into this supply region through the Grimsby High lift pumping station, located at a certain height H_{Gp} above sea level. With the use of water reservoirs and pumping stations, the drinking water is stored in the region and distributed over the three towns.

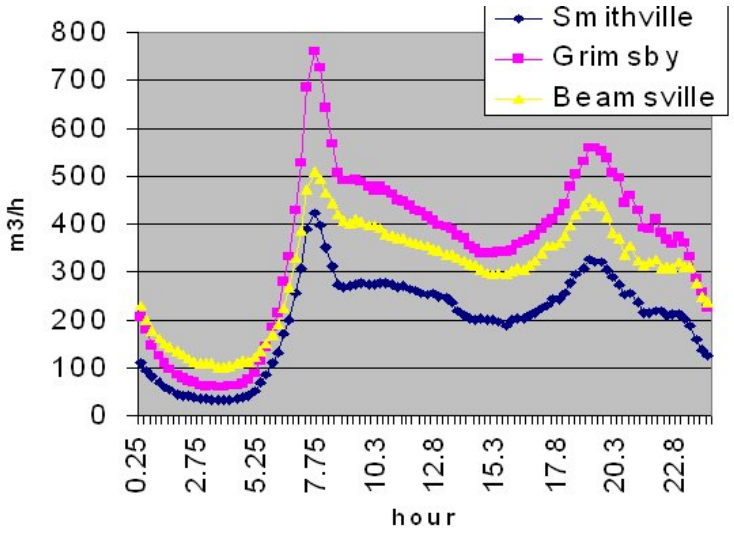

Figure 6.2: The drinking water demand for each of the three towns.

The pumping stations at Smithville, Park road and Beamsville are located at different heights, H_{Sp}, H_{Pp} and H_{Bp}, respectively. The pumping stations each contain a certain number of pumps, see Figure 6.1, which have different capacities and which can either be turned on or off every 15 minutes. Also, the pressure difference *before* and *behind* the pumps, the so-called head loss, determines the operation of the pump. When switching on a pump, this head loss first has to be overcome before the water starts flowing through the pump.

The reservoirs have different capacities as denoted in Figure 6.1. Moreover, there are restrictions on the minimum level, 75 %, and maximum level, 95 %, that the reservoirs are allowed to contain.

Finally, operating the pumps costs energy which in turn costs money. The cost of energy is known; it varies through the day and is different on weekends, see Figure 6.3. Under the given restrictions and water demand, we were asked for an optimal solution such that the cost of energy is minimal. In other words, DHV would like to know, for a period of 48 hours in advance, at which moment the pumps should be turned on or off. Each pump can be switched on or off every 15 minutes. Since there are 15 pumps in the system, there are within the time-frame of 48 hours, $2^{(4*48*15)}$ switching possibilities. Obviously, studying this complete system with all of these possibilities is not possible, so other approaches need to be taken.

One of the pumping strategies, for example, is to fill the reservoirs during the night

Table 1: Energy prices

Day of the Week	Time	Time-of-Use Period	Time-of-Use Price* (cents/kWh)
Weekends & holidays	All day	Off-peak	3.0
Summer Weekdays (May 1st - Oct 31st)	7:00 a.m. to 11:00 a.m.	Mid-peak	7.0
	11:00 a.m. to 5:00 p.m.	On-peak	8.7
	5:00 p.m. to 10:00 p.m.	Mid-peak	7.0
	10:00 p.m. to 7:00 a.m.	Off-peak	3.0
Winter Weekdays (Nov 1st - Apr 30th)	7:00 a.m. to 11:00 a.m.	On-peak	8.7
	11:00 a.m. to 5:00 p.m.	Mid-peak	7.0
	5:00 p.m. to 8:00 p.m.	On-peak	8.7
	8:00 p.m. to 10:00 p.m.	Mid-peak	7.0
	10:00 p.m. to 7:00 a.m.	Off-peak	3.0

Figure 6.3: The energy prices.

when the energy is cheapest. In this way, there is sufficient water supply to satisfy the peak in the drinking water demand of the towns in the morning. In this way the pumps are operated less during the times when the energy is most expensive. However, it is not at all clear that this is the optimal solution since the energy price varies through the day and several other constraints need to be satisfied. Also, filling the reservoirs up to the maximum level could result in a surplus of water stored in the reservoirs. Moreover, how to operate the pumps such that this strategy is fulfilled, is also not known.

We use several analytic and computer aided approaches to tackle the problem. First, we use the fact that the Grimsby drinking water supply region can be split up into several independent modules (OPIRS in Figure 6.1). In section 2, we study one such a module analytically. For this module, a set of algebraic-differential equations is derived which is then optimized by using a Lagrange multiplier. In section 3, we analyze optimal pump rates for the pump stations. Then, in section 4, we develop a method that, given a certain flow rate going to a pump station, determines the combination of which pumps should be switched on and which ones off to give this flow rate. Finally, using this result, a feedback controller is proposed in section 5.

6.2 Analytic approach to flow control

6.2.1 Modular approach to water pump optimization

In a network of water pumps, reservoirs, and supply regions, it is possible to identify a general module. The entire network can then be interpreted as a network of such modules.

The general module is displayed in Figure 6.4.

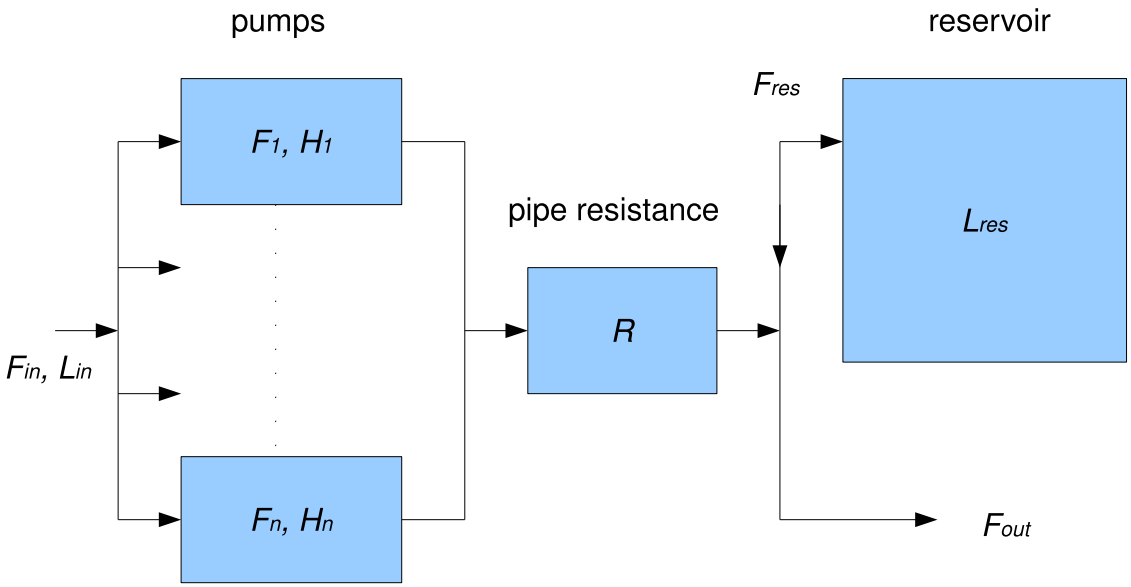

Figure 6.4: A module consisting of a set of n water pumps, a pipe with resistance R and a reservoir.

The module consists of n water pumps that can be switched either on or off, with no intermediate states. The flow through pump i is given by $F_i \geq 0$ in m^3/h. The sum of these flows is necessarily equal to F_{in}. Behind each pump, the water flows through a pipe which has a characteristic resistance R (h^2/m^5). The flow is then split into the flow demand by the supply region (or to another module), F_{out} (≥ 0), and a flow F_{res} into the reservoir. The reservoir flow F_{res} can be negative, representing a flow from the reservoir to F_{out}. We have the following continuity equation

$$F_{\text{in}} = \sum_{i=1}^{n} F_i = F_{\text{out}} + F_{\text{res}}. \tag{6.1}$$

The water height above sea level in the reservoir is denoted by L_{res} in m. It can thus be compared to the water level just before the pumps, L_{in}, also absolute above sea level. The

water level satisfies the differential equation

$$A\frac{dL_{\text{res}}}{dt} = F_{\text{res}}, \tag{6.2}$$

where A is the surface area of the reservoir. The head H in m (a measure for pressure) required to transfer the water from L_{in} to L_{res} is given by

$$H = L_{\text{res}} - L_{\text{in}} + R\Big(\sum_{i=1}^{n} F_i\Big)^2 = L_{\text{res}} - L_{\text{in}} + RF_{\text{in}}^2. \tag{6.3}$$

For all the pumps that are switched on, a nonlinear relation holds between the head over the pump and flow F_i through the pump. This relation is given by

$$H = -\alpha_i F_i^2 + \gamma_i, \quad i = 1, \ldots, n. \tag{6.4}$$

Here (α_i, γ_i) are positive constants that characterize pump i for $i = 1, \ldots, n$. The pressure over all pumps is equal, which explains why H is independent of i. If a pump i is switched off it gives rise to the flow $F_i = 0$.

The requested flow F_{out} and the water level L_{in} are assumed to be given functions of time. If we decide which pumps are turned on, equations (6.1), (6.3) and (6.4) give $m+2$ relations in the $m+2$ unknowns $F_{\text{res}}, (F_i), H$, where m is the number of pumps that are switched on. Roughly speaking, by the implicit function theorem this determines F_{res} locally as a function of $L_{\text{in}}, L_{\text{res}}$ and F_{out}. The dynamics of L_{res} are then described by (6.2), and from L_{res} all the other dynamics follow.

6.2.2 Analytic approach to flow control

The energy price is given by $c(t)$ as a function of time (in ct/Kwh). We have $P = P(F_{\text{in}}, H)$ which expresses the power consumed by the pumps in kW. An approximation for P is

$$P(F_{\text{in}}, H) = kF_{\text{in}}H, \tag{6.5}$$

where $k > 0$ is a constant that relates to the efficiency of the pumps. In practice k depends on F_{in}. Altogether, this enables us to express the pumping power in monetary units as a function of time. Our goal is to minimize the total monetary costs over a time span T. To make analysis possible, we make the following simplifying assumptions.

Assumption 6.1. We can obtain any flow F_{in} by switching pumps on or off.

Assumption 6.2. There are no constraints on allowed reservoir levels.

There are some objections. Since pumps are only switched on/off at discrete times (e.g. at most every 15 minutes), and since pumps have their limits, Assumption 6.1 does not hold in practice. Furthermore, water levels should be kept in a [75 %, 95%] range of the reservoir capacity, so Assumption 6.2 can not hold in practice. We discuss these objections later.

For any given F_{in}, we can in principle compute the resulting head through either equation (6.3) or, in case we switch only one pump on, (6.4). It makes more sense to use (6.3) since the required head is expressed through this equation, which is of crucial importance. Equation (6.4) actually looses its meaning when we use assumption 6.1: we assume that the given head/flow combination can be delivered by a certain combination of pumps and are thus indifferent about the specific pump characteristics.

Next we derive an optimization problem to minimize the required energy. We treat F_{in} as the control variable and use the notation $u = F_{\text{in}}$, $x = L_{\text{res}}$.

The total energy used over time span T is given by

$$E[x,u] = \int_{t_0}^{t_0+T} c(t)P(F_{\text{in}},H)\ dt = \int_{t_0}^{t_0+T} c(t)ku(t)\Big(x(t) - L_{\text{in}}(t) + Ru(t)^2\Big)\ dt, \quad (6.6)$$

where we used (6.3) to express H as a function of (u,x), and (6.5) to calculate the power used by the pumps. Equation (6.2) translates into the constraint

$$\dot{x}(t) = \Big(u(t) - F_{\text{out}}(t)\Big)/A, \quad t \in [t_0, t_0+T]. \quad (6.7)$$

We require that, after time T, the reservoir level x is equal to its starting value at t_0, that is

$$x(t_0) = x(t_0+T). \quad (6.8)$$

Using Lagrangian multiplier $\lambda(\cdot)$ to include the constraint (6.7) we obtain the following Lagrangian,

$$\begin{aligned} L(x,u,\lambda,\mu) &= E(x,u) + \int_{t_0}^{t_0+T} \lambda(t)\Big((u(t) - F_{\text{out}}(t))/A - \dot{x}(t)\Big)\ dt + \mu(x(t_0+T) - x(t_0)) \\ &= \int_{t_0}^{t_0+T} c(t)ku(t)\Big(x(t) - L_{\text{in}}(t) + Ru(t)^2\Big) + \lambda(t)\Big((u(t) - F_{\text{out}}(t))/A - \dot{x}(t)\Big)\ dt \\ &\qquad + \mu(x(t_0+T) - x(t_0)) \\ &= \int_{t_0}^{t_0+T} c(t)ku(t)\Big(x(t) - L_{\text{in}}(t) + Ru(t)^2\Big) + \lambda(t)\Big(u(t) - F_{\text{out}}(t)\Big)/A \\ &\quad +\dot{\lambda}(t)x(t)\ dt \qquad + (\mu - \lambda(t_0+T))x(t_0+T) - (\mu - \lambda(t_0))x(t_0), \end{aligned}$$

where we used partial integration in the last step. At an extremum, small variations of x (with fixed boundary values $x(t_0) = x(t_0 + T)$), u, λ and μ should have no influence on the value of the Lagrangian. By formally differentiating with respect to x we obtain the condition

$$c(t)ku(t) + \dot{\lambda}(t) = 0, \quad t \in [t_0, t_0 + T]. \tag{6.9}$$

Differentiating with respect to u gives

$$c(t)k\Big(x(t) - L_{\text{in}}(t) + 3Ru(t)^2\Big) + \lambda(t) = 0, \quad t \in [t_0, t_0 + T]. \tag{6.10}$$

From (6.10) we derive that, for $t \in [t_0, t_0 + T]$,

$$u(t) = \begin{cases} \sqrt{\frac{1}{3R}\Big(L_{\text{in}}(t) - x(t) - \frac{\lambda(t)}{c(t)k}\Big)}, & \text{if } c(t)k(L_{\text{in}}(t) - x(t)) - \lambda(t) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6.11}$$

Recall the differential equation (6.7) for x. We have now obtained the coupled set of differential equations

$$\begin{aligned} \dot{\lambda}(t) &= -c(t)ku(t), \\ \dot{x}(t) &= (u(t) - F_{\text{out}})/A, \qquad\qquad t \in [t_0, t_0 + T], \end{aligned}$$

subject to boundary condition (6.8).

As an example we solved these equations numerically for the Smithville reservoir, starting from $t_0 = 0$ over a time period $T = 24$ h and assuming a fixed water level L_{in}. The solutions are depicted in Figure 6.5.

We see that there is no flow when energy is most expensive. Unfortunately the reservoir limits $[230, 240]$ are significantly exceeded. So Assumption 6.2 is strong. Otherwise the results seem realistic, which gives confidence in our methods. It is interesting that the flow rate is not periodic. Furthermore it is of interest why the function of flow rate with respect to time has (at certain time intervals) the form of the square root function.

The violation of assumption 2, required for a global optimum, implies that the water level constraint reduces optimality of the controlled system. This conclusion can help for future design considerations.

In future research, reservoir limits can be included in the optimization either by using slack variables or by using a penalty function. The same can be done to include maximum flow rates, in order to relax assumption 6.1. Another interesting topic of further research is to optimize a coupled set of modules, so that an entire network of pumps, pipes and reservoirs can be controlled optimally.

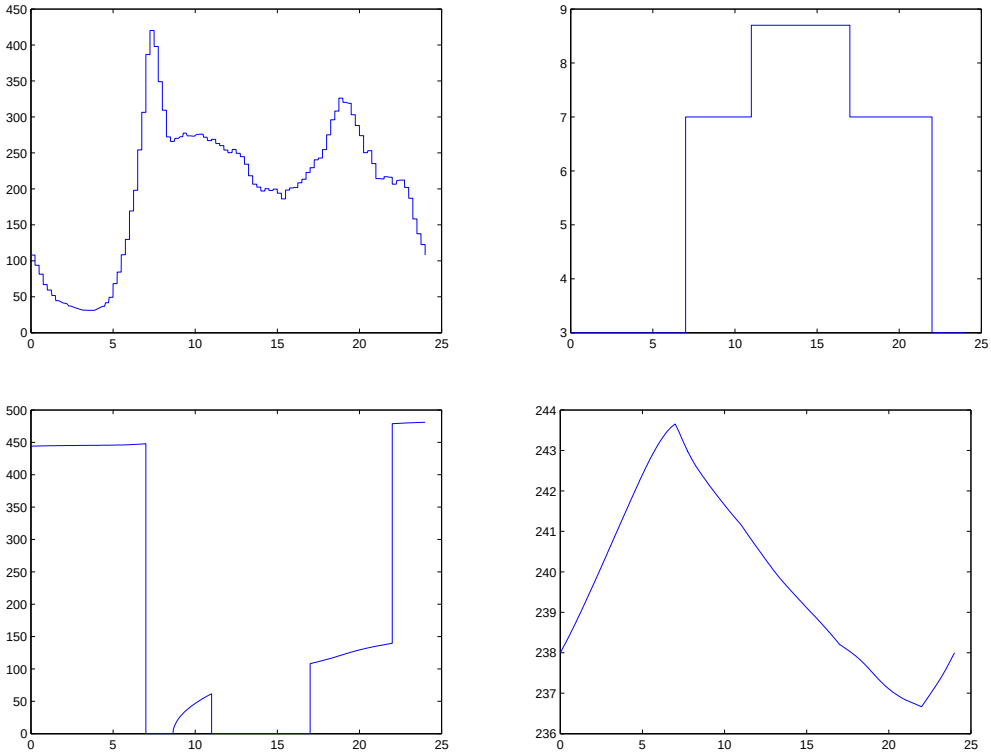

Figure 6.5: Upper left: Water demand F_{out} in Smithville. Upper right: Energy costs $c(t)$ per kWh. Bottom left: Optimal water flow $u = F_{\text{in}}$ through pumps. Bottom right: Reservoir level $x = L_{\text{res}}$ in case of optimal flow.

6.3 Optimal Pump Rates for Four Stations

6.3.1 A simple model

In this section we calculate optimal pump rates for the four pump stations at Park Road, Grimsby High, Smithville, and Beamsville. First, a model is defined for the dynamic behavior of the water levels in the four corresponding reservoirs. This model can easily be obtained by performing a mass-balance equation for each reservoir. Denote with $x_i(t)$ the

water level in each of the reservoirs so that the following 4-state model can be defined:

$$\begin{aligned}
\dot{x}_1(t) &= \frac{u_1(t) - d_{Gr}(t) - u_2(t) - u_4(t)}{A_{PRdRes}} \\
\dot{x}_2(t) &= \frac{u_2(t) - u_3(t)}{A_{SmRes}} \\
\dot{x}_3(t) &= \frac{u_3(t) - d_{SmV}(t)}{A_{SmTow}} \\
\dot{x}_4(t) &= \frac{u_4(t) - d_{BV}(t)}{A_{HxRes}}
\end{aligned} \tag{6.12}$$

where $d_{Gr}(t),\ d_{SmV}(t),\ d_{BV}(t)$ are the demand curves for Grimsby, Smithville, and Beamsville, respectively (these are assumed known and given) $[m^3/hr]$; A_{PRdRes}, A_{SmRes},A_{HxRes},A_{SmTow} are the local surface areas of the three reservoirs and Smithville Tower $[m^2]$; $u_1(t),\ u_2(t),\ u_3(t),\ u_4(t)$ are the pump rates at Grimsby High Pumping station, Park Road, Smitville Pumping Station, and Beamsville pumping station, respectively $[m^3/hr]$. Finally, the states $x_1(t),\ x_2(t),\ x_3(t),\ x_4(t)$ are the water levels in the three reservoirs and at Smithville tower.

An important note on the required pump rates (to be solved for) is the assumption of a continuous variable $u_i(t)$ for all four pumping stations. This is a simplification which could be relaxed at a later stage of the project, but for now it is very convenient to allow a continuous variable since the optimal control algorithm at our disposal can directly be applied. A realization of the optimal pump rates is deferred to section 6.4. In addition to the above dynamic constraints we also introduce four state constraints on the water levels. After some discussion with the problem owner we decided to maintain the water levels in a bandwidth of 25% to 95% of the maximum levels allowed. With regard to the pump rates it should be noted that we assume only positive values of the inputs $u_i(t)$ for a simple reason: the pump rates are not allowed to pump in reverse direction.

6.3.2 The goal function

The provided problem description included a clear goal, namely to minimize on the electricity price for operation of the four pumping stations. The following table of electricity prices was included:

Electricity Price [ct/kWh]	Time Slot
3.0	22:00 – 7:00 hr (next day)
7.0	7:00 – 11:00 and 17:00 – 22:00 hr
8.7	11:00 – 17:00 hr

The trade-off that needs to be optimized in this case is storage of water in the reservoirs that can be stocked at cheap electricity time-slots, whilst not increasing the water levels too much since additional head builds up when the reservoirs are filled with water and this hampers the pumps in their task (thereby reducing the flow rates). For each pump station the consumed power is

$$P(t) = C\,H(t)\,u(t)$$

where $P(t)$ is the power [kW], C is a constant characterizing the pump efficiency, and $u(t)$ is the flow rate $[m^3/hr]$. The hydrologic head as experienced by a pump is given by

$$H(t) = \Delta L + x(t) + R\,u^2(t)$$

where ΔL is the elevation difference between two pump stations, $x(t)$ is the water level in the reservoir, and R is the hydrologic resistance of the piping network. Let $p_E(t)$ denote the pricing of electricity [ct/kW]. Then our problem is to minimize the total monetary costs over 24 hours:

$$\int_0^{24} p_E(\tau)\,P(\tau)d\tau$$

6.3.3 Results

To obtain some first results the above problem was programmed in Matlab, making use of the so-called TomLab PROPT toolbox for optimal control. The software allowed all constraints (both input and state constraints) to be included. To force a cyclic solution, and not to obtain so-called 'greedy control', terminal constraints were included so that the final water levels in the reservoirs are exactly the same as the initial water levels. In Figure 6.6 the optimization results are presented in three graphs. In the first graph we see the water-levels in the reservoirs. It is immediately clear that Smith Tower with a relatively small capacity is used as a storage during off-peak hours and this clearly pays off in terms of electricity use.

Grimsby High lifting station has the highest pump rates which can be expected since it has such a central position as a gateway to the three communities. It is clear from the results that our calculated strategy anticipates on low electricity prices by pumping intensively during the off-peak hours. Also, the pumps do not switch off completely in the most expensive hours, indicating that hydrologic head buildup is circumvented.

Of course, the above results are just a starting point that should be elaborated upon at a later stage. More refinement in, for example, the hydrologic resistance R for the piping

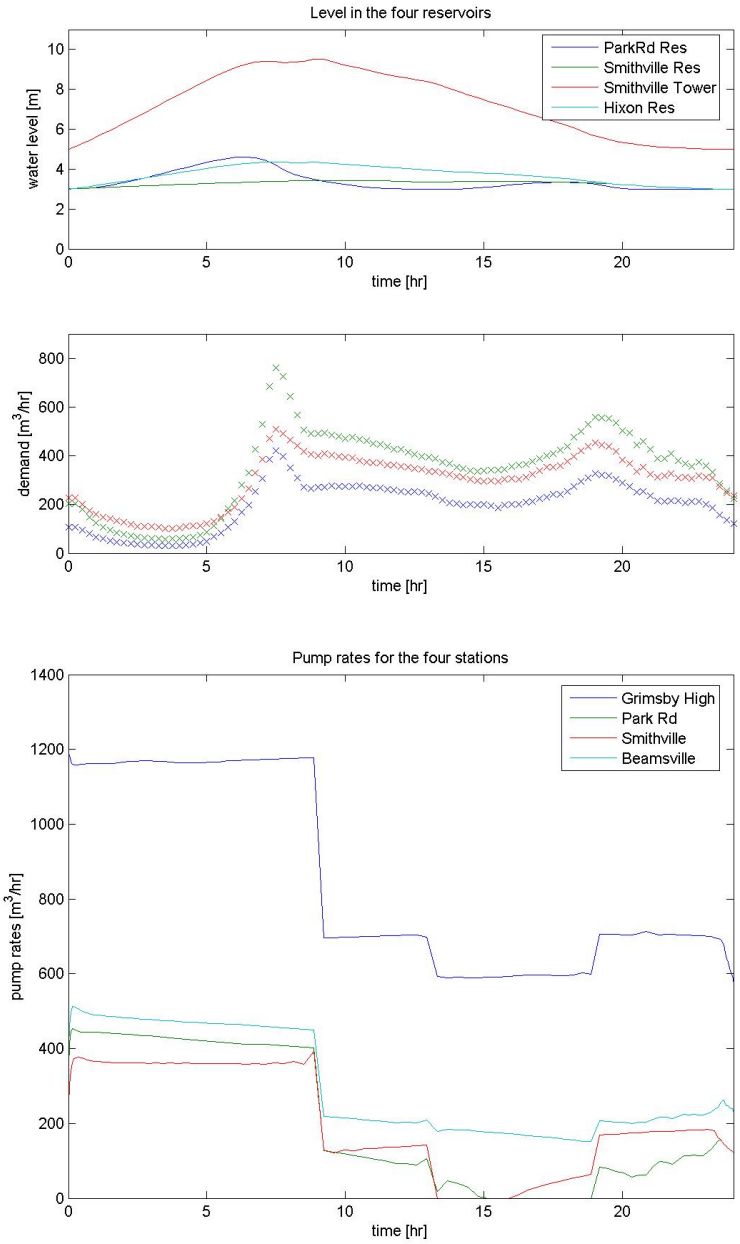

Figure 6.6: Optimization results for pump-stations

networks could be taken into account and, also, the on-off switching nature of the controls.

6.4 Conversion of continuous flow rates into pumping combinations

6.4.1 Introduction

In the previous sections, continuous-time methods are used to control drinking-water supply systems. These methods assume that the flow rates for each pumping station is a continuous control input that can be controlled directly. However, the on- and off switching of the pumps make it a discrete quantity. In this section, a method is developed that computes for any given continuous-time flow rate a combination of switched on pumps, that results in a flow rate that is most similar to the one that was given. Throughout this section, we assume that there are no transient effects, i.e. when a pump is switched on or off, the resulting flow is immediately in steady state.

6.4.2 Modular flow model for given pump states

First, a single pump is considered. The pumping pressure can be represented by a quantity in meter. The head H is

$$H = \frac{p}{\rho g}, \tag{6.13}$$

with p the pressure , ρ the density and g the acceleration of gravity. The sum of the head and the physical height difference (generalized head) determines the flow rate. This relation can be inverted: if the flow rate is given, then the generalized head can be calculated via a Bernoulli equation.

A pump P is considered as an object with two member functions: P.head2flow(H) calculates the flow rate for a given generalized head H, and P.flow2head(F) calculates the generalized head for a given flow rate. The same analysis applies to pipes. Pumps and pipes are examples of a network. Each network object N has member functions N.head2flow(H) and N.flow2head(F).

Networks can be build recursively from parallel connections and serial connections, and we treat them separately.

- For a parallel connection N with subnetworks $S[1], \ldots, S[n]$, the flow resulting from a given head is calculated by adding the flows through the subnetworks:

$$N.\text{head2flow}(H) = \sum_{i=1}^{n} S[i].\text{head2flow}(H) \tag{6.14}$$

	α	γ	max. head	max. flow
pump 1	$-1/2$	2	2	2
pump 2	$-1/3$	3	3	3

Table 6.1: Characteristics of pumps in calculation example

The function $H = N.\text{flow2head}(F)$ is now evaluated by iteratively searching H such that $N.\text{head2flow}(H) = F$. For this, we used an algorithm that solves one nonlinear equation with one unknown.

- For a serial connection N with subnetworks $S[1], \ldots, S[n]$, the head over network N is calculated by adding the heads over the subnetworks:

$$N.\text{flow2head}(F) = \sum_{i=1}^{n} S[i].\text{flow2head}(F) \tag{6.15}$$

The function $F = N.\text{head2flow}(H)$ is evaluated by iteratively searching a flow F such that $N.\text{flow2head}(F) = H$. Note that the generalized head over a network N between two reservoirs is known, because it equals the height difference ΔL (m) between the reservoir levels.

Now, for any given pump state, the resulting flow rates and the pumping pressures can be found by evaluating $N.\text{head2flow}(\Delta L)$. This gives a table with all possible flow realizations and their corresponding pump states.

6.4.3 Model application

As an example, we investigate a pumping station with two unequal pumps. Assume that the pumps both satisfy $\text{flow2head}(F) = -\alpha F^2 + \gamma$, but with different characteristics α and γ, listed in Table 6.1. These pumps are connected in parallel, and the pumping station is connected in series with a pipe satisfying $\text{flow2head}(F) = -0.5F^2$. Assume that this network is connected with two reservoirs having a water level difference of 1. The flow rates that can be realized are shown in Table 6.2. There are two pumps that can be either on or off, resulting in four possibilities. It is interesting that there is no solution if both pumps are turned on. The reason is that the largest pump would produce a larger head than the maximum allowed for the smallest pump (Table 6.1).

	Pump states		head of	head of
flow	pump 1	pump 2	pumps	pipe
0	off	off	0	0
1	on	off	1.5	-0.5
1.5	off	on	2.2	-1.2
–	on	on	–	–

Table 6.2: Possible flows in calculation example; if both pumps are on, there is no solution for the head and the flow rate

The drinking water supply system for Smithville, Beamsville and Grimsby is shown in Figure 6.7. To apply our algorithm, the system is divided in five subsystems. If the dynamics of the reservoir levels is considered small compared to the static height differences between the reservoirs, then subsystem 4 and subsystem 5 in Figure 6.7 are driven by a fixed height difference. These systems are independent on the subsystems 1 – 3 (provided that shared reservoirs are not empty). Subsystems 1 – 3 are not independent, because they depend on the generalized head in connection point C. To solve this, we make use of an extra equation that expresses mass conservation

$$F_1 - F_2 - F_3 = 0. \tag{6.16}$$

6.4.4 Discussion

By applying recursion, the multivariate problem can be solved using a numerical method that solves scalar equations. The recursive algorithm applies to a general class of drinking water supply systems, including the Smithville, Beamsville and Grimsby situation. The algorithm was applied to a calculation example with unequal pumps. In this example, one pump state combination was impossible, because it resulted in a larger head than possible for the one pump. This boundary indicates that the algorithm should be used with caution.

The result is a table that contains all possible steady states for the flow rates in the system. This table can be used to approximate a continuous-time control input (flow rate) by a discrete one (pump combination).

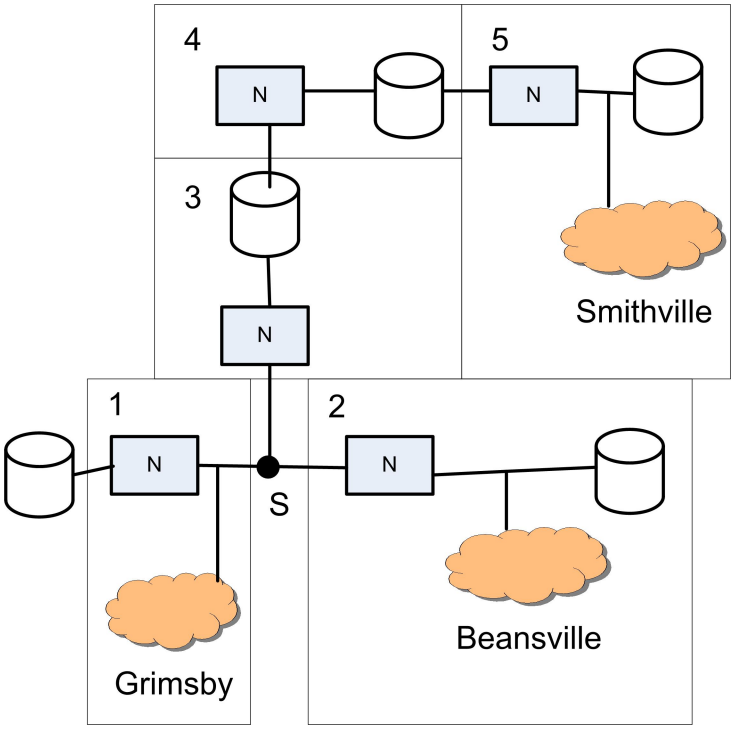

Figure 6.7: Network representation of supply to Smithville, Beamsville and Grimsby, with subsystems 1 – 5

6.5 Local linear feedback control

6.5.1 Problem definition

The starting point of this section is the optimal state and input trajectories that were derived in 6.3. An optimal control method is called open loop, which means that it computes the trajectories beforehand. This provides an ideal choice of input for the undisturbed system, but in general the robustness with respect to model errors and disturbances, such as deviations from the predicted water demand, is hard to guarantee. For example, although the optimal input trajectories are refreshed each 15 minutes, which can be seen as state feedback control, there is no 'integrating action'. This means that the for a model error or disturbance that generates an output error that is constant in time, the controller keeps responding in the same way and thus keeps making the same error. More concrete, if the optimal controller keeps predicting an input that results in a too low water level in the basins, there is no mechanism that adjusts the input for that constant output error.

6.5.2 Approach

A feedback controller is proposed to make the optimally controlled system more robust against the above mentioned perturbations. The purpose of this controller is to drive the system output (water height in the basins) to the output trajectory that was predicted by the optimal controller by using a feedback mechanism that adjusts the input (pumping power). There are many ways to design a feedback controller, and they come from different fields. We choose model based linear feedback control, since this is a widely successfully applied approach that has a strong mathematical foundation. Further, it is a textbook subject and design tools are widely available.
The approach is the following. First, since linear control design is mathematically only possible for linear systems, the system is linearized around the optimal state and input trajectories. The new variables are the deviations from the optimal variables, and they are defined as

$$\begin{aligned} \tilde{x}(t) &= x(t) - x_{opt}(t) \\ \tilde{u}(t) &= u(t) - u_{opt}(t) \\ \tilde{y}(t) &= y(t) - y_{opt}(t), \end{aligned} \tag{6.17}$$

where x is the water height in the basins, y the measured output (for example the water height in some of the basins), and u the pump rates. The subscript "opt" refers to the optimal trajectories derived in section 6.3. Inserting (6.17) in the model equations (6.12) gives a system of the form

$$\begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t) + Ee(t) \\ y(t) &= Cx(t), \end{aligned} \tag{6.18}$$

with e the disturbances in the water demand curves $d(t)$, and A, B, C and E system matrices. The system (6.18) is in a form that allows the design of a feedback controller, e.g. via H_∞ theory.

6.5.3 Discussion

It is shown that given an optimal input and state trajectory, a robust linear feedback control design is possible. The design itself is omitted, but as mentioned before this is a textbook subject. The controller acts locally in time, as opposed to the optimal controller, so it does not look ahead to save energy costs. For example, it does not shift pumping duties to the night time because the power is then cheaper. In theory, each time that

the optimal trajectory is refreshed, the trajectories that the state and input are linearized around changes, resulting in a different controller each 15 minutes. So A, B, C and E change. However, since the controller will be robust against model errors, this is not necessary as long as the differences stay reasonably small.

6.6 Conclusions

Given the fluctuating energy prices and the drinking water consumption in the Grimsby area, where each of the 15 pumps are either switch on or off, in total $2^{4x48x15}$ possible trajectories result. Hence, a solution via enumeration is infeasible and thus there is a need for approximations.

At first, we consider smooth (continuous-time) pump functions and one head-flow relationship per pumping station. Hence, a set of algebraic-differential equations with constraints result. This allows us to use Lagragian theory for dynamic systems, also known as the minimum principle of Pontryagin. After all, a two-point boundary value problem (TPBVP) in terms of the states and co-states results. Recall that in this problem, the flow generated by a pumping station is the control input and the height in a reservoir is the state of the system. Given the energy-related goal function, together with the input and state constraints, numerical solutions to the TPBVP have been found. If, however, we consider the unconstrained problem for a single pumping station configuration, (semi-)analytical solutions result.

In a second step and given a required head-flow combination as a function of time, as found after solving the TPBVP, an optimal pump configuration can be selected. Careful analysis of the Grimsby region, under quasi-steady state assumptions, also shows four archetypical modeling problems, which can be solved in a modular approach.

The direct link between energy costs and the flow signals allow a direct physical interpretation. The unconstrained problem for Beamsville, while considering only the running costs, clearly shows that enlarging the Hixon reservoir and pumping capacity is profitable.

Further research is needed to analyze the problem with respect to the sub-optimal solutions found in this work, to set-up a generic framework for the dynamic optimization of any drinking water network and to come up with real-time solutions.