DAILY TEMPERATURE OPTIMISATION IN GREENHOUSE BY REINFORCEMENT LEARNING

Marc Tchamitchian *,[1] **Constantin Kittas** **
Thomas Bartzanas ** **Christos Lykas** **

* *Unité PSH, Bât B - INRA Domaine St Paul*
84914 Avignon CX 9 - France
** *Lab. of Agric. Construct. & Environ. Control*
School of Agric. Sciences - Univ. Thessaly
Fytokou Street - 38 446 Nea Ionia Magnisias - Greece

Abstract: The goal of this study is to show the usefulness of reinforcement learning (RL) to solve a common greenhouse climate optimisation problem. The problem is to minimise the daily heating cost while achieving simultaneously two agronomic goals, namely maintaining a good crop growth and an appropriate development rate. The complexity of the problem is due to the very different time constants of these two biological processes. First, a simple model for greenhouse roses is presented that simulates the daily crop growth and development. Second, the RL method is presented, in its application to this problem. Finally, optimisation results are presented and discussed. *Copyright© 2005 IFAC*

Keywords: Greenhouse, Climate control, Rose, Temperature, Heating, Reinforcement learning

1. INTRODUCTION

Greenhouse crops can be manipulated by the control of their environment, including climate, fertigation and biotic populations. Heating is, after labour, the main cost of greenhouse crops, but it is also one of the main controls used to steer the crop behaviour according to the grower's objectives. The relation between heating and crop response is not straightforward, nor simple. The heating mainly modifies the air temperature, but also affects the air water vapour deficit (VPD) but the resulting greenhouse climate is also dependent on the outside climate and the crop activity (transpiration mainly). The response of the crop to heating is in reality a response to the greenhouse air temperature and VPD and, depending on the heating system, to the heating intensity itself. Determining the adequate temperature according to a goal on the crop behaviour is therefore a complex task, which is mainly achieved by the grower.

In practice, two main temperature set-points are used, one for the night, one for the day. Although only two set-points are used, the greenhouse climate does not remain stable, especially during daytime, due to the evolution of outside conditions. Furthermore, it has been shown that crops can withstand deviations from optimal conditions for a while and recover later, provided that these episodes do not last more than a few days (Bredmose and Nielsen, 2004; Bakker and

[1] This research has been carried out during the sabbatical of Dr Tchamitchian at the Univ. of Thessaly, funded by the ENTER program of the General Secretariat for Research and Technology of the Greek Ministry of Development

van Uffelen, 1988; Heuvelink, 1989). It is therefore possible to optimise the daily evolution of the temperature in the greenhouse so as to maintain a given average and also to limit the energetic cost of this average.

The temperature optimisation problem has received a lot of attention in the past years. Within the studies devoted to the daily optimisation, most have concentrated on optimising the temperature to maximise growth rate or a function of the crop harvest (which may include economic optimisation by balancing harvest and costs), which in fact is very close to maximising the growth of the crop for vegetable crops. Slow crop processes, like development rates, are rather difficult to take into account because of the limited time horizon of such optimisation. Moreover, dynamic models of the crop including two very different time scales are not easy to solve with the mathematical methods of Automatics like optimal control. Therefore development is very seldom included in the optimisation criteria for daily temperature optimisation.

The goal of this work is to overcome the above mentioned limitations in solving the daily greenhouse temperature optimisation problem, stated as an energy saving problem. It implies to redesign the criterion used to define the optimal solution and therefore the model of the bio-physical system. The role of the model of the bio-physical system is both to predict the response of the system to the controls applied to it and to provide the necessary information to build the criterion. New optimisation methods based on the results of the simulations themselves rather than on the formal structure of the model will be applied. These methods are computer intensive but have fewer requirements on the model structure. Their usefulness and efficiency will be demonstrated here on an application to the control the greenhouse climate for rose production.

This paper is organised as follows: first, a description of the reinforcement learning method (RL) is given because it has implications on the structure of the model. Second the greenhouse rose crop model is described, as well as the chosen criterion. Third, optimisation results are given and discussed. Finally, the application of RL to the greenhouse climate control problem is discussed.

2. REINFORCEMENT LEARNING

Reinforcement learning is defined as a type of problem to solve rather than by an algorithm (Sutton and Barto, 1998). It can be defined as solving the problem of deciding what to do in a given situation by trial and error, that is by interacting with the environment. Most commonly, it is applied to Markov Decision Problems (MDP). An MDP is an extension of Markov chains including actions and rewards (Puterman, 1994). In an MDP, the system can be in a given state of the set of possible *states*, $\mathcal{S}$, and will move to another state in $\mathcal{S}$ with a given probability because an action was taken in the set of possible actions, $\mathcal{A}$. With each state, a reward is associated that measures how satisfactory is the state with respect to the goal assigned to the control problem. In an MDP, the transition probability from state s_i to state s_{i+1} given the action a_i only depends on s_i and a_i. What previous actions were taken and which previous states the system was in are of no importance, only the current state and taken action are necessary to move forward. The second property of MDPs is that the criterion to maximise is the sum of the local rewards obtained after each decision step. Given this system, the problem is to find the sequence of actions that maximises the criterion. It corresponds to mapping from $\mathcal{S}$ into $\mathcal{A}$. This sequence of actions is called a policy, noted Π. Because the system evolves in a stochastic environment, the state it will be in at a given time step is not known. Therefore a policy is constituted by a set of subpolicies defined for each time step and mapping the each possible states of this time step into an action. It is very close to the feedback loops of classical control where the action is decided based on the current system state using the feedback law.

Reinforcement learning consists in exploring by simulation the outcome of policies generated either randomly or according to some law, in order to *learn* the transition probabilities from a given state into another granted a given action, transitions which are not known *a priori*. The R-learning algorithm learns the estimates of the average value ρ and of the relative value functions $R_i(s,a)$. ρ is defined as:

$$\rho = E\left(\frac{1}{N}\sum_{i=1}^{N} r_i\right) \qquad (1)$$

where N is the number of decision steps (time steps)and r_i is the local reward at step i. $R_i(s,d)$ is:

$$R_i(s,a) = E\left(\sum_{j=0}^{N} r_j - \rho | s_i = s, a_i = a\right) \qquad (2)$$

where s is a state (s_i at step i) and a an action. Given this knowledge, the optimal policy is found by choosing, at each decision step, the action that maximises the R_i function:

$$\forall i, \forall s \in \mathcal{S}i \; \Pi i(s) = \underset{a \in \mathcal{A}i}{\arg\max}\, R_i(s,a) \qquad (3)$$

In finite horizon and discrete state and decision space, the R-learning method used in this study is summarised in algorithm 1.

```
$\rho \leftarrow 0$;
$\forall i \in \{1,\ldots,n_a+1\}, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}\ \ R_i(s,a) \leftarrow 0$;
for $n \leftarrow 1$ to $n_{max}$ do
    Set initial state of the system;
    Run the model using either a random policy or a policy determined according to eqn. 3 (greedy policy);
    for $i \leftarrow n_a$ to $1$ do
        $R_i(s_i,a_i) \leftarrow R_i(s_i,a_i) + \alpha_n \left(r_i - \rho + \max_{a'} R_{i+1}(s_{i+1},a') - R_i(s_i,a_i)\right)$;
        if policy is greedy then
            $\rho \leftarrow \rho + \beta_n \left(r_i - \rho + \max_{a'} R_i(s_{i+1},a') - R_i(s_i,a_i)\right)$;
        end
    end
end
// $n_a$ is the number of decision steps
```

Algorithm 1: R-learning algorithm in the finite horizon case

As will be detailed later, the only reward (r_i) is obtained at the end of the control horizon. Therefore all r_i are null except for r_{final} which is equal to the value of the performance criterion (see below). The average value function for $n_a + 1$ is also always null. Two coefficients appear, α and β. They are learning coefficients and decay in time with the number of visits for each pair (s, a) in the case of α or with the iteration number (β).

Although the present description addresses a discrete representation of $\mathcal{S}$ and of $\mathcal{A}$, it is easy to transpose this approach to continuous domains. However, in the present study, a discrete representation of the system states and of actions has been adopted.

3. GREENHOUSE ROSE CROP MODEL

3.1 Production goal: the optimisation criterion

Greenhouse roses are grown for their flowers, of course. Hence, the main concerns of growers is to obtain flowers of a good quality (long stems, long post harvest life *a.o.*) and at a regular pace or at given dates depending on the obligations contracted with resellers. Hence, dry matter accumulation, although important, is as important than development rates. Because it is rather difficult to unify the measures for growth, for development and for energy consumption (minimising the energy cost of production is our primary goal as mentioned in the introduction), the criterion has been formulated as the sum of these three components, each expressed by comparison to a reference value:

$$Y_\Pi = \frac{1}{\sum_{gde} \alpha} \left[\alpha_g \frac{G_\Pi}{G_{ref}} + \alpha_d \frac{D_\Pi}{D_{ref}} + \alpha_e \left(2 - \frac{E_\Pi}{E_{ref}} \right) \right] \quad (4)$$

where $\alpha_{g,d,e}$ are weighting coefficients allowing to modify the role of each of the three components of Y, G, D, E are measures of the growth, development and energy (respectively) achieved or used during the control horizon and where Π means achieved through the application of the policy Π, while ref denotes the reference policy. The reference policy is the policy that the grower would have chosen on the day at hand, expressed as a set of set points for day and night. To maximise the criterion while minimising the energy used, the third term in Y has been negated and added to 2 so that if the policy Π is the same as the reference policy all the three terms in Y are equal to one. If Y_Π is more than one, then Π is better than the reference policy and can be kept, otherwise Π can be discarded. Thanks to this formulation of the criterion, it is possible to only build three submodels simulating crop growth, development rates and energy consumption instead of a full autonomous model of the crop.

3.2 Greenhouse rose crop model

The control horizon has been fixed to 24 hours and starts at dawn to encompass a daylight period and then a night period, a sequence easier to describe by growth models. This control horizon is discretised in hourly steps where decisions about the heating or ventilation intensities are to be taken.

The dry matter accumulation of a rose crop is modelled rather classically (Dayan *et al.*, 2003), by accumulating photosynthesis and accounting maintenance respiration and, if possible, growth and growth respiration. Parameter values for leaf photosynthesis are taken from Gonzalez-Real and Baille (2000). The structure of the growth submodel is described in details in Seginer *et al.* (1994). It consists of two compartments, one for transient carbon pool (fed by photosynthesis, depleted by respirations and growth) and one for structural dry matter (fed by growth). At the start of each day, the assumption will be made that the transient reserve compartment is empty, an optimal situation for the coming day. Giving a value to the level of the transient pool is rather difficult because is depends on the use that will be made of it, and this use is a function of the future temperatures experienced by the crop. Therefore, the value of the growth is only appreciated at the end of the control horizon and no local reward will be associated to the hourly decisions. Because the criterion Y uses the growth, it is also only

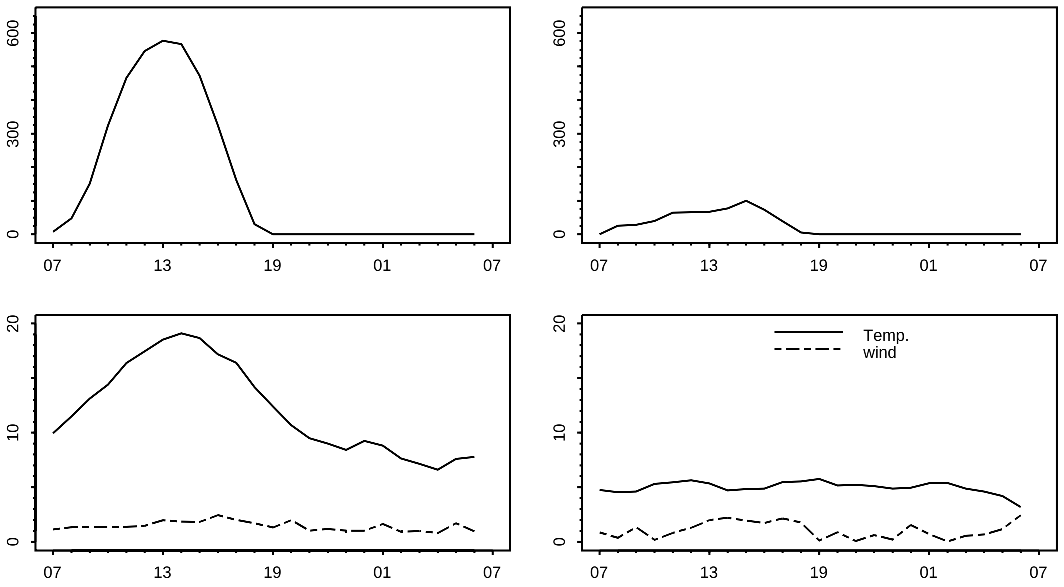

Fig. 1. Outside climate used as weather forecast. Upper row: global radiation; lower row: temperature and wind speed. Left column: March 1st, 2004; right column: March 4th, 2004.

defined at the end of the control horizon? No local rewards will be associated to development and energy consumption too. This submodel must be initialised with the crop LAI and dry matter. To do so, empirical relations mapping the shoot density and leaf number per shoot to the crop LAI have been established using measurements carried out at the experimental farm of the University of Thessaly, Volos, Greece (Lykas, pers. comm.) These measurements were also the base to Leaf Area Ratio function allowing to deduce the crop dry matter from its LAI.

A simple measure of the development rate has been adopted in this study. It consists in estimating the ratio of the degree day provided during the control horizon to the degree day from emergence of a shoot to its harvest. If the hypothesis that the age distribution of flowering shoots in the greenhouse is uniform cannot be made, then the grower has to supply a simple estimation of the age distribution of the flowering shoots, like the ratio of shoots bearing a bud to those not bearing one (younger). Degree day values of the life of flowering shoot, including some intermediate stages were taken from Pasian and Lieth (1996).

To determine the greenhouse temperature from the heating or ventilation intensity decided upon at each time step, a greenhouse energy model has been used. Because of the discretisation of the time adopted in the present study, a simple static model has been selected (Gutman *et al.*, 1993):

$$T_g = T_o + \frac{bRg + H}{U + V} \qquad (5)$$

where $T_{g,o}$ are the greenhouse and outside temperatures (respectively), Rg the incoming solar radiation, b an efficiency coefficient, H the heating intensity, U the overall energy loss by convection at the walls and V the energy loss by ventilation. U and V are proportional to wind speed. This complete model is described in more details in Tchamitchian and Kittas (2004).

3.3 Weather forecasts: a stochastic environment

For the model to run, *a priori* knowledge of the outside conditions (radiation, temperature, wind speed) is necessary. The RL approach is fit for finding a optimal policy in an uncertain environment: randomly chosen values for the outside conditions are used at each trial of a policy to update the values of the average value and of the relative value functions. However, to limit the range of outside conditions to consider, which will reduce the computational load of the RL approach, weather forecasts can be used as an indicator of the probable future climate. The outside condition values used are drawn randomly according to a uniform distribution centred on the weather forecasts with a standard deviation chosen by the user (depending on the quality of the forecasts for the local situation). Values not physically possible are filtered out, like negative wind speed or radiation, or non zero radiation values at night time.

4. RESULTS AND DISCUSSION

4.1 Trial setup

The optimisations have been carried out using historical data of two late winter in Velestino

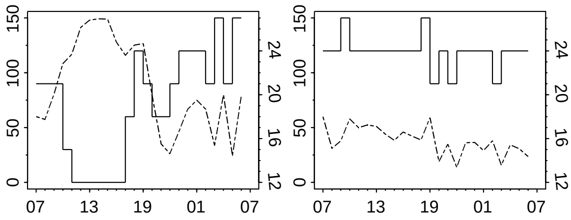

Fig. 2. Optimal heating (left axis & plain line) and resulting greenhouse temperature (right axis & dashed line). Left graph: March 1st; right graph: March 4th, 2004.

(experimental farm of the University of Thessaly, Greece, near Volos) days as weather forecasts. They are presented in figure 1. Given that forecasts are used as a guide to generate random values, using historical data should not alter the conclusions that will be drawn from the results of the optimisation unless the forecasts are biased.

The reference policy adopted has been obtained from local advisors to rose growers. It consists in a night-time set point of 17°C and a daytime set point of 22°C. The greenhouse temperature model is used to compute the energy cost of this policy in order to obtain the E_{ref} reference value. If the night set point cannot be maintained because the boiler cannot deliver the necessary energy, then the instantaneous temperature in the greenhouse is taken as that resulting from the maximum heating instead as the set point. Similarly, day time temperature is corrected of the ventilation cannot maintain the set point. Using this new temperature pattern, the crop model is applied to determine the reference growth and development elements of the criterion.

Learning starts after this step. For a given number of first iterations, the policy to test is chosen randomly in order to explore the states and decision domains. Afterwards, random and greedy policy alternate randomly (greedy policy evaluation is needed to determine the value of ρ, see algorithm description). The best results and also the less sensitive to changes in outside climate, reference policy or initial system states are obtained with a high proportion of greedy policy evaluation (about 80%) and with a minimum number of trials of 200,000 (two hundred thousand). The current implementation of the algorithm takes 90 μs per iteration on a Pentium 3 (700 MHz) computer, and 60% of that time is spent running the model *per se*; 200,000 iterations therefore takes about 20 seconds.

4.2 results

The final results are n_a multidimensional grids (R_i) with two entries, the first being the triplet representing the system state and the second being the decisions that can be taken. In each cell thus defined stands the corresponding relative value ($R - i(s, d)$). These grids and the knowledge of the system states at a given decision step are sufficient to determine the optimal policy according to equation (3), almost like a classical feedback control in which the current action is a function of the observed system states. Figure 2 shows the heating policies and resulting greenhouse temperatures when applying the *optimal* policy to the weather forecasts.

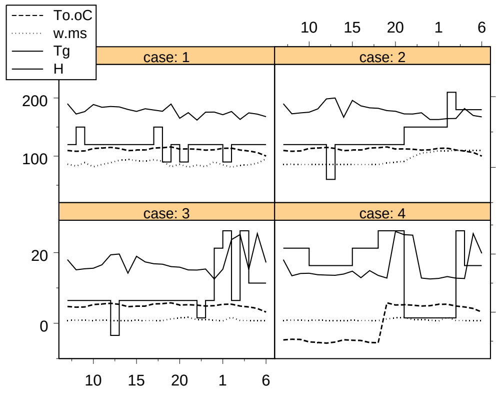

Fig. 3. Optimal policy variations in response to changes in outside weather forecasts. H = heating, stepwise line, right axis. Tg = greenhouse temperature, plain line, left axis. To.oC, outside temperature. w.ms, wind speed.

It can be seen that on March 1st where global radiation provides natural heating during the daylight period, heating is seldom used while at night it used to maintain the temperature well above the lower limit set in this case (14°C). On the contrary, on March 4th where the global radiation provides very little energy, the heating is used throughout the day to maintain the minimum temperature requested by the user (16°C at daylight time, 14°C at night). In both cases, the optimal policy improves the control of the greenhouse and of the crop. The average value for the final criterion is 1.12 on March 1st, 1.04 on March 4th, that is 12% and 4% better than the reference policy respectively (with $\{\alpha_g, \alpha_d, \alpha_e\} = \{1., 1., 2.\}$). The main result is that the R-learning method has been able to exploit the model and to define a control policy consistent with both the knowledge in the model (fortunately!) and common agronomic knowledge. Variations around the climate of March 4th has been studied and are presented in figure 3. Case 1 is the original case, already presented in figure 2. In case 2, the wind speed has been increased at night. More heating is applied during this period to maintain the greenhouse temperature at the minimum level requested by the user, thus decreasing the improvement achieved by the optimisation; the average value of the final criteria is close to one (not better nor worse than the reference policy). In case 3, wind speed has been reduced. Less heating is applied during daytime, more at night. This results in higher temperatures at night than during daytime, which was possible because of the overlap in the day and night bounds that were specified ($\{18, 26\}$ at day, $\{14, 20\}$ at night). Finally, in case 4, daytime temperatures have been drastically reduced. The result is more heating during daytime, again to maintain the temperature above the specified limit. In case 3, the average value of the criterion is less than one, indicating that the proposed policy yields worse results than the reference, which could lead the system to fall back on the reference policy. In case 4, the proposed strategy is about 8% better than the reference.

The previous optimisations have been carried out with an equal weight for the crop related parts of the criterion, the sum of these being equal to the weight set on heating: energy savings was therefore as meaningful as the global measure of the crop performance. A change in the relative weights of these three criterion elements modifies the resulting policies. However, growth and development tend to oppose one to the other: to increase development, high temperatures are needed. However, high temperatures induce a increased rate of respiration and may result in the depletion of the carbon pool compartment in the growth model, a situation which limits the growth. Putting the emphasis on the crop behaviour by deceasing the weight of the energy savings must therefore be made with caution to avoid this situation.

5. CONCLUSION

The goal of this study was to prove the applicability and effectiveness of reinforcement learning to the greenhouse climate control problem. Although simple, the model used encompasses the two main aspects of crop processes: growth and development. The results obtained so far shows that the R-learning algorithm described here can adequately solve the problem and provides the equivalent of a feedback control law that can be applied to greenhouses. One of the key features of the method is that it is designed for finding optimal policies under uncertain future, which is the case here. In that respect it is far more advantageous than, for example, optimal control approach which find an optimal solution for a *given* future rather than for a range of *possible* futures.

The next steps of the study are first to complete and validate the crop model to make it more accurate and second to adopt a continuous approach to the problem.

REFERENCES

Bakker, J. C. and J. A. M. van Uffelen (1988). The effects of diurnal temperature regimes on growth and yield of glasshouse sweet pepper. *Netherland J. Agric. Sci.*

Bredmose, N. and J. Nielsen (2004). Effect of thermoperiodicity and plant population density on stem and flower elongation, leaf development, and specific fresh weight in single stemmed rose (*Rosa hybrida* L. plants. *Scientia Hortic.* **100**(1–4), 169–182.

Dayan, E., E. Presnov and M. Fuchs (2003). Prediction and calculation of morphological characteristics and distribution of assimilates in the ROSGRO model. *Math. Comput. Simul.* **65**(1–2), 101–116.

Gonzalez-Real, M. M. and A. Baille (2000). Changes in leaf photosynthetic parameters with leaf position and nitrogen content within a rose plant canopy (*Rosa hybrida*). *Plant Cell Environ.* **23**(4), 351–363.

Gutman, P. O., P. O. Lindberg, I. Ioslovich and I. Seginer (1993). A non-linear optimal greenhouse control problem solved by linear programming. *J. Agric. Eng. Res.* **55**(4), 335–351.

Heuvelink, E. (1989). Influence of day and night temperature on the growth of young tomato plants. *Scientia Hortic.* **38**, 11–22.

Pasian, C. C. and J. H. Lieth (1996). Prediction of rose shoot development: Model validation for the cultivar 'cara mia' and extension to the cultivars 'royalty' and 'sonia'. *Scientia Hortic.* **66**(1–2), 117–124.

Puterman, R. L. (1994). *Markov decision processes.* John Wiley & Sons. New-York, NY (USA).

Seginer, I., C. Gary and M. Tchamitchian (1994). Optimal temperature regimes for a greenhouse crop with a carbohydrate pool: a modelling study. *Scientia Hortic.* **60**, 55–80.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: an introduction.* MIT Press. Cambridge, MA (USA).

Tchamitchian, M. and C. Kittas (2004). An alternative rose crop model for greenhouse temperature control. In: *26th Annual Congress.* Hellenic Society for Biological Sciences. University of Thessaly, Volos (GRC). p. 13 pp.