THERMAL INSTABILITY

AND

THE CONVECTIVE STABILITY OF STELLAR CHROMOSPHERES

Thesis by

Richard John Defouw

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1970

(Submitted January 19, 1970)

ACKNOWLEDGMENTS

I am indebted to Dr. W. Sargent for an extremely valuable conversation. In this conversation, at an orals party a few years ago, Dr. Sargent explained to me the phenomenon of thermal instability, which is the basis of this thesis. In addition, Dr. Sargent suggested that I make the acquaintance of Dr. Goldreich.

Peter Goldreich has been my research adviser during my entire stay at Caltech. I can only say that I couldn't have had a better adviser.

I have had several discussions with Dr. H. Zirin. I would like to thank him for his interest and also for teaching me a good deal of astrophysics.

Dr. E. Spiegel has kindly read several of my papers, and his suggestions have been very helpful.

I have benefited from consultations with Drs. R.F. Christy, G.B. Field, R. Howard, and R.B. Leighton.

I would like to thank Dr. Oke for all his help and, in particular, for giving me the opportunity to see what teaching is like.

I have appreciated the hospitality and assistance offered by Mrs. Eleanor Ellison, Lilo Hauck, and Helen Holloway.

ABSTRACT

It is generally believed that stellar chromospheres are stable against convection since the Schwarzschild criterion indicates stability when temperature increases with height. It is shown, however, that the Schwarzschild criterion does not apply to chromospheres because it ignores the possibility of thermal instability. In the absence of a magnetic field, thermally unstable regions of a chromosphere will be overstable if the temperature inversion is sufficiently steep. This overstability may explain the origin of a certain class of oscillations in the solar chromosphere. Thermally unstable regions containing magnetic fields are monotonically unstable for all values of the temperature gradient. It is suggested that this monotonic instability of magnetic regions is responsible for spicule formation in the solar chromosphere. Elementary considerations of thermal balance predict that the temperature gradient should diverge at levels of marginal stability. The chromospheric region of spicule formation should therefore be bounded below by an abrupt temperature jump.

The above results are derived by analyzing the stability of a simple model chromosphere in which all

neutral hydrogen atoms are assumed to be in the ground state. Although the model chromosphere emits only free-bound radiation, its thermal instability is caused by the same ionization effects which lead to instability in plasmas emitting line radiation. Thermally unstable regions of a stellar chromosphere, although not represented in detail by the model, should behave in a similar fashion.

Table of Contents

BACKGROUND AND SUMMARY

The concept of thermal instability is very simple. Consider a gas which is losing energy to its environment by one or more processes and which, at the same time, is gaining energy by another set of processes. Suppose that this gas is initially in a steady state with the rate of energy output equal to the rate of energy input. Now imagine that the temperature of the gas is increased by a small amount. If, as a result of this temperature rise, the rate of energy output exceeds the rate of energy input, the gas will lose energy and cool down to its original temperature. The gas is then said to be thermally stable. If, on the other hand, the temperature rise causes the rate of energy output to fall below the input rate, heat will accumulate in the gas and its temperature will continue to increase. This type of behavior is called thermal instability. Of course, one can imagine that the steady state of the gas is perturbed by decreasing rather than increasing the temperature. In this case, the gas is thermally unstable if the temperature change raises the cooling (energy output) rate above the heating (energy input) rate.

These ideas can be expressed more concisely as follows. We define the generalized heat loss function, $\mathcal{L}$, as the rate of energy output per unit mass (ergs/gram/second)

minus the rate of energy input per unit mass. In a steady state, the energy output equals the energy input so that $\mathcal{L} = 0$. The gas cools if $\mathcal{L} > 0$ and heats up if $\mathcal{L} < 0$. Therefore, a gas is thermally unstable if

$$\frac{\partial \mathcal{L}}{\partial T} < 0, \qquad (1)$$

where T is the temperature. This instability criterion was derived by Parker (1953).

As a specific example, consider a partially ionized hydrogen gas. An energetic electron in this gas will occasionally collide with a neutral atom and lose some of its kinetic energy by exciting the atom from the ground state to some higher level. In general, the excited atom will return spontaneously in one or more steps to the ground state by emitting photons, the total energy of which equals the kinetic energy lost by the incident electron. The photons escape provided the gas is transparent. Thus the gas cools by converting thermal (kinetic) energy to excitation energy and then to radiant energy, which leaves the system. In a steady state, this energy output must be balanced by some form of energy input. For simplicity, we shall not discuss the details of any particular mechanism of energy input. Instead, we simply assume that the input balances the output initially and that the input rate is unaffected by any perturbation of the gas. In this

hypothetical situation, the gas is evidently unstable if an increase in temperature reduces the rate of energy output.

Now an increase in temperature of the hydrogen gas will affect the energy output in two ways. First, the number of electrons with sufficient energy to excite hydrogen atoms will increase. This effect will tend to increase the cooling rate. However, the number of electrons with enough energy to ionize hydrogen will also increase. Therefore the number of neutral atoms will decrease. The second effect of an increase in temperature is therefore a reduction in the number of target atoms for inelastic collisions; this effect tends to reduce the cooling rate. When the hydrogen gas is largely neutral, the first effect is more important; an increase in temperature is accompanied by an increase in cooling rate, and the gas is stable. On the other hand, when the gas is already mostly ionized, the increased ionization which follows a rise in temperature causes the cooling rate to fall below the input rate. We therefore expect thermal instability when the gas is sufficiently ionized.

Several authors have pointed out that the instability criterion (1) is not quite correct because it overlooks the following simple fact: the rise in pressure which accompanies an increase in temperature causes a gas to expand. Owing to this decrease in density, the rate of

inelastic collisions is reduced. Thus an increase in temperature will, by this mechanism, always tend to reduce the cooling rate, as required for thermal instability. The instability criterion which takes account of density variations is (Field 1965)

$$\frac{\partial \mathcal{L}}{\partial T} - \frac{\rho}{T}\frac{\partial \mathcal{L}}{\partial \rho} < 0, \qquad (2)$$

where ρ is the density. Since the rate of binary collisions per unit volume is proportional to ρ^2, the energy output per unit mass is proportional to ρ. Thus $\partial \mathcal{L}/\partial \rho > 0$, in general, so that criterion (2) indicates that density variations have a destabilizing effect, as expected.

Stellar photospheres, which are in or near local thermodynamic equilibrium, are thermally stable (this is proved in Section IVb of Part I). However, sufficiently ionized regions of stellar chromospheres are probably unstable, and coronas may be as well. In this thesis I try to answer the following question: If some region of a stellar atmosphere is thermally unstable, how will this instability be manifested?

Athay and Thomas (1956) suggested that thermal instability governs the temperature structure of an atmosphere. They argued that, as one ascends in a chromosphere, the temperature increases slowly in stable regions but that in unstable regions the temperature rises swiftly to the next region of stability. In this way Athay and Thomas hoped to account

for the temperature plateaus and abrupt temperature jumps which are thought to exist in the solar atmosphere. I will now show very simply that this behavior is not a result of thermal instability.

In a steady state, the energy output equals the energy input, i.e., $\mathcal{L} = 0$. Since $\mathcal{L} = 0$ for all heights z, it is also true that $d\mathcal{L}/dz = 0$, or

$$\mathcal{L}_T \frac{dT}{dz} + \mathcal{L}_\rho \frac{d\rho}{dz} = 0, \tag{3}$$

where subscripts have been used to denote the partial derivatives. If P is the pressure and R is the gas constant, the equation of state is

$$P = R\rho T, \tag{4}$$

logarithmic differentiation of which yields

$$\frac{1}{P}\frac{dP}{dz} = \frac{1}{\rho}\frac{d\rho}{dz} + \frac{1}{T}\frac{dT}{dz}. \tag{5}$$

The equation of hydrostatic balance is

$$\frac{dP}{dz} = -\rho g, \tag{6}$$

where g is the gravity. From equations (3), (5), and (6) we find that the temperature gradient of an atmosphere in thermal and hydrostatic balance is

$$\frac{dT}{dz} = \frac{\rho^2 g \mathcal{L}_\rho}{P\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)}. \tag{7}$$

We have already noticed that $\mathcal{L}_\rho > 0$ in general, and, by the instability criterion (2), we see that $\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho > 0$

in a thermally stable gas. According to equation (7), therefore, $dT/dz > 0$ in thermally stable regions of an atmosphere. As one ascends through such a region, the temperature and therefore the degree of ionization of the gas increase, and one approaches a region of thermal instability. At the level of marginal stability, the denominator of equation (7) vanishes and the temperature gradient diverges. At this point, conduction (which has been neglected) is obviously important, and equation (7) does not really apply. However, it is clear that the abrupt temperature jumps in the solar atmosphere are easily explained in terms of the requirement of thermal balance without invoking the concept of thermal instability.

Equation (7) has been derived in this introduction because, in addition to explaining an important feature of the solar atmosphere, it underlies both the weakest point and one of the strong points of the theory to be described below. The discussion of this equation will therefore be resumed later.

I suggest that thermal instability in a stellar atmosphere leads to convective motions. As an element of fluid heats up and expands, it experiences a buoyancy force and rises. Similarly, a cooling element of fluid condenses and falls. In Part I the standard treatment of thermal instability (Field 1965) is used to analyze the convective stability of a stellar atmosphere. It is shown that a

thermally unstable atmosphere is always convectively unstable, regardless of the atmospheric temperature gradient. In a sufficiently steep temperature inversion, this "thermal-convective instability" takes the form of exponentially amplifying oscillations (overstability). In the presence of a magnetic field, however, monotonic instability (by which I mean the ordinary type of instability in which a disturbance grows with time in a monotonic rather than oscillatory fashion) is possible for any value of the temperature gradient.

After making the calculations presented in Part I, I wanted to determine the nature of thermal instability in the solar chromosphere. The two possibilities were:

(a) The primary radiation losses from the chromosphere could result from radiative captures of electrons by protons. Since the emission rate per gram for this process is proportional to $\rho T^{-1/2}$, previous authors have decided on the basis of criterion (1) or (2) that a plasma radiating mainly by free-bound transitions is thermally unstable.

(b) Energy could be radiated from the chromosphere mainly in spectral lines (of hydrogen); thermal instability would then result from the ionization effect described earlier.

Although the radiation by free-bound transitions is proportional to $\rho T^{-1/2}$, it should be noted that this radiation does not always represent a loss of thermal

energy. The energy of a recombination photon equals the kinetic energy of the recombining electron plus the energy required to ionize the neutral atom. For this and other reasons I decided to investigate in detail the thermal stability of an idealized hydrogen plasma consisting of protons, electrons, and hydrogen atoms in the ground state. Since excited bound levels are excluded, this model plasma emits only free-bound radiation (the free-free emission is neglected). The stability analysis, which is given in Part II, shows that the model plasma is thermally unstable if and only if

$$\left(\mathcal{L}_x - \frac{\rho}{1+x}\mathcal{L}_\rho\right)\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) - \left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)\left(\mathcal{J}_x - \frac{\rho}{1+x}\mathcal{J}_\rho\right) < 0, \quad (8)$$

where x is the fraction of hydrogen which is ionized and $\mathcal{J}$ is the number of ionizations minus the number of recombinations per gram per second. When ionization is caused by atom-electron collisions and recombination is radiative, inequality (8) is satisfied only if $T > 17500\,^{\circ}K$. Thus, even though the plasma emits only free-bound radiation, it is stable for $T < 17500\,^{\circ}K$.

The instability (for $T > 17500\,^{\circ}K$) of the model plasma results from the same ionization effect described earlier. The primary cooling mechanism is the inelastic collision process of collisional ionization. An increase in temperature increases the ionization rate, but the cooling rate soon falls owing to the reduction in the concentration of

neutral atoms. Thus, although the model plasma emits no line radiation, the instability mechanism is essentially the same as in a plasma radiating via bound-bound transitions. In other words, there is no special type of thermal instability associated with free-bound radiation. The answer to the question regarding the cause of instability in the chromosphere is simply that instability results from ionization effects, regardless of whether most of the radiation is emitted in lines or free-bound continua.

The thermal-convective instability of the model hydrogen plasma is analyzed in Part III. This analysis confirms most of the conclusions of Part I. For example, the model plasma is monotonically unstable when subjected to both gravitational and magnetic fields for all values of the temperature gradient provided the plasma is thermally unstable in the absence of the fields. In a gravitational field alone, the thermally unstable plasma is overstable in a sufficiently steep temperature inversion. However, even when inequality (8) is satisfied, the model plasma is stable in a gravitational field (with no magnetic field) for sufficiently small values of the density and the temperature gradient, contrary to the results of the simplified analysis of Part I.

The model hydrogen plasma is obviously not an accurate representation of a stellar chromosphere. However, the model probably does contain the essential physics as far

as thermal and thermal-convective instability are concerned (see Section 4 of Part III). I therefore submit Part III as an analysis in the first approximation of the convective stability of stellar chromospheres.

In the past, astronomers have assumed that the convective stability of stellar chromospheres is governed by the Schwarzschild criterion, according to which there is instability only if the temperature decreases with increasing height more rapidly than the so-called adiabatic lapse rate. In chromospheres, of course, the temperature increases with height, and the Schwarzschild criterion predicts extreme stability. For this reason the observation of gas jets, called spicules, in the solar chromosphere has been very perplexing. I suggest that spicules are produced by monotonic thermal-convective instability. Theory and observation are compared in Section 6 of Part III. The main points to note are:

(a) The theory predicts the existence of gas jets only in regions containing appreciable magnetic field. Spicules are observed only at the cell boundaries of the chromospheric network, which is where the magnetic field is concentrated.

(b) Instability requires that the temperature exceed some critical value. The value of T_{crit} for the simple model plasma is 17500°K, but more detailed considerations indicate that T_{crit} is about 12000°K in the solar chromo-

sphere. Probably the best estimates of spicule temperatures are the values 14000 to 17000 °K found by Beckers (1968).

(c) The theory predicts that the unstable region of spicule formation should be bounded below by an abrupt temperature jump (see equation [7] or Section 5 of Part III). In fact, some astronomers have concluded that spicules begin to appear just at the height at which there is a steep temperature rise. Other astronomers, however, feel that the evidence for lack of spicules below this height is weak. Nevertheless, I mention this point because this prediction of the theory was the only one made before the author was aware of the corresponding observation.

The main shortcoming of the thermal-convective theory of spicule formation is that the temperature structure of the unstable region is completely unknown. No particular value of the temperature gradient is required by the theory, but if the gradient is too large, the potentially unstable region will be so thin that disturbances will be smoothed away by conduction. A theory for the structure of the unstable region must obviously include conduction (see equation [7]) and may also require a reliable calculation of the energy input (see Section 5 of Part III). Thus, it is unlikely that this difficulty will be resolved in the near future.

Field-free regions of the solar chromosphere are

almost certainly not monotonically unstable, but overstability is possible and may explain the origin of a certain class of chromospheric oscillations. This application of the theory is particularly tentative, and the reader is referred to Section 6 of Part III for further discussion.

The chromosphere of the sun is the only chromosphere discussed in this thesis simply because it is the only one for which detailed dynamical observations are available. It must be emphasized that the physical processes involved in the thermal-convective theory are quite general, and if the theory applies to the solar chromosphere, it must also apply to the chromospheres of many other stars.

The only source of information concerning chromospheric dynamics for stars besides the sun is the phenomenon of eclipses of relatively small early-type stars by the extended atmospheres of late-type giants in systems such as ζ Aurigae and 31 Cygni (see the review by Wilson [1964]). Whether the observed differential motions can be interpreted in terms of the thermal-convective theory depends on whether hydrogen ionization in the chromosphere of the giant is collisional or whether it is due mainly to the ionizing radiation from the early-type companion. Apparently this point is subject to some debate.

Each of the following three parts has been written as a self-contained paper. The reader interested in the theory of thermal instability may wish to read only Part II,

while solar physicists may be interested only in Part III.

References

Athay, R.G. and Thomas, R.N. 1956, Ap.J., 123, 299.

Beckers, J.M. 1968, Solar Phys., 3, 367.

Field, G.B. 1965, Ap.J., 142, 531.

Parker, E.N. 1953, Ap.J., 117, 431.

Wilson, O.C. 1964, in Stellar Atmospheres (ed. by J.L. Greenstein), University of Chicago Press, p. 436.

PART I

THERMAL-CONVECTIVE INSTABILITY

ABSTRACT

The Schwarzschild criterion for convection is generalized to include departures from adiabatic motion. It is demonstrated that a thermally unstable atmosphere is also convectively unstable, regardless of the atmospheric temperature gradient. If the latter is sufficiently subadiabatic (e. g., if the temperature increases rapidly with height), convection sets in as exponentially growing oscillations. In the presence of a magnetic field, a thermally unstable atmosphere is monotonically unstable, although overstability is also possible if the temperature gradient is subadiabatic. The effects of conduction, viscosity, opacity, and rotation are evaluated. In this paper the assumption is made that the radiative cooling (or the source function) depends only on the local values of density and temperature.

I. INTRODUCTION

The term "thermal instability" is sometimes used to denote the instability of a fluid layer heated from below. In the present paper the type of instability in which motions are driven by buoyancy forces will be called convective instability. The term "thermal instability" will be reserved for the instability first discussed by Parker (1953) and extensively explored by Field (1965).

Parker considered a heat equation of the form

$$c_v \frac{dT}{dt} = - \mathcal{L}, \tag{1}$$

where c_v is the specific heat at constant volume, T is the temperature, and t is the time. The quantity $\mathcal{L}$ is the energy lost minus the energy gained per gram per second. In equilibrium, of course, this so-called heat loss function vanishes. By convention, $\mathcal{L}$ does not include energy transfer by conduction. The effect of conduction was investigated by Parker but will be omitted from the present discussion (see §IV). If the equilibrium temperature of a uniform medium is perturbed by an amount θ, equation (1) states that, to first order,

$$c_v \frac{\partial \theta}{\partial t} = - \mathcal{L}_T \theta, \tag{2}$$

where $\mathcal{L}_T$ is the derivative of $\mathcal{L}$ with respect to T evaluated in the equilibrium state. It is evident from equation (2) that the temperature perturbation will grow exponentially if

$$\mathcal{L}_T < 0. \tag{3}$$

Weymann (1960) pointed out that the condition (3) for thermal

instability applies only to isochoric perturbations. But density perturbations are to be expected owing to the strong tendency of a fluid to remain in pressure equilibrium. A work term must therefore be added to equation (1), and the density (ρ) dependence of the heat loss function must be taken into account. For an ideal gas, the resulting condition for thermal instability is (Field 1965)

$$\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho < 0, \tag{4}$$

where the subscripts denote partial derivatives.

The occurrence in thermal instability of density perturbations in pressure equilibrium with their surroundings clearly has a bearing on thermal convection. The main purpose of this paper is to show that a stellar atmosphere, for example, which is thermally unstable is also, in fact, convectively unstable, regardless of the atmospheric temperature gradient. The convective instability of a thermally unstable atmosphere will be called thermal-convective instability.

Most of the previous work on thermal instability has concerned the behavior of perturbations of an initially uniform medium in the absence of a gravitational field. One exception is Field's (1965) analysis of the effect of density stratification in a gravitational field. However, Field considered only modes with no horizontal motions, and he omitted the acceleration term in the equation of vertical motion. Both restrictions preclude convective-type motions. Instead, thermal instability manifests itself in Field's model by either an overall expansion or an overall contraction of the atmosphere.

In the opinion of the present author, thermal instability of a stellar atmosphere is more probably manifested as thermal-convective

instability than as an overall expansion or contraction. In §II the parcel method is used to demonstrate the existence and properties of thermal-convective instability. In this idealized calculation all sources of dissipation, such as viscosity and conduction, are neglected. The effects of molecular transport processes, opacity, rotation, and magnetic fields are considered in §IV. It will be necessary to be able to distinguish ordinary instability, in which a perturbation increases monotonically with time, from overstability. The former type of instability will be called monotonic instability since the term "dynamical instability" which is sometimes used may be confusing in the present context. Also, following conventional usage, I will characterize a temperature gradient as subadiabatic if the temperature either increases with height or decreases less rapidly than the adiabatic lapse rate.

II. PARCEL METHOD

Consider a parcel of gas of density ρ^*, temperature T^*, and pressure P^*. The ambient values of the gas variables will be denoted by the corresponding symbols without asterisks. Since the parcel will quickly reach pressure equilibrium with its surroundings, we have

$$P^* = P , \tag{5}$$

which for an ideal gas means that

$$\frac{\rho^* - \rho}{\rho} + \frac{T^* - T}{T} = 0 , \tag{6}$$

where we have assumed that the parcel differs only slightly from its environment.

The heat equation for the parcel can be written in the form

$$\frac{dP^*}{dt} - \frac{\gamma P^*}{\rho^*}\frac{d\rho^*}{dt} + (\gamma-1)\rho^* \mathcal{L}(\rho^*, T^*) = 0, \quad (7)$$

where d/dt is the convective derivative following the motion of the parcel and γ is the ratio of specific heats (assumed constant). We have assumed that the heat loss function of the parcel depends only on its density and temperature. Since the atmosphere is initially in equilibrium, we have $\mathcal{L}(\rho, T) = 0$ so that

$$\begin{aligned}\mathcal{L}(\rho^*, T^*) &= \mathcal{L}_\rho(\rho^*-\rho) + \mathcal{L}_T(T^*-T) \\ &= -\frac{T}{\rho}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)(\rho^*-\rho), \quad (8)\end{aligned}$$

where use has been made of equation (6). In view of equations (5) and (8), equation (7) may be written

$$\frac{dP}{dt} - \frac{\gamma P}{\rho}\frac{d\rho^*}{dt} - (\gamma-1)T\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)(\rho^*-\rho) = 0 \quad (9)$$

to first order in deviations from the equilibrium state.

Now let w be the upward (z) velocity of the parcel so that

$$\frac{dP}{dt} = w\frac{dP}{dz} \quad \text{and} \quad \frac{d\rho}{dt} = w\frac{d\rho}{dz}. \quad (10)$$

With the aid of equations (10) and a little algebra, we may rewrite equation (9) in the following form:

$$\frac{d}{dt}(\rho^*-\rho) + \frac{1}{c_p}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)(\rho^*-\rho) - (\beta-\beta_{ad})\frac{\rho}{T}w = 0, \quad (11)$$

where c_p is the specific heat at constant pressure, β is the atmospheric lapse rate,

$$\beta = \frac{dT}{dz} = T\left(\frac{1}{P}\frac{dP}{dz} - \frac{1}{\rho}\frac{d\rho}{dz}\right), \tag{12}$$

and

$$\beta_{ad} = \left(\frac{\gamma-1}{\gamma}\right)\frac{T}{P}\frac{dP}{dz} \tag{13}$$

is the adiabatic temperature gradient.

The equation of motion of the parcel is

$$\rho^* \frac{dw}{dt} = -(\rho^* - \rho)g, \tag{14}$$

where g is the acceleration due to gravity. Solving equation (14) for $\rho^* - \rho$ and substituting into equation (11), we get (to first order)

$$\frac{\partial^2 w}{\partial t^2} + \frac{1}{C_p}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)\frac{\partial w}{\partial t} + \frac{g}{T}(\beta - \beta_{ad})w = 0. \tag{15}$$

Equation (15) is satisfied by $w \sim e^{nt}$, where the growth rate is

$$n = -\frac{1}{2C_p}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right) \pm \left[\frac{1}{4C_p^2}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)^2 - \frac{g}{T}(\beta - \beta_{ad})\right]^{1/2}. \tag{16}$$

Equation (16) contains several well known results. Setting $g = 0$, we obtain the growth rate for perturbations of a uniform medium:

$$n(g=0) = -\frac{1}{C_p}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right). \tag{17}$$

Comparison with Field's (1965) analysis shows that equation (17) is a good approximation to the growth rate of the condensation (or rarefaction) mode provided the growth rate divided by the perturbation wavenumber is much

less than the speed of sound. In other words, the analysis presented here is valid when the size of the parcel is much less than the sound speed multiplied by the time scale, which is the condition for pressure equilibrium. In addition, the parcel size must be much less than a scale height for the ambient gas variables to be uniquely defined.

If, on the other hand, we assume adiabatic motion, equation (16) yields

$$n(\mathcal{L}=0) = \pm\left[-\frac{g}{T}(\beta-\beta_{ad})\right]^{1/2}, \qquad (18)$$

which means that there will be a value of n which is real and positive if and only if $\beta < \beta_{ad}$ (Schwarzschild criterion for convective instability). If $\beta > \beta_{ad}$, equation (18) gives the Brunt-Väisälä frequency of gravity oscillations.

In addition to these well known results, equation (16) shows that a thermally unstable atmosphere, which satisfies inequality (4), is also unstable against convection. This result, in the form just stated, is completely independent of the atmospheric temperature gradient. However, if the latter is strongly subadiabatic, convection will set in as exponentially amplifying oscillations (overstability). If inequality (4) is not satisfied, convective stability is determined by the Schwarzschild criterion.

III. DESCRIPTION OF THE INSTABILITY

The close connection between thermal and convective instability demonstrated in §II results from an intimate relation between the thermal and dynamic behavior of a parcel which we now analyze in detail. For the sake of clarity, we transpose terms in equation (11) to get

$$\frac{d}{dt}(\rho^* - \rho) = -\frac{1}{C_p}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)(\rho^* - \rho) + (\beta - \beta_{ad})\frac{\rho}{T}w. \qquad (19)$$

Let us first suppose that the rate of change of density excess or deficit is determined primarily by the effects of heating or cooling of the parcel. Then the first term on the right side of equation (19) dominates the second term. If inequality (4) is satisfied, a density perturbation will grow monotonically and exponentially. As we see from equation (14), the parcel will either fall or rise, depending on whether it is more or less dense than its surroundings. According to equation (16), this case of monotonic instability obtains if, in addition to inequality (4), we have

$$\frac{1}{4C_p^2}\left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)^2 \geq \frac{g}{T}(\beta - \beta_{ad}). \qquad (20)$$

When inequality (4) is satisfied but inequality (20) is not, equation (16) indicates overstability. In order to understand the overstability, consider first adiabatic motion in an atmosphere which is stable according to the Schwarzschild criterion. Then the first term on the right side of equation (19) vanishes, and the second term has the same sign as the upward velocity w. If a parcel of gas is given a slight push upwards, $w > 0$ and the parcel becomes more dense than its surroundings. Its upward velocity therefore diminishes (equation [14]). When the parcel has reached its highest level ($w = 0$), its density excess is a maximum. As the parcel descends ($w < 0$), its density excess decreases and eventually changes sign. For this case of $\mathcal{L} = 0$, the parcel performs simple harmonic oscillations.

Now let us follow the parcel motion taking into account the first term on the right side of equation (19). When the parcel is given an

upward push, its density rises above the ambient density as before. Then the first term on the right side of equation (19) reinforces the second term (provided inequality [4] is satisfied). The parcel experiences an increasing density excess while its upward velocity diminishes. When the parcel has attained its maximum elevation ($w = 0$), its density excess continues to increase owing to the cooling term in equation (19). Thus, as the parcel falls back, its velocity of descent is greater than in the adiabatic case. Instead of executing simple harmonic motion, the parcel shoots below its original position with increased speed, and an oscillation with exponentially increasing amplitude has begun.

IV. BOUSSINESQ THEORY

So far we have studied thermal-convective instability by following the motion of a fluid parcel. The parcel calculation of §II was idealized in that all sources of dissipation were ignored. In this section we examine the dissipative effects of viscosity and conduction as well as the effects of opacity, rotation, and magnetic fields. Since gradients of physical variables are important in these considerations, it is convenient to switch from the Lagrangian parcel approach to an Eulerian description. Apart from terms in the heat loss function, $\mathcal{L}$, the equations used in this section are those of standard convection theory (Chandrasekhar 1961). We employ the Boussinesq approximation (Boussinesq 1903) modified so as to apply to thin layers of compressible fluids (Spiegel and Veronis 1960).

Of the various boundary conditions used in ordinary convection theory, we shall adopt those corresponding to free boundaries since these are the simplest and also the most appropriate for stellar atmospheres

(Spiegel 1965). Thus at both the upper and lower boundaries of the infinite horizontal layer of fluid we require

$$\left.\begin{aligned} \theta &= 0 \\ w &= 0 \\ \frac{\partial^2 w}{\partial z^2} &= 0 \end{aligned}\right\}, \tag{21}$$

where θ is the temperature perturbation and w is the vertical (z) component of the fluid velocity.

a) Effects of Viscosity and Conduction

The first law of thermodynamics may be written in the form

$$c_v \frac{dT}{dt} = -\mathcal{L} + \frac{K}{\rho}\nabla^2 T + \frac{P}{\rho^2}\frac{d\rho}{dt}, \tag{22}$$

where K is the thermal conductivity (assumed constant). Viscous dissipation has been omitted from equation (22) as it is a quantity of second order. We expand $\mathcal{L}$ as in equation (8) and substitute the density perturbation from the Boussinesq equation of state:

$$\frac{\delta\rho}{\rho} = -\alpha\theta. \tag{23}$$

Equation (23) is equivalent to equation (6), and we see that the coefficient of thermal expansion is $\alpha = 1/T$. Next we evaluate the vertical gradient of the unperturbed density with the aid of equation (12) and the equation of hydrostatic equilibrium. After these substitutions, equation (22) becomes

$$\frac{\partial\theta}{\partial t} + \frac{1}{c_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho)\theta - \chi\nabla^2\theta + (\beta + \frac{g}{c_p})w = 0, \tag{24}$$

where $\chi = K/c_p\rho$ is the coefficient of thermometric conductivity.

In the Boussinesq approximation, the continuity equation and the linearized equation of motion yield (Chandrasekhar 1961, p. 21)

$$\frac{\partial}{\partial t}\nabla^2 w = g\alpha\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}\right) + \nu\nabla^4 w, \tag{25}$$

where ν is the kinematic viscosity and x and y are the horizontal coordinates.

Equations (24) and (25), as well as the boundary conditions (21), are satisfied if both θ and w vary as

$$e^{nt}\, e^{i(k_x x + k_y y)} \sin k_z z, \tag{26}$$

where k_z is an integral multiple of π divided by the thickness of the fluid layer. When expression (26) is substituted into equations (24) and (25), these equations become a pair of homogeneous, simultaneous, algebraic equations. Setting the determinant of the coefficients equal to zero, we obtain a quadratic characteristic equation, the solutions of which are

$$n = -\frac{1}{2}\left[\frac{1}{c_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + (\chi+\nu)k^2\right] \pm \left\{\frac{1}{4}\left[\frac{1}{c_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + (\chi-\nu)k^2\right]^2 - \left(\beta + \frac{g}{c_p}\right)\Gamma\right\}^{1/2}, \tag{27}$$

where

$$k^2 = k_x^2 + k_y^2 + k_z^2 \tag{28}$$

and

$$\Gamma = g\alpha\,\frac{k_x^2 + k_y^2}{k^2}\,. \tag{29}$$

Apart from the terms in $\mathcal{L}$, equation (27) is just the result of Rayleigh's (1916) original investigation modified by effects of compressibility (Spiegel and Veronis 1960). If we set $\chi = \nu = 0$ in equation (27), we recover equation (16) with an additional factor of $(k_x^2 + k_y^2)/k^2$ in the temperature gradient term.

Equation (27) indicates that there will be some form of instability for any value of β if

$$\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + (1+p)\chi k^2 < 0, \qquad (30)$$

where $p = \nu/\chi$ is the Prandtl number. Thus conduction and viscosity stabilize high wavenumber perturbations. However, there is always thermal-convective instability if inequality (4) is satisfied and the dimensions of the perturbation are sufficiently large. For the astrophysically interesting case of small Prandtl number, the perturbation wavelengths must be large enough so that the time scale for conductive smoothing of temperature fluctuations, of order $(\chi k^2)^{-1}$, is greater than the growth time in the absence of conduction. Finally, we note from equation (27) that the instability takes the form of overstability if the temperature gradient is sufficiently subadiabatic, as we saw in §II.

b) Radiative Transfer Effects

The use of a heat loss function, $\mathcal{L}(\rho, T)$, which depends only on the local values of density and temperature requires that the gas is optically thin. We now examine the effects of radiative transfer which arise when self-absorption is not negligible.

We shall treat radiative transfer with the aid of the Eddington approximation. The usefulness of this approximation in gas dynamics has been pointed out by Vincenti and Kruger (1965) and by Unno and Spiegel (1966) (when Spiegel [1964] criticized the use of the Eddington approximation in convection studies, he was actually referring to the Rosseland diffusion approximation). These authors emphasize that the Eddington approximation is accurate in both the optically thick and optically thin limits. Using this approximation, one may derive the following heat equation for a grey gas:

$$\rho c_v \frac{dT}{dt} - \frac{P}{\rho}\frac{d\rho}{dt} = \nabla\cdot\left\{\frac{1}{3\kappa\rho}\nabla\left[\frac{1}{\kappa\rho}\left(\rho c_v \frac{dT}{dt} - \frac{P}{\rho}\frac{d\rho}{dt}\right) + 4\pi S\right]\right\}, \tag{31}$$

where κ is the mass absorption coefficient and S is the source function. Equation (31) is an elementary generalization of a formula given by Unno and Spiegel (1966).

In the same way we derived equation (24) from equation (22), we find from equation (31)

$$3\kappa^2\rho^2\frac{\partial\theta}{\partial t} - \frac{4\pi\kappa}{c_p}\left(S_T - \rho^{\alpha}S_{\rho}\right)\nabla^2\theta - \frac{\partial}{\partial t}\nabla^2\theta + \left(\beta + \frac{g}{c_p}\right)\left(3\kappa^2\rho^2 w - \nabla^2 w\right) = 0, \tag{32}$$

where we have assumed that S and κ are independent of height in the equilibrium state. Although we have assumed in equation (32) that the source function depends only on density and temperature, the analysis is easily generalized to include a dependence of S on the radiation field.

Substitution of expression (26) into equations (25) and (32) results in a pair of simultaneous algebraic equations. The roots of

the corresponding secular equation are

$$n = -\frac{1}{2}\left[\frac{4\pi K(S_T - \rho\alpha S_\rho)}{C_p(1+3K^2\rho^2/k^2)} + \nu k^2\right] \pm \left\{\frac{1}{4}\left[\frac{4\pi K(S_T - \rho\alpha S_\rho)}{C_p(1+3K^2\rho^2/k^2)} - \nu k^2\right]^2 - \left(\beta + \frac{g}{C_p}\right)\Gamma\right\}^{1/2}. \quad (33)$$

If we set $\nu = 0$ and consider the special case of local thermodynamic equilibrium ($S = \frac{\sigma}{\pi} T^4$), equation (33) becomes equivalent to equation (58) of Spiegel (1964) provided we make allowance for the known relationship between results based on the Eddington approximation and those based on the exact integral formulation (Unno and Spiegel 1966).

From equation (33) we see that the real part of n will be positive if

$$\frac{4\pi K(S_T - \rho\alpha S_\rho)}{C_p(1+3K^2\rho^2/k^2)} + \nu k^2 < 0 \quad (34)$$

regardless of the atmospheric temperature gradient. The source functions which lead to thermal-convective instability are therefore those for which

$$S_T - \frac{\rho}{T} S_\rho < 0, \quad (35)$$

which is a generalization of inequality (4).

c) Effect of Rotation

For a fluid rotating with angular velocity Ω about a vertical axis, the continuity equation and the equation of motion yield the following linearized Boussinesq equations (Chandrasekhar 1961, pp. 88 and 89):

$$\frac{\partial \zeta}{\partial t} = \nu \nabla^2 \zeta + 2\Omega \frac{\partial w}{\partial z} \tag{36}$$

and

$$\frac{\partial}{\partial t}\nabla^2 w = g\alpha\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}\right) + \nu\nabla^4 w - 2\Omega\frac{\partial \zeta}{\partial z}, \tag{37}$$

in which ζ is the vertical component of the vorticity. The variable ζ can be eliminated by differentiating equation (37) with respect to time and substituting equation (36). Substitution of equation (37) into the resulting equation yields

$$\left(\frac{\partial^2}{\partial t^2}\nabla^2 + 4\Omega^2\frac{\partial^2}{\partial z^2} - 2\nu\frac{\partial}{\partial t}\nabla^4 + \nu^2\nabla^6\right)w - g\alpha\left(\frac{\partial}{\partial t} - \nu\nabla^2\right)\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}\right) = 0. \tag{38}$$

We now assume that w and θ vary as in expression (26). Equation (38) and the heat equation (24) then imply the following secular equation:

$$\begin{aligned} &n^3 + \left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + (\chi + 2\nu)k^2\right]n^2 \\ &+\left[4\Omega^2 k_z^2/k^2 + (\beta + g/C_p)\Gamma + \frac{2\nu k^2}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + (2\chi+\nu)\nu k^4\right]n \\ &+ (4\Omega^2 k_z^2/k^2 + \nu^2 k^4)\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] \\ &+\nu k^2(\beta + g/C_p)\Gamma = 0. \end{aligned} \tag{39}$$

We shall limit our detailed discussion to inviscid fluids. In this case equation (39) reduces to

$$\begin{aligned} &n^3 + \left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right]n^2 \\ &+\left[4\Omega^2 k_z^2/k^2 + (\beta + g/C_p)\Gamma\right]n + 4\Omega^2(k_z^2/k^2)\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] = 0. \end{aligned} \tag{40}$$

If

$$\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2 < 0, \tag{41}$$

the constant term in equation (40) is negative, which means that this equation has a positive real root. Therefore inequality (41) is a sufficient condition for monotonic instability of a rotating inviscid fluid for all values of β.

To first order in the quantity on the left side of inequality (41), the roots of equation (40) are

$$n = \frac{-4\Omega^2(k_z^2/k^2)[(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho)/c_p + \chi k^2]}{4\Omega^2 k_z^2/k^2 + (\beta + g/c_p)\Gamma} \tag{42}$$

and

$$n = \frac{-(\beta+g/c_p)\Gamma[(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho)/c_p + \chi k^2]}{2[4\Omega^2 k_z^2/k^2 + (\beta+g/c_p)\Gamma]} \pm i[(\beta+g/c_p)\Gamma + 4\Omega^2 k_z^2/k^2]^{1/2}. \tag{43}$$

According to equation (43), overstability is possible if inequality (41) holds and $\beta > \beta_{ad} = -g/c_p$. In ordinary convection theory ($\mathcal{L} = 0$), of course, we expect overstability only if $\beta < \beta_{ad}$.

The root corresponding to monotonic thermal-convective instability is given by equation (42). If we compare this equation with the equation which results when expression (26) is substituted into equation (24), we find that $w = 0$ when $n = 0$. This result is necessary to avoid a conflict with the Taylor-Proudman theorem (cf. Chandrasekhar 1961). However, when inequality (41) is satisfied, it is easily shown from equations (24), (26), and (42) that w no longer vanishes.

d) Effect of a Magnetic Field

We now suppose that the fluid layer is permeated by a vertical magnetic field of strength B. In this case the continuity equation and the equation of motion may be used, in the linear Boussinesq approximation, to derive the following equation (Chandrasekhar 1961, p. 162):

$$\frac{\partial}{\partial t}\nabla^2 w = \frac{B}{4\pi\rho}\frac{\partial}{\partial z}\nabla^2 b_z + g\alpha\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}\right) + \nu\nabla^4 w, \qquad (44)$$

where b_z is the vertical component of the field perturbation and is governed by the induction equation

$$\frac{\partial b_z}{\partial t} = B\frac{\partial w}{\partial z} + \frac{\eta}{4\pi}\nabla^2 b_z, \qquad (45)$$

in which η is the electrical resistivity. Proceeding as in the case of rotation, we combine equations (44) and (45) into the single equation

$$\left[\frac{\partial^2}{\partial t^2} - v_A^2\frac{\partial^2}{\partial z^2} - (\nu+\eta/4\pi)\frac{\partial}{\partial t}\nabla^2 + \frac{\nu\eta}{4\pi}\nabla^4\right]\nabla^2 w - g\alpha\left(\frac{\partial}{\partial t} - \frac{\eta}{4\pi}\nabla^2\right)\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}\right) = 0, \qquad (46)$$

where $v_A = B/(4\pi\rho)^{1/2}$ is the Alfvén speed.

The heat equation should now be modified owing to the anisotropy of the thermal conductivity of an ionized gas in a magnetic field. For simplicity, however, we shall retain equation (24). The correct results can be obtained from the equations which follow by replacing χk^2 by $\chi_\perp(k_x^2 + k_y^2) + \chi k_z^2$, where $\chi_\perp$ is the thermometric conductivity perpendicular to the field and is often negligible.

After substitution of expression (26), equations (24) and (46) yield the characteristic equation

$$n^3 + \left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2 + (\nu + \eta/4\pi)k^2\right] n^2$$
$$+ \left\{ V_A^2 k_z^2 + (\beta + g/C_p)\Gamma + (\nu+\eta/4\pi)k^2\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] \right.$$
$$\left. + \frac{\nu\eta}{4\pi} k^4 \right\} n + \left(V_A^2 k_z^2 + \frac{\nu\eta}{4\pi} k^4\right)\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] \qquad (47)$$
$$+ \frac{\eta}{4\pi} k^2 (\beta + g/C_p)\Gamma = 0.$$

In many cases of astrophysical interest the effects of viscosity and resistivity are negligible. Setting $\nu = \eta = 0$ in equation (47), we get

$$n^3 + \left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] n^2$$
$$+ \left[V_A^2 k_z^2 + (\beta + g/C_p)\Gamma\right] n + V_A^2 k_z^2 \left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] = 0. \qquad (48)$$

This equation becomes identical to equation (40) if we replace v_A by $2\Omega/k$. Inequality (41) is a sufficient condition for monotonic instability in the presence of a magnetic field. To first order in the quantity on the left side of this inequality, the roots of equation (48) are

$$n = \frac{-f}{f+1}\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right] \qquad (49)$$

and

$$n = \frac{-1}{2(f+1)}\left[\frac{1}{C_p}(\mathcal{L}_T - \rho\alpha\mathcal{L}_\rho) + \chi k^2\right]$$
$$\pm i\left[(\beta + g/C_p)\Gamma + V_A^2 k_z^2\right]^{1/2}, \qquad (50)$$

where

$$f = \frac{V_A^2 k_z^2}{(\beta + g/C_p)\Gamma} \qquad (51)$$

is the square of the ratio of an Alfvén frequency to the Brunt-Väisälä frequency. Comparing equations (49) and (50), we see that, when

inequality (41) is fulfilled, the overstable mode should dominate for $f \ll 1$ (assuming, of course, that $[\beta + g/c_p]\Gamma + v_A^2 k_z^2 > 0$) while the monotonically unstable mode dominates for $f \gg 1$.

The existence of monotonic instability when $\eta = 0$ would appear to conflict with the magnetic analogue of the Taylor-Proudman theorem (cf. Chandrasekhar 1961). The resolution of this apparent conflict is the same as in the case of rotation.

The value of the thermal expansion coefficient, α, is determined by considerations of pressure equilibrium. One might therefore expect that α should be influenced by the magnetic contribution to the pressure. In fact, Field (1965) found that when the wave vector of the perturbation is perpendicular to the field, α is reduced by the factor $(1 + v_A^2/v_s^2)^{-1}$, where v_s is the isothermal sound speed. However, Field also found that α is unaffected by the field when the angle between the wave vector and the field is not precisely $\pi/2$. The origin of this result lies, paradoxically, in the importance of the field. Motions are constrained to be primarily along the field, and, since these motions do not change the field, there are no magnetic pressures to be overcome. Similarly, thermal-convective motions will be primarily parallel to the field so that we may assume $\alpha = 1/T$ even in the presence of a magnetic field. For our case of a vertical initial field, this assumption would appear to break down at the boundaries, where we have required $w = 0$. In a real atmosphere, however, convective motions will penetrate stable layers so that the fluid velocity is never required to be strictly perpendicular to the field.

V. CONCLUSION

The convective stability of a star has customarily been decided on the basis of the Schwarzschild criterion. One of the fundamental assumptions used in deriving this criterion is that the motion is adiabatic. In the interior of a star, where the photon mean free path is small, this assumption is justified and the Schwarzschild criterion is applicable. However, in the outer layers of a stellar atmosphere, effective heat transfer is no longer prevented by opacity, and departures from adiabatic motion can be significant.

In this paper we have seen that if an atmosphere is thermally stable, its convective stability is determined by Schwarzschild's criterion, but if it is thermally unstable, it must also be unstable against convection, regardless of the temperature gradient. Thus the Schwarzschild criterion gives a sufficient condition for convective instability but not a necessary condition.

We found that a thermally unstable atmosphere with a sufficiently steep temperature inversion is overstable. This conclusion must be regarded as tentative owing to the local nature of the analysis. Whereas the fluid layer in the theory was isolated from its surroundings by artificial boundary conditions, oscillations of an overstable layer will, in reality, feed energy into waves in adjacent layers and will thereby be damped to some extent. Whether oscillations actually develop in spite of this damping can be decided only by a nonlocal analysis of the medium in question. However, when rotation or a magnetic field is present, the local analysis predicted monotonic instability, and this is less likely to be affected by adjacent stable layers.

It is probable that stellar chromospheres and coronas, and possibly the interstellar medium, exhibit thermal-convective instability. The chromosphere and corona of the sun will be discussed in detail elsewhere. However, before these applications are considered, the atomic physics of thermal instability must be examined in detail. In a forthcoming paper I show that the assumption that the heat loss function depends only on the density and temperature is very restrictive (even for optically thin media) and, in fact, is valid only when the heat loss is free-free emission from a fully ionized gas. Although the present theory strictly applies only to this case, it will be shown in the paper on the solar atmosphere that many of the conclusions of the present paper are of general validity.

I am indebted to Dr. W. L. W. Sargent for introducing me to the concept of thermal instability. I would like to thank Drs. P. Goldreich and E. Spiegel for reading the manuscript and offering some helpful suggestions.

REFERENCES

Boussinesq, J. 1903, Théorie Analytique de la Chaleur (Gauthier-Villars, Paris), vol. 2, §261.

Chandrasekhar, S. 1961, Hydrodynamic and Hydromagnetic Stability (Oxford: Clarendon Press).

Field, G. B. 1965, Ap. J., 142, 531.

Parker, E. N. 1953, Ap. J., 117, 431.

Rayleigh, Lord 1916, Phil. Mag., Series 6, 32, 529.

Spiegel, E. A. 1964, Ap. J., 139, 959.

_____. 1965, Ap. J., 141, 1068.

Spiegel, E. A., and Veronis, G. 1960, Ap. J., 131, 442.

Unno, W., and Spiegel, E. A. 1966, Pub. Astr. Soc. Japan, 18, 85.

Vincenti, W. G., and Kruger, C. H. 1965, Introduction to Physical Gas Dynamics (New York: John Wiley and Sons).

Weymann, R. 1960, Ap. J., 132, 452.

PART II

THERMAL INSTABILITY OF A MODEL HYDROGEN PLASMA

ABSTRACT

The thermal stability of a hydrogen plasma is analyzed with the aid of a simple model atom which possesses a continuum of unbound states but only one bound level (the ground state). If ionization is collisional and recombination is radiative, the model plasma is thermally unstable if and only if its kinetic temperature exceeds 17500 ^{o}K. Although this plasma emits only free-bound radiation (the free-free emission is neglected), the instability is caused by the same ionization effects which lead to instability in plasmas emitting line radiation. The instability criterion and the growth rate (if it is not too large) are not affected by the presence of a uniform magnetic field unless the motions are constrained to be perpendicular to the field.

The applicability of steady-state ionization equations to calculations of thermal instability is examined.

I. INTRODUCTION

The first treatment of thermal instability was by Parker (1953). If we define the heat loss function, $\mathcal{L}$, as the energy lost (by radiation) minus the energy gained per gram of matter per second, Parker's criterion for thermal instability is simply

$$\left(\frac{\partial \mathcal{L}}{\partial T}\right)_\rho < 0, \tag{1}$$

where T is the temperature and ρ is the density. Parker applied criterion (1) to an optically thin hydrogen plasma, which emits free-free, free-bound, and bound-bound radiation. Owing to the incompleteness of criterion (1), however, his analysis of the various emission mechanisms was incorrect.

The first complete discussion of the thermal stability of free-free emission was given by Field (1965). Field showed that when density variations are taken into account, the criterion for thermal instability of an ideal gas becomes

$$\left(\frac{\partial \mathcal{L}}{\partial T}\right)_p = \left(\frac{\partial \mathcal{L}}{\partial T}\right)_\rho - \frac{\rho}{T}\left(\frac{\partial \mathcal{L}}{\partial \rho}\right)_T < 0, \tag{2}$$

where P is the pressure. For later reference, we note that if the instability develops sufficiently slowly (so that pressure equilibrium is maintained), the growth rate predicted by Field's analysis is

$$n = -\frac{1}{C_p}\left(\frac{\partial \mathcal{L}}{\partial T}\right)_p, \tag{3}$$

where c_p is the specific heat at constant pressure. Since the rate of free-free emission per unit mass in a fully ionized gas is proportional to $\rho T^{1/2}$, the isobaric criterion (2) shows that a plasma radiating mainly by free-free transitions is thermally unstable whereas the isochoric criterion (1) incorrectly predicts stability.

Although the case of free-free emission from a fully ionized gas is now well understood, no completely satisfactory treatment of thermal instability associated with free-bound or bound-bound emission has yet been given. The shortcomings of the analyses which have appeared are explained in §§IV and V.

In §§II and III the stability of a plasma emitting free-bound radiation is studied with the aid of a simple model hydrogen atom. The model atom, which has been used in other contexts by Curtis (1963) and by Dietz and House (1965), has a continuum of unbound states, but the only bound level is the ground state. The model hydrogen plasma therefore consists of protons, electrons, and hydrogen atoms in the ground state. This plasma can emit both free-free and free-bound radiation, but we suppress the bremsstrahlung in order to focus attention on the recombination radiation.

In §IV the thermal instability of the model plasma is shown to be of a more general nature than the simple atomic model would seem to permit. In fact, the instability is the same in its essential physics as the instability of a plasma emitting line radiation. A separate treatment of thermal instability associated with bound-bound emission is therefore unnecessary, although additional calculations are clearly required for accurate results in any particular case of practical interest.

The standard method of treating systems emitting line radiation is discussed in §V. In §VI we investigate the influence of a magnetic field on the instability of the model plasma.

II. STABILITY ANALYSIS

Let the number of hydrogen atoms, neutral or ionized, per gram be N (Avogadro's number). Let x denote the fraction of atoms which which are ionized, so that in one gram there are Nx protons, Nx electrons, and N(1-x) neutral hydrogen atoms. If T is the kinetic temperature (assumed identical for all species) and χ is the ionization potential of hydrogen, the internal energy per gram of the model plasma is

$$U = N(1+x)\tfrac{3}{2}kT + Nx\chi, \tag{4}$$

where k is Boltzmann's constant. The rate of change of internal energy is therefore

$$\frac{dU}{dt} = N(1+x)\tfrac{3}{2}k\frac{dT}{dt} + N(\tfrac{3}{2}kT+\chi)\frac{dx}{dt}. \tag{5}$$

In this paper thermal conduction is regarded as a complicating factor of minor theoretical significance in spite of the fact that it is sometimes of major importance in practice. The effect of conduction is simply to smooth temperature fluctuations and to stabilize perturbations of the shortest wavelengths. Since the influence of conduction is well known

(Parker 1953; Field 1965), it is neglected in the present analysis although it could easily have been included. The first law of thermodynamics may then be written in the form

$$\frac{dU}{dt} = -\mathcal{L} + \frac{P}{\rho^2}\frac{d\rho}{dt}, \tag{6}$$

where the pressure is given by

$$P = N(1+x)\rho kT. \tag{7}$$

The heat loss function, $\mathcal{L}$, equals the rate of free-bound emission per gram minus the rate of energy input. The mechanism of heat input is unspecified but is assumed to proceed at a constant rate such that $\mathcal{L} = 0$ in the unperturbed state. We assume that the plasma is optically thin to the free-bound radiation so that all recombination photons leave the system. The heat loss function then depends only on the local values of x, ρ, and T. For small departures from the equilibrium state, we have

$$\mathcal{L} = \mathcal{L}_x x_1 + \mathcal{L}_\rho \rho_1 + \mathcal{L}_T T_1, \tag{8}$$

where the subscripts x, ρ, and T denote partial derivatives and x_1, for example, is the departure of the ionization level from its value in the unperturbed state.

The degree of ionization is governed by

$$N\frac{dx}{dt} = \mathcal{J}, \tag{9}$$

where $\mathcal{J}$ (the ionization function) is the number of ionizations minus the number of recombinations per gram per second and is a function of x, ρ, and T. Initially, $\mathcal{J} = 0$, but after the plasma has been perturbed we write

$$\mathcal{J} = \mathcal{J}_x x_1 + \mathcal{J}_\rho \rho_1 + \mathcal{J}_T T_1 , \tag{10}$$

where subscripts have the same meaning as in equation (8).

The density variations are calculated from the continuity equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \underset{\sim}{u}) = 0 , \tag{11}$$

while the fluid velocity $\underset{\sim}{u}$ is governed by the equation of motion

$$\rho \frac{d\underset{\sim}{u}}{dt} + \nabla P = 0 , \tag{12}$$

where we have neglected viscosity and gravitation. Magnetic fields are assumed to be absent but will be included in §VI.

We now consider one-dimensional perturbations in which all quantities depend on the coordinate z and the time t. From equations (5), (6), and (8), we obtain the linearized heat equation

$$\begin{aligned} &N\left(\tfrac{3}{2}kT + \chi\right)\frac{\partial x_1}{\partial t} + \mathcal{L}_x x_1 - \frac{P}{\rho^2}\frac{\partial \rho_1}{\partial t} \\ &+ \mathcal{L}_\rho \rho_1 + N(1+x)\tfrac{3}{2}k\frac{\partial T_1}{\partial t} + \mathcal{L}_T T_1 = 0. \end{aligned} \tag{13}$$

The differential form of the equation of state (7) is

$$\frac{P_1}{P} - \frac{x_1}{1+x} - \frac{\rho_1}{\rho} - \frac{T_1}{T} = 0. \tag{14}$$

From equations (9) and (10) we get the linearized ionization equation

$$N\frac{\partial x_1}{\partial t} - \mathcal{J}_x x_1 - \mathcal{J}_\rho \rho_1 - \mathcal{J}_T T_1 = 0. \tag{15}$$

The linearized forms of equations (11) and (12) are

$$\frac{\partial \rho_1}{\partial t} + \rho \frac{\partial w}{\partial z} = 0 \tag{16}$$

and

$$\frac{\partial P_1}{\partial z} + \rho \frac{\partial w}{\partial t} = 0, \tag{17}$$

where w is the z-component of $\underset{\sim}{u}$.

We now expand each of the perturbation variables in plane waves of the form $\exp(nt + i\kappa z)$. Then equations (13) - (17) become a set of five simultaneous, homogeneous, algebraic equations. The requirement for the existence of a nontrivial solution, that the determinant of the coefficients vanish, leads to the characteristic equation

$$\tfrac{3}{2}Nn^4 + \left\{\mathcal{L}_T + \chi\mathcal{J}_T + \tfrac{3}{2}k\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\right\}\frac{n^3}{k(1+x)}$$
$$+\left[\frac{1}{N(1+x)k}\left(\mathcal{L}_x\mathcal{J}_T - \mathcal{L}_T\mathcal{J}_x\right) + N^2(1+x)\tfrac{5}{2}kTK^2\right]n^2$$
$$+\left\{T\mathcal{L}_T - \rho\mathcal{L}_\rho + \chi\left(T\mathcal{J}_T - \rho\mathcal{J}_\rho\right) + \tfrac{5}{2}kT\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\right\}NK^2n$$
$$+ TK^2\Delta = 0, \qquad (18)$$

where

$$\Delta = \left(\mathcal{L}_x - \frac{\rho}{1+x}\mathcal{L}_\rho\right)\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) - \left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)\left(\mathcal{J}_x - \frac{\rho}{1+x}\mathcal{J}_\rho\right). \qquad (19)$$

The quartic equation (18) has a positive real root when the last term is negative. Therefore, a sufficient condition for instability is

$$\Delta < 0. \qquad (20)$$

When the growth rate of the instability is small, it may be easily computed as follows. First, we note that when $\mathcal{L}$ and $\mathcal{J}$ vanish, the solutions of equation (18) are two imaginary roots, corresponding to isentropic sound waves, and $n = 0$. We are interested in the roots which vanish as $\mathcal{L}$ and $\mathcal{J}$ tend to zero. When $\mathcal{L}$ and $\mathcal{J}$ (and therefore n) are small, these roots are evidently the solutions of the quadratic equation

$$N^2(1+x)\tfrac{5}{2}k\,n^2$$
$$+\left\{\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho + \chi\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) + \tfrac{5}{2}k\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\right\}Nn$$
$$+ \Delta = 0. \qquad (21)$$

We note that pressure equilibrium is maintained if the growth rate is not too large. In fact, if we set $P_1 = 0$ in equation (14) and assume an e^{nt} behavior for the remaining variables, equations (13)-(15) lead to the characteristic equation (21).

III. THE HEAT LOSS AND IONIZATION FUNCTIONS

In the preceding analysis we treated $\mathcal{L}$ and $\mathcal{I}$ as arbitrary functions of x, ρ, and T. We now consider the forms of these functions which are of most interest in practice.

For a hydrogen plasma emitting free-bound radiation, the heat loss function is

$$\mathcal{L} = 3.26 \cdot 10^{-6} N^2 x^2 \rho k T^{-1/2} - \text{const}, \qquad (22)$$

where "const" refers to the unspecified energy input and is such that $\mathcal{L} = 0$ in the unperturbed state. For $T \gtrsim 10^5\ ^\circ K$, a term in $x^2 \rho T^{1/2}$ should be added to equation (22) to account for free-free emission. However, this emission is neglected in the present analysis.

We assume that all ionization is produced by collisions. The number of ionizations per gram per second is then

$$I = 1.23 \cdot 10^{-5} N^2 x(1-x) \rho \frac{k}{\chi} T^{1/2} e^{-\chi/kT}. \qquad (23)$$

The case of photoionization by an external radiation field is treated in the Appendix. Next, we suppose that the density is sufficiently low that three-body recombination is negligible compared to radiative recombination,

for which the rate is

$$R = 3.26 \cdot 10^{-6} N^2 x^2 \rho T^{-3/2} e^{\chi/kT} E_1(\chi/kT), \quad (24)$$

where E_1 represents the first exponential integral. The ionization function used in the stability analysis is

$$\mathcal{J} = I - R, \quad (25)$$

which vanishes in the unperturbed state.

Using equations (22)-(25), we find that the instability condition (20) may be written as follows:

$$2\left(\frac{\chi}{kT} + 1\right) - \frac{e^{-\chi/kT}}{E_1(\chi/kT)} - \frac{3}{2}\left(\frac{1+x}{2+x}\right)\frac{1}{1-x} < 0, \quad (26)$$

where x is to be determined from the equilibrium ionization equation ($\mathcal{J} = 0$):

$$\frac{1-x}{x} = 0.265 \frac{\chi}{kT^2} e^{2\chi/kT} E_1(\chi/kT). \quad (27)$$

When $\chi/kT \gg 1$, the instability condition (26) becomes

$$\frac{3}{2}\left(\frac{1+x}{2+x}\right)\frac{1}{1-x} > \frac{\chi}{kT} + 1, \quad (28)$$

and the ionization equation (27) becomes

$$\frac{1-x}{x} = \frac{0.265}{T} e^{\chi/kT}\left(1 - \frac{kT}{\chi}\right). \quad (29)$$

From relations (28) and (29) it follows that a hydrogen plasma satisfying the assumptions of this analysis is thermally unstable for $T > 17500\ ^{o}K$ (which corresponds to $x > 0.90$).

From a mathematical point of view, condition (20) is sufficient but not necessary for equation (18) to have a root in the right half of the complex n plane. However, it can be shown from the Hurwitz-Routh criterion (cf. Kaplan 1962) and the expressions for $\mathcal{L}$ and $\mathcal{J}$ adopted in this section that inequality (20) is, in fact, also a necessary condition for instability. Thus the model plasma is stable for $T < 17500\ ^{o}K$.

IV. PHYSICAL BASIS OF THE INSTABILITY

In order to analyze the physics of the instability treated in §§II and III, we consider the following form of the energy equation:

$$\frac{dU_{th}}{dt} = -\mathcal{L}_{th} + \frac{P}{\rho^2}\frac{d\rho}{dt} \qquad (30)$$

where U_{th} is the thermal energy per gram and $\mathcal{L}_{th}$ is the loss minus the gain of thermal energy per gram per second. The loss component of $\mathcal{L}_{th}$ for the model plasma consists of two parts. First, each radiative recombination is associated with a loss of thermal energy equal to the kinetic energy of the recombining electron. Second, each collisional ionization reduces the energy in the thermal field by χ. The rates at which thermal energy is lost by these processes are $\mathcal{L}_{fb} - R\chi$ (where $\mathcal{L}_{fb}$ represents the first term on the right side of equation [22]) and $I\chi$, respectively. We therefore have

$$\mathcal{L}_{th} = (\mathcal{L} - R\chi) + I\chi = \mathcal{L} + \mathcal{I}\chi . \qquad (31)$$

Substituting equation (31) into equation (30), we recover equation (6) by virtue of equations (5) and (9).

An error made in previous treatments of thermal instability associated with free-bound emission is the implicit assumption that $\mathcal{L}_{th} = \mathcal{L}$. This assumption is valid in the steady state in which there is an ionizing collision for each radiative recombination so that the total energy of a recombination photon actually represents a loss of thermal energy. When this steady state is disturbed, however, $\mathcal{I} \neq 0$ and therefore $\mathcal{L}_{th} \neq \mathcal{L}$.

In addition to assuming $\mathcal{L}_{th} = \mathcal{L}$, most treatments of free-bound emission consider the degree of ionization, x, to be a constant. Criterion (1) or (2) and equation (22) for $\mathcal{L}$ then predict that a plasma emitting mainly free-bound radiation is always thermally unstable. On the other hand, if we use $\mathcal{L}_{th}$ instead of $\mathcal{L}$ but continue to treat x as a constant, criteria (1) and (2) predict stability. In fact, for $\chi/kT \gg 1$, equations (22)-(25) and (31) yield

$$\frac{\partial}{\partial T}\mathcal{L}_{th} - \frac{\rho}{T}\frac{\partial}{\partial \rho}\mathcal{L}_{th} = kR\left(\frac{\chi}{kT} + 1\right)\left(\frac{\chi}{kT} - \frac{3}{2}\right) > 0. \qquad (32)$$

The stability which appears to be indicated by equation (32) and criterion (2) is easily understood. A temperature rise increases the rate $(I\chi)$ at which electrons lose kinetic energy by inelastic ionizing collisions with hydrogen atoms. Furthermore, although an increase in temperature reduces the recombination rate ($R \sim T^{-1/2}$ for $kT \ll \chi$),

the mean thermal energy lost per recombination (kT) increases, with the net result that the rate of loss of thermal energy via radiative recombination also increases. Thus the $I\chi$ and $\mathcal{L}_{fb} - R\chi$ components of $\mathcal{L}_{th}$ both vary with temperature in a way which promotes thermal stability. Although the density dependence of $\mathcal{L}_{th}$ favors instability, it is not strong enough to counteract the stabilizing temperature dependence of $I\chi$.

If we ignore variations in the degree of ionization, therefore, we are forced to conclude that a plasma emitting mainly free-bound radiation is thermally stable. It follows that the instability of the model hydrogen plasma for $T > 17500\ ^{o}K$ must result from changes in x. To see how the ionization level affects stability, we note that the instability condition (20) can be rewritten in the form

$$\left(\frac{\partial}{\partial x}\mathcal{L}_{th} - \frac{\rho}{1+x}\frac{\partial}{\partial \rho}\mathcal{L}_{th}\right)\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) - \left(\frac{\partial}{\partial T}\mathcal{L}_{th} - \frac{\rho}{T}\frac{\partial}{\partial \rho}\mathcal{L}_{th}\right)\left(\mathcal{J}_x - \frac{\rho}{1+x}\mathcal{J}_\rho\right) < 0. \quad (33)$$

From equations (23)-(25) we see that $\mathcal{J}_x < 0$ while $\mathcal{J}_\rho = 0$. In view of equation (32), the second term in condition (33) promotes stability.

Since $\mathcal{J}_T > 0$, any negative terms in the first factor of condition (33) contribute to instability. It is evident from equation (23) that the derivative of the $I\chi$ component of $\mathcal{L}_{th}$ with respect to x is negative for $x > 0.5$. The mechanism of this destabilizing influence is that the increase in the degree of ionization which follows a rise in temperature (since $\mathcal{J}_T > 0$) means that there are fewer targets for inelastic ionizing collisions so that the rate of cooling via this process must decline ($I_x < 0$).

The second term in the first factor of inequality (33) also promotes instability since $\frac{\partial}{\partial \rho}\mathcal{L}_{th} > 0$. In this case the physical mechanism is that an increased level of ionization is associated with a decrease in density owing to the tendency toward pressure equilibrium. The density decrease reduces the rate of inelastic collisions (both ionizations and recombinations) and therefore the cooling rate.

The ionization effects just described for the model hydrogen plasma are also operative in a plasma radiating mainly in spectral lines. In such a plasma, cooling is effected primarily by the process of collisional excitation, and thermal instability results from variations in the excitation rate caused by changes in the degree of ionization of the target atoms. This type of instability does not differ in principle from the instability of the model plasma, in which collisional ionization not only affects the degree of ionization but also happens to be the inelastic collision process which contributes most of the cooling.

V. THE STEADY-STATE IONIZATION EQUATION

In considerations of thermal instability associated with line radiation, it is customary to calculate "cooling curves" which give the radiation rate as a function of temperature. To compute the emission for a particular value of T, one first solves the steady-state ionization equation ($\mathcal{J} = 0$) to find the concentration of the emitting ion. The heat loss function can then be considered to be a function of ρ and T only since the level of ionization is a known function of T (and possibly ρ).

The use of the steady-state ionization equation requires that the

time required to reach ionization equilibrium be much less than the time scale of the gas motions. However, the macroscopic time scale (i.e., the growth time for the instability) is obviously controlled by the time scale of the "microscopic" processes. Moreover, we have seen in §IV that departures from $\mathcal{J} = 0$ can be critical. It is therefore necessary to examine the use of the steady-state ionization equation in more detail.

We now evaluate the usefulness of growth rates which have been computed with the aid of steady-state ionization equations. Although these growth rates usually have referred to systems emitting line radiation, we shall restrict our considerations to the model hydrogen plasma since, as we saw in §IV, this model contains the essential physics. We now calculate the growth rate for the model plasma assuming both pressure equilibrium and ionization equilibrium. We have already given the growth rate under the assumption of pressure equilibrium alone (equation [21]). Comparison of the two expressions for the growth rate will reveal the consequences of using the steady-state ionization equation.

If we set $P_1 = 0$ in equation (14), the resulting equation can be written in the form

$$\left(\frac{\partial \rho}{\partial T}\right)_p = -\frac{\rho}{T} - \frac{\rho}{1+x}\left(\frac{\partial x}{\partial T}\right)_p . \tag{34}$$

If $\mathcal{J}$ is to vanish identically, we require that

$$\mathcal{J}_T\, dT + \mathcal{J}_\rho\, d\rho + \mathcal{J}_x\, dx = 0. \tag{35}$$

Equations (34) and (35) yield

$$\left(\frac{\partial \rho}{\partial T}\right)_P = -\frac{\rho}{T} + \rho \frac{\mathcal{J}_T - \rho \mathcal{J}_\rho / T}{(1+X)\mathcal{J}_X - \rho \mathcal{J}_\rho} \tag{36}$$

and

$$\left(\frac{\partial X}{\partial T}\right)_P = -\frac{\mathcal{J}_T - \rho \mathcal{J}_\rho / T}{\mathcal{J}_X - \rho \mathcal{J}_\rho / (1+X)} \,. \tag{37}$$

Substituting equations (36) and (37) into

$$\left(\frac{\partial \mathcal{L}}{\partial T}\right)_P = \mathcal{L}_T + \mathcal{L}_\rho \left(\frac{\partial \rho}{\partial T}\right)_P + \mathcal{L}_X \left(\frac{\partial X}{\partial T}\right)_P \,, \tag{38}$$

we find from equation (3) that the growth rate is

$$n = \frac{\Delta}{c_p \left[\mathcal{J}_X - \rho \mathcal{J}_\rho / (1+X)\right]} \,. \tag{39}$$

Since $\mathcal{J}_X < 0$ and $\mathcal{J}_\rho = 0$, this growth rate is positive if and only if the instability condition (20) is satisfied. Furthermore, numerical computation shows that equation (39) reproduces the appropriate root of equation (21) with satisfactory accuracy. We conclude, therefore, that the use of the steady-state ionization equation does not lead to significant errors in calculations of thermal instability. However, this result does not apply when the gas is in a gravitational field (Defouw 1970b).

The discussion of this section has referred primarily to the method ordinarily used to treat thermal instability of a plasma emitting line radiation. As mentioned in §IV, most previous treatments of free-bound emission have assumed that the ionization level remains constant.

Athay and Thomas (1956; see also Thomas and Athay 1961), however, have included the effects of variable ionization by a method equivalent to the one used in this section. Although this part of their work has apparently been overlooked, they also found that a plasma emitting mainly free-bound radiation is thermally unstable only if the kinetic temperature exceeds some critical value.

VI. EFFECT OF A MAGNETIC FIELD

Since most astrophysical plasmas are subject to magnetic and/or gravitational fields, the effect of these fields on thermal instability is of great interest. For the simple case in which ionization effects are negligible, the effects of magnetic and gravitational fields have been investigated by Field (1965) and Defouw (1970a), respectively. However, the effect of these fields on ionization-induced thermal instability remains to be demonstrated. The thermal instability of the model hydrogen plasma in a gravitational field will be analyzed in another paper (Defouw 1970b). In this section we study the stability of the model plasma in an initially uniform magnetic field.

The equation of motion is

$$\rho \frac{d\underset{\sim}{u}}{dt} + \nabla\left(P + \frac{B^2}{8\pi}\right) - \frac{1}{4\pi}\left(\underset{\sim}{B}\cdot\nabla\right)\underset{\sim}{B} = 0, \qquad (40)$$

where the magnetic field $\underset{\sim}{B}$ satisfies

$$\frac{d\underset{\sim}{B}}{dt} + \underset{\sim}{B}\nabla\cdot\underset{\sim}{u} - \left(\underset{\sim}{B}\cdot\nabla\right)\underset{\sim}{u} = 0 \qquad (41)$$

when the resistivity is negligible. We now linearize equations (40) and (41) and expand the perturbation variables in plane waves of the form $\exp(nt + i\,\underset{\sim}{\kappa}\cdot\underset{\sim}{r})$. If $\underset{\sim}{B}$ now represents the uniform initial field and $\underset{\sim}{b}$ is the perturbation of the field, we have

$$\rho n\underset{\sim}{u} + i\kappa\underset{\sim}{\kappa}P_1 + \frac{i}{4\pi}\kappa(\underset{\sim}{B}\cdot\underset{\sim}{b}) - \frac{i}{4\pi}\kappa B\cos\theta\,\underset{\sim}{b} = 0 \qquad (42)$$

and

$$n\underset{\sim}{b} + iB(\underset{\sim}{\kappa}\cdot\underset{\sim}{u}) - i\kappa B\cos\theta\,\underset{\sim}{u} = 0, \qquad (43)$$

where θ is the angle between $\underset{\sim}{\kappa}$ and $\underset{\sim}{B}$. We now take the scalar product of equation (42) with $\underset{\sim}{\kappa}$ to get

$$i\kappa^2 P_1 + \rho n(\underset{\sim}{\kappa}\cdot\underset{\sim}{u}) + i\frac{\kappa^2}{4\pi}(\underset{\sim}{B}\cdot\underset{\sim}{b}) = 0, \qquad (44)$$

where we have used $\underset{\sim}{\kappa}\cdot\underset{\sim}{b} = 0$ (since $\nabla\cdot\underset{\sim}{b} = 0$). Taking the scalar product of both equations (42) and (43) with $\underset{\sim}{B}$, we find

$$i\kappa B\cos\theta\,P_1 + \rho n(\underset{\sim}{B}\cdot\underset{\sim}{u}) = 0 \qquad (45)$$

and

$$iB^2(\underset{\sim}{\kappa}\cdot\underset{\sim}{u}) - i\kappa B\cos\theta\,(\underset{\sim}{B}\cdot\underset{\sim}{u}) + n(\underset{\sim}{B}\cdot\underset{\sim}{b}) = 0. \qquad (46)$$

The continuity equation (11) takes the form

$$n\rho_1 + i\rho(\underset{\sim}{\kappa}\cdot\underset{\sim}{u}) = 0. \tag{47}$$

Equations (13)-(15) and (44)-(47) form a simultaneous set of seven homogeneous algebraic equations (after the e^{nt} time behavior is substituted in equations [13] and [15]) in the seven unknowns x_1, ρ_1, T_1, P_1, $\underset{\sim}{\kappa}\cdot\underset{\sim}{u}$, $\underset{\sim}{B}\cdot\underset{\sim}{u}$, and $\underset{\sim}{B}\cdot\underset{\sim}{b}$. Setting the determinant of the coefficients equal to zero, we obtain the characteristic equation

$$\begin{aligned}
&N^2(1+x)\tfrac{3}{2}kn^6 + \left[\mathcal{L}_T + (\tfrac{3}{2}kT+\chi)\mathcal{I}_T - \tfrac{3}{2}k(1+x)\mathcal{I}_x\right]Nn^5 \\
&+\left[\mathcal{L}_x\mathcal{I}_T - \mathcal{L}_T\mathcal{I}_x + N^2(1+x)\tfrac{3}{2}k\kappa^2(v_A^2+a_s^2)\right]n^4 \\
&+\left\{v_A^2\left[\mathcal{L}_T + (\tfrac{3}{2}kT+\chi)\mathcal{I}_T - \tfrac{3}{2}k(1+x)\mathcal{I}_x\right]\right. \\
&\left.+\,a_T^2\left[\mathcal{L}_T - \tfrac{\rho}{T}\mathcal{L}_\rho + \chi(\mathcal{I}_T - \tfrac{\rho}{T}\mathcal{I}_\rho) + \tfrac{5}{2}kT\mathcal{I}_T - \tfrac{5}{2}k(1+x)\mathcal{I}_x\right]\right\}N\kappa^2n^3 \\
&+\left\{v_A^2\left[\mathcal{L}_x\mathcal{I}_T - \mathcal{L}_T\mathcal{I}_x + N^2(1+x)\tfrac{5}{2}k\,a_T^2\kappa^2\cos^2\theta\right] + a_T^2\Delta\right\}\kappa^2n^2 \\
&+\left\{\mathcal{L}_T - \tfrac{\rho}{T}\mathcal{L}_\rho + \chi(\mathcal{I}_T - \tfrac{\rho}{T}\mathcal{I}_\rho) + \tfrac{5}{2}k\left[T\mathcal{I}_T - (1+x)\mathcal{I}_x\right]\right\}Na_T^2v_A^2\kappa^4\cos^2\theta\, n \\
&+\,a_T^2v_A^2\kappa^4\cos^2\theta\,\Delta = 0,
\end{aligned} \tag{48}$$

where $v_A = B/(4\pi\rho)^{1/2}$ is the Alfvén speed, $a_T = (P/\rho)^{1/2}$ is the isothermal speed of sound, and $a_s = (5P/3\rho)^{1/2}$ is the isentropic sound speed.

Equation (48) has a positive real root if the last term is negative.

Hence inequality (20) is a sufficient condition for instability in a magnetic field provided $\cos\theta \neq 0$.

If $\mathcal{L}$ and $\mathcal{J}$ vanish identically, equation (48) becomes

$$n^2\left[n^4+(v_A^2+a_S^2)\kappa^2 n^2+a_S^2 v_A^2\kappa^4\cos^2\theta\right]=0. \quad (49)$$

The zeros of the quartic correspond to magnetohydrodynamic waves. However, we are interested in the roots which vanish for $\mathcal{L}$ and $\mathcal{J}$ equal to zero. These roots are of first order in $\mathcal{L}$ and $\mathcal{J}$, and it is easily verified from equation (48) that they satisfy equation (21) when $\mathcal{L}$ and $\mathcal{J}$ are small and $\cos\theta \neq 0$. We conclude, therefore, that the instability condition and also the growth rate of the instability (if it is not too large) are unaffected by the presence of a magnetic field provided the motions are not constrained to be exactly perpendicular to the field. Field (1965) also obtained this result and offered an explanation which applies equally well to the present calculation.

We now consider the special case $\theta = \pi/2$. When $\cos\theta = 0$, equation (48) becomes a quartic, the constant term of which is negative when

$$\left[\mathcal{L}_X-\frac{\rho}{1+X}(1+v_A^2/a_T^2)^{-1}\mathcal{L}_\rho\right]\left[\mathcal{J}_T-\frac{\rho}{T}(1+v_A^2/a_T^2)^{-1}\mathcal{J}_\rho\right]$$
$$-\left[\mathcal{L}_T-\frac{\rho}{T}(1+v_A^2/a_T^2)^{-1}\mathcal{L}_\rho\right]\left[\mathcal{J}_X-\frac{\rho}{1+X}(1+v_A^2/a_T^2)^{-1}\mathcal{J}_\rho\right]<0. \quad (50)$$

Apart from the factor $(1+v_A^2/a_T^2)^{-1}$, the instability condition (50) is identical to inequality (20). It can be shown (Field 1965) that this factor simply takes account of the magnetic contribution to the pressure. For

the model plasma, $\mathcal{L}_\rho > 0$, $\mathcal{J}_T > 0$, $\mathcal{J}_x < 0$, and $\mathcal{J}_\rho = 0$. With these relations, we see from inequality (50) that the magnetic field has a stabilizing influence on perturbations with $\theta = \pi/2$.

VII. CONCLUSION

There are two basic types of thermal instability due to variations in cooling rate. First, as shown by Field (1965), a fully ionized gas emitting bremsstrahlung is thermally unstable because a temperature increase (say) is accompanied by a density decrease and, therefore, by a reduction in the rate of free-free transitions. Second, a partially ionized plasma, whether it radiates primarily in spectral lines or free-bound continua, can be thermally unstable if the ionization of the atoms which act as targets for inelastic collisions is primarily collisional and therefore a rapidly increasing function of temperature. In this paper we have given a detailed analysis of the simplest possible case of this ionization-induced type of thermal instability.

We found that the model hydrogen plasma is unstable for $T > 17500\ ^{\circ}K$. The precision with which this value has been quoted does not imply that it is an accurate determination of the critical temperature for a real hydrogen plasma. Athay (Thomas and Athay 1961, p. 163) concluded that the solar chromosphere becomes thermally unstable at a temperature somewhere in the range 12000 to 14000 $^{\circ}K$. These values are certainly more realistic than the one found for the simple model plasma because they take account, in a crude fashion, of the effects of excited bound states. The treatment of this paper could, of course, be extended to include excited atomic levels.

However, the difficulty of simultaneously considering the time-dependent diffusion of the line photons makes such an extension a rather ambitious project.

I am indebted to Peter Goldreich for a number of discussions on this subject. I wish to thank both Dr. Goldreich and Dr. E. A. Spiegel for reading a draft of this paper.

APPENDIX

IONIZATION BY AN EXTERNAL RADIATION FIELD

In a dilute plasma subjected to an intense ultraviolet radiation field, photoionization dominates collisional ionization and the ionization rate is

$$I = N(1-x)C, \tag{A1}$$

where C is a constant determined by the given radiation field and the atomic absorption coefficient. If $\mathcal{E}$ represents the mean energy of photons absorbed by the plasma, equation (22) must be replaced by

$$\mathcal{L} = 3.26\cdot10^{-6}N^2x^2\rho kT^{-1/2} - N(1-x)C\mathcal{E} - \text{const}, \tag{A2}$$

where "const" refers to nonradiative heat input.

Using equations (A1), (A2), (19), (24), and (25), we find that the instability condition (20) takes the form

$$\mathcal{E} - \chi < kT \tag{A3}$$

for $\chi/kT \gg 1$. Thus the plasma is thermally unstable only if the mean energy of a photoelectron is less than the mean kinetic energy of a recombining electron.

REFERENCES

Athay, R. G., and Thomas, R. N. 1956, Ap. J., 123, 299.

Curtis, G. W. 1963, Thesis, University of Colorado.

Defouw, R. J. 1970a, submitted to Ap. J.

_____. 1970b, in preparation.

Dietz, R. D., and House, L. L. 1965, Ap. J., 141, 1393.

Field, G. B. 1965, Ap. J., 142, 531.

Kaplan, W. 1962, Operational Methods for Linear Systems (Addison-Wesley).

Parker, E. N. 1953, Ap. J., 117, 431.

Thomas, R. N., and Athay, R. G. 1961, Physics of the Solar Chromosphere (New York: Interscience).

PART III

THE ORIGIN OF SOLAR SPICULES AND SOME RELATED PHENOMENA

Abstract. The convective stability of a simple model chromosphere is investigated. The model chromosphere consists of protons, electrons, and hydrogen atoms in the ground state; ionization is collisional and recombination is radiative. The analysis indicates stability when the kinetic temperature (T) is less than 17500 K (assuming T increases with height). However, for T > 17500 K, the model chromosphere is overstable in the absence of magnetic fields provided the temperature inversion is sufficiently steep. For smaller values of the temperature gradient, field-free regions are stable if the density is small and monotonically unstable if it is large. In the presence of a magnetic field,the model chromosphere is monotonically unstable for T > 17500 K, regardless of the temperature gradient.

The convective instability of the model chromosphere results from the fact that the plasma is thermally unstable for T > 17500 K. Thermally unstable regions of the solar atmosphere, although not represented in detail by the model, should behave in a similar fashion.

Field-free regions of the solar chromosphere are probably not monotonically unstable, but overstability is possible and may explain the origin of chromospheric oscillations with periods less than 200 sec. It is suggested that spicules result from the monotonic instability of magnetic regions. A similar instability in the corona may be responsible for the large Doppler spreading of radar echoes.

Elementary considerations of thermal balance predict that the temperature gradient should diverge at levels of marginal stability. The chromospheric region of spicule formation and the corona should therefore both be bounded below by abrupt temperature jumps.

1. Introduction

According to the Schwarzschild criterion, the solar chromosphere is extremely stable against convection. For this reason, a number of authors have suggested that chromospheric spicules are caused by photospheric disturbances. However, it will be shown in this paper that the Schwarzschild criterion does not apply to the outer solar atmosphere and that spicules may, in fact, result from convective instability of the chromosphere.

Standard convection theory treats the atoms of a gas as mass points with no internal structure. But the internal structure of atoms is fundamental to the energy budget of the chromosphere (since cooling is effected by inelastic atomic collisions) and therefore should be included in calculations of thermal convection.

In this paper the convective stability of a partially ionized hydrogen atmosphere is analyzed. The effects of atomic structure are explored with the aid of a simple model hydrogen atom in which the ground state is the only bound level. Ionization is caused by atom-electron collisions, and recombination is radiative. Furthermore, the chromospheric regions of interest are assumed to be optically thin in the Lyman continuum so that all recombination photons escape. I do not claim that all these assumptions are strictly applicable to the solar chromosphere. However, the physical processes are believed to be sufficiently representative that meaningful results can be obtained (see Section 4).

The stability analysis is performed in Sections 2 and 3.

Section 5 is devoted to elementary considerations of thermal balance. The predictions of the theory are compared with observations of the chromosphere in Section 6, and a brief discussion of the corona is given in Section 7. Throughout this paper the ordinary type of instability, in which a perturbation increases monotonically with time, is called monotonic instability in order to distinguish it from overstability.

2. Stability Analysis

Let the number of hydrogen atoms (neutral or ionized) per gram be N and let x denote the fraction of atoms which are ionized. In one gram of plasma there are then Nx protons, Nx electrons, and $N(1-x)$ neutral hydrogen atoms. If T is the plasma kinetic temperature (assumed identical for all species) and χ is the ionization potential of hydrogen, the internal energy per gram is

$$U = N(1+x)\tfrac{3}{2}kT + Nx\chi , \qquad (1)$$

where k is Boltzmann's constant.

Now let $\mathcal{L}$ denote the heat loss function, the energy lost minus the energy gained per gram per second. In the present analysis, $\mathcal{L}$ equals the rate of free-bound emission per gram minus the rate of energy input due to dissipation of mechanical radiation from the photosphere. Since the theory of chromospheric heating is not very trustworthy at the present time, it will be assumed that the energy input (per gram) proceeds at a constant

rate at any given point. The heat loss function is therefore

$$\mathcal{L} = 3.26 \times 10^{-6} N^2 x^2 \rho k T^{-1/2} - \text{const}, \qquad (2)$$

where ρ is the density. The "const" in Equation (2), representing the input, is such that $\mathcal{L} = 0$ in the unperturbed state. The perturbation in $\mathcal{L}$ results only from variations in the recombination radiation and is given to first order by

$$\mathcal{L} = \mathcal{L}_x x_1 + \mathcal{L}_\rho \rho_1 + \mathcal{L}_T T_1 , \qquad (3)$$

where the subscripts x, ρ, and T denote partial derivatives and x_1, for example, is the departure of the ionization level at a fixed point from its value in equilibrium.

The time variation of the internal energy is related to the heat loss function by the first law of thermodynamics

$$\frac{dU}{dt} = -\mathcal{L} + \frac{P}{\rho^2}\frac{d\rho}{dt} , \qquad (4)$$

where the pressure is given by

$$P = N(1+x)\rho k T . \qquad (5)$$

Note that thermal conduction has been neglected in Equation (4). Conduction could easily have been included, of course, but its effects have already been evaluated for a related problem in paper I (Defouw,

1970a).

From Equations (1), (3), and (4) we obtain the linearized heat equation

$$N\left(\tfrac{3}{2}kT+\chi\right)\frac{\partial x_1}{\partial t} + \mathcal{L}_x x_1 - \frac{P}{\rho^2}\frac{\partial \rho_1}{\partial t} + \mathcal{L}_\rho \rho_1$$
$$+ N(1+x)\tfrac{3}{2}k\frac{\partial T_1}{\partial t} + \mathcal{L}_T T_1 \qquad (6)$$
$$+\left[N(1+x)\tfrac{3}{2}k\beta + N\left(\tfrac{3}{2}kT+\chi\right)\alpha - \frac{P}{\rho^2}\frac{d\rho}{dz}\right]w = 0,$$

where w is the vertical (z) component of the fluid velocity, $\beta = dT/dz$ is the temperature gradient, and $\alpha = dx/dz$ is the ionization gradient of the undisturbed chromosphere. Note that the physical variables are assumed to be independent of the horizontal coordinates in the unperturbed chromosphere.

The ionization equation may be written in the general form

$$N\frac{dx}{dt} = \mathcal{J}, \qquad (7)$$

where $\mathcal{J}$ (the ionization function) is the number of ionizations minus the number of recombinations per gram per second. For collisional ionization and radiative recombination, we have

$$\mathcal{J} = 1.23\times10^{-5}N^2x(1-x)\rho\frac{k}{\chi}T^{1/2}e^{-\chi/kT}$$
$$- 3.26\times10^{-6}N^2x^2\rho T^{-3/2}e^{\chi/kT}E_1(\chi/kT), \qquad (8)$$

where E_1 represents the first exponential integral. In the unperturbed state, $\mathcal{J} = 0$. For small perturbations we expand $\mathcal{J}$ in a Taylor series so that, to first order,

$$\mathcal{J} = \mathcal{J}_x X_1 + \mathcal{J}_\rho \rho_1 + \mathcal{J}_T T_1 , \tag{9}$$

where the conventions are the same as in Equation (3). From Equations (7) and (9) we obtain the linearized ionization equation

$$N \frac{\partial X_1}{\partial t} + N \alpha w = \mathcal{J}_x X_1 + \mathcal{J}_\rho \rho_1 + \mathcal{J}_T T_1 . \tag{10}$$

If the perturbation evolves sufficiently slowly, pressure equilibrium will be maintained and the pressure perturbation P_1 can be neglected (see Defouw, 1970b, hereafter called paper II). The differential form of Equation (5) then yields the modified Boussinesq equation of state:

$$\rho_1 = - \frac{\rho}{T} T_1 - \frac{\rho}{1+X} X_1 . \tag{11}$$

From the exact differential form of Equation (5) applied to the undisturbed atmosphere and the equation of hydrostatic equilibrium, we get

$$\frac{1}{\rho} \frac{d\rho}{dz} = - \frac{\rho g}{P} - \frac{\beta}{T} - \frac{\alpha}{1+X} , \tag{12}$$

where g is the gravity. The requirement that the ionization function

vanish at all heights in the undisturbed state yields the relation

$$\mathcal{J}_x \alpha + \mathcal{J}_\rho \frac{d\rho}{dz} + \mathcal{J}_T \beta = 0. \tag{13}$$

Substituting Equation (12) into Equation (13), we find

$$\alpha = \frac{(1+x)\rho^2 g \mathcal{J}_\rho}{P[(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho]} - \frac{T\mathcal{J}_T - \rho\mathcal{J}_\rho}{(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho} \frac{(1+x)\beta}{T}, \tag{14}$$

while substitution of this equation into Equation (12) yields

$$\frac{1}{\rho}\frac{d\rho}{dz} = \frac{-1}{(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho}\left\{\left[(1+x)\mathcal{J}_x - T\mathcal{J}_T\right]\frac{\beta}{T} + \frac{\rho g}{P}(1+x)\mathcal{J}_x\right\}. \tag{15}$$

With the aid of Equations (11), (14), and (15), we may rewrite the energy equation (6) in the form

$$\begin{aligned} &N\left(\tfrac{5}{2}kT + \chi\right)\frac{\partial x_1}{\partial t} + \left(\mathcal{L}_x - \frac{\rho}{1+x}\mathcal{L}_\rho\right)x_1 \\ &+ N(1+x)\tfrac{5}{2}k\frac{\partial T_1}{\partial t} + \left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)T_1 \\ &+ \left\{g\left[\mathcal{J}_x + N\left(\tfrac{3}{2}kT + \chi\right)\frac{\rho^2}{P}\mathcal{J}_\rho\right]\right. \\ &\left. - N\beta\left[\chi\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) + \tfrac{5}{2}kT\mathcal{J}_T - \tfrac{5}{2}k(1+x)\mathcal{J}_x\right]\right\}\frac{(1+x)w}{(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho} = 0. \end{aligned} \tag{16}$$

After substitution of Equations (11) and (14), the ionization equation (10) becomes

$$N\frac{\partial x_1}{\partial t} - \left(\mathcal{I}_x - \frac{\rho}{1+x}\mathcal{I}_\rho\right)x_1 - \left(\mathcal{I}_T - \frac{\rho}{T}\mathcal{I}_\rho\right)T_1 \tag{17}$$

$$+\left[\frac{\rho^2 g}{P}\mathcal{I}_\rho - \beta\left(\mathcal{I}_T - \frac{\rho}{T}\mathcal{I}_\rho\right)\right]\frac{N(1+x)w}{(1+x)\mathcal{I}_x - \rho\mathcal{I}_\rho} = 0.$$

In Equations (16) and (17) the vertical ionization gradient α has been expressed in terms of the temperature gradient β by means of Equation (14). In principle, β could also be eliminated with the aid of the condition, analogous to Equation (13), that the heat loss function vanish at each height in the equilibrium state. However, we will see in Section 5 that factors not included in the stability analysis, such as conduction and the height variation of the heat input, can be critical in determining the value of β. Hence it is preferable to treat β as a free parameter.

We now derive a Boussinesq equation of motion appropriate to the present problem. In the absence of a magnetic field, the inviscid equation of motion is

$$\rho\frac{d\underset{\sim}{u}}{dt} = -\nabla P + \rho\underset{\sim}{g}, \tag{18}$$

where $\underset{\sim}{u}$ is the fluid velocity and $\underset{\sim}{g} = (0, 0, -g)$. We replace ρ by some representative value of the density, ρ_o, except in the gravity term, where we let $\rho = \rho_o + \rho_1$. Then, after linearization and substitution of Equation (11), Equation (18) becomes

$$\frac{\partial \underset{\sim}{u}}{\partial t} = -\nabla\left(\frac{P}{\rho_o}\right) + \underset{\sim}{g}\left(1 - \frac{T_1}{T} - \frac{x_1}{1+x}\right). \tag{19}$$

We now apply the curl operator to Equation (19) twice in succession, making use of the identity

$$\nabla \times (\nabla \times \underset{\sim}{A}) = \nabla(\nabla \cdot \underset{\sim}{A}) - \nabla^2 \underset{\sim}{A} \tag{20}$$

and the Boussinesq continuity equation

$$\nabla \cdot \underset{\sim}{u} = 0. \tag{21}$$

The vertical component of the resulting equation is

$$\frac{\partial}{\partial t} \nabla^2 w = \frac{g}{T} \nabla_h^2 T_1 + \frac{g}{1+x} \nabla_h^2 x_1 \,) \tag{22}$$

where

$$\nabla_h^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \tag{23}$$

is the horizontal Laplacian operator (Equations [23], [25], and [28] are the only equations in this paper where x represents one of the horizontal coordinates rather than the level of ionization).

Equations (16), (17), and (22) are to be solved in conjunction with the conditions that

$$\left.\begin{aligned} x_1 &= 0 \\ T_1 &= 0 \\ w &= 0 \end{aligned}\right\} \tag{24}$$

at both the upper and lower boundaries of the infinite horizontal layer of plasma under consideration.

Equations (16), (17), and (22) and the boundary conditions (24) are satisfied if x_1, T_1, and w each vary as

$$e^{nt}\, e^{i(\kappa_x x + \kappa_y y)} \sin \kappa_z z \,, \tag{25}$$

where κ_z is an integral multiple of π divided by the vertical extent of the chromospheric layer. In fact, when expression (25) is used in Equations (16), (17), and (22), these equations become a set of three homogeneous, simultaneous, algebraic equations. The requirement for the existence of a nontrivial solution, that the determinant of the coefficients vanish, leads to the characteristic equation

$$\begin{aligned}
&N^2(1+x)\tfrac{5}{2}kn^3 + \left\{\mathcal{L}_T - \tfrac{\rho}{T}\mathcal{L}_\rho + \chi\left(\mathcal{J}_T - \tfrac{\rho}{T}\mathcal{J}_\rho\right) + \tfrac{5}{2}k\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\right\} Nn^2 \\
&+ \left\{ \Delta + \frac{N\Gamma g\chi\left(T\mathcal{J}_T + \tfrac{3}{2}\rho\mathcal{J}_\rho\right)}{\chi\left(T\mathcal{J}_T - \rho\mathcal{J}_\rho\right) + \tfrac{5}{2}kT\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]} \right. \\
&\qquad \left. - \tfrac{5}{2}kN^2\Gamma(1+x)(\beta - \beta_{ad})\left[\frac{T\mathcal{J}_T - (1+x)\mathcal{J}_x}{(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho}\right] \right\} n \\
&+ \frac{\Gamma g}{1+x}\, \frac{\left(T\mathcal{J}_T + \tfrac{3}{2}\rho\mathcal{J}_\rho\right)\left[(1+x)\mathcal{L}_x - T\mathcal{L}_T\right]}{\chi\left(T\mathcal{J}_T - \rho\mathcal{J}_\rho\right) + \tfrac{5}{2}kT\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]} \\
&+ \frac{N\Gamma(\beta - \beta_{ad})}{(1+x)\mathcal{J}_x - \rho\mathcal{J}_\rho}\left[\left[(1+x)\mathcal{L}_x - T\mathcal{L}_T\right]\left(\mathcal{J}_T - \tfrac{\rho}{T}\mathcal{J}_\rho\right)\right. \\
&\left. - \left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\left\{\chi\left(\mathcal{J}_T - \tfrac{\rho}{T}\mathcal{J}_\rho\right) + \tfrac{5}{2}k\left[T\mathcal{J}_T - (1+x)\mathcal{J}_x\right]\right\}\right] = 0 \,,
\end{aligned} \tag{26}$$

where

$$\Delta = \left(\mathcal{L}_x - \frac{\rho}{1+x}\mathcal{L}_\rho\right)\left(\mathcal{J}_T - \frac{\rho}{T}\mathcal{J}_\rho\right) - \left(\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho\right)\left(\mathcal{J}_x - \frac{\rho}{1+x}\mathcal{J}_\rho\right), \tag{27}$$

$$\Gamma = \frac{g}{T}\,\frac{K_x^2 + K_y^2}{K_x^2 + K_y^2 + K_z^2}\,, \tag{28}$$

and

$$\beta_{ad} = \frac{-g}{N(1+x)k}\left\{\frac{kT(1+x)\mathcal{J}_x + (\chi + \frac{3}{2}kT)\rho\mathcal{J}_\rho}{\frac{5}{2}kT\left[(1+x)\mathcal{J}_x - T\mathcal{J}_T\right] - \chi\left(T\mathcal{J}_T - \rho\mathcal{J}_\rho\right)}\right\} \tag{29}$$

is the adiabatic lapse rate of the partially ionized atmosphere.

Let a_j be the coefficient of n^j in Equation (26). This equation will have a positive real root if $a_o < 0$. When $\mathcal{L} = 0$, a_o has the same sign as $\beta - \beta_{ad}$ since $\mathcal{J}_T > 0$, $\mathcal{J}_x < 0$, and $\mathcal{J}_\rho = 0$ (see Equation [8]). Hence for the case of adiabatic motion we have instability when $\beta < \beta_{ad}$, which is the Schwarzschild criterion for convection applied to an ionization zone.

We restrict our further considerations to the chromospheric case, in which $\beta > 0$ so that $\beta - \beta_{ad} > 0$ and $a_o > 0$ (even for $\mathcal{L} \neq 0$). According to the Hurwitz-Routh criterion (cf. Kaplan, 1962), Equation (26) then has a root in the right half of the complex n plane if and only if

$$a_1 a_2 - a_o a_3 < 0 \tag{30}$$

and/or

$$a_1 < 0 . \tag{31}$$

Substituting the coefficients from Equation (26) and using Equations (2) and (8), we find that inequalities (30) and (31) are equivalent to

$$c_1 \rho^2 + c_2(\beta - \beta_{ad}) > 1 \tag{32}$$

and

$$c_3 \rho^2 - c_4(\beta - \beta_{ad}) > 1 , \tag{33}$$

respectively. The c_j are functions of g, T, and $(\kappa_x^2 + \kappa_y^2)/\kappa_z^2$ only. The signs of c_1, c_2, and c_3 are all opposite that of the quantity Δ defined by Equation (27); $c_4 > 0$ always. It can be shown that c_1 and c_3 are very nearly equal so that condition (32) is satisfied whenever condition (33) is. Therefore, inequality (32) is a necessary and sufficient condition for instability of the model chromosphere in the absence of magnetic fields.

In view of the signs of c_1 and c_2, a necessary condition for instability is

$$\Delta < 0. \tag{34}$$

From Equations (2) and (8) it can be shown that inequality (34) is satisfied if and only if $T > 17500$ K (paper II). We conclude that regions of the model chromosphere with $T < 17500$ K are stable

while regions with $T > 17500$ K are unstable if the temperature gradient or the density is large enough to satisfy inequality (32).

Having shown that Equation (26) has a root with positive real part when inequality (32) holds, we now investigate the imaginary part of this root. According to Weiss (1964), the unstable root indicated by inequality (30) has a nonzero imaginary part if $a_0 a_2 > 0$ and $a_1 > 0$. The first of these conditions is satisfied when $\beta - \beta_{ad} > 0$, and the second is fulfilled if and only if

$$c_3 \rho^2 - c_4(\beta - \beta_{ad}) < 1 . \tag{35}$$

In a sufficiently steep temperature inversion, inequalities (32) and (35) are both satisfied (we are assuming $\Delta < 0$), and the unstable root indicated by the former is complex, corresponding to overstability. However, if the temperature gradient is not too large and the density is high enough, inequality (33) holds and monotonic instability, rather than overstability, results.

3. The Effect of a Magnetic Field

In chromospheric regions permeated by a magnetic field $\underset{\sim}{B}$, the equation of motion is

$$\rho \frac{d\underset{\sim}{u}}{dt} = -\nabla\left(P + \frac{B^2}{8\pi}\right) + \frac{1}{4\pi}(\underset{\sim}{B}\cdot\nabla)\underset{\sim}{B} + \rho \underset{\sim}{g} . \tag{36}$$

As in the nonmagnetic case, we replace ρ in this equation by a

constant, ρ_o, except in the gravity term, where we let $\rho = \rho_o + \rho_1$. Furthermore, we shall use Equation (11) for ρ_1 since it can be shown that the magnetic field does not affect the pressure balance provided the growth rate is small and the motion is not exactly perpendicular to the field (see papers I and II). The linearized form of Equation (36) is then

$$\frac{\partial \underset{\sim}{u}}{\partial t} = -\nabla\left(\frac{P}{\rho_o} + \frac{\underset{\sim}{B}\cdot\underset{\sim}{b}}{4\pi\rho_o}\right) + \frac{1}{4\pi\rho_o}(\underset{\sim}{B}\cdot\nabla)\underset{\sim}{b} + \underset{\sim}{g}\left(1 - \frac{T_1}{T} - \frac{X_1}{1+X}\right), \quad (37)$$

where $\underset{\sim}{B}$ represents the uniform field of the undisturbed state and $\underset{\sim}{b}$ is the perturbation of this field. We now apply the curl operator to Equation (37) twice in succession and use Equations (20) and (21) as well as $\nabla \cdot \underset{\sim}{b} = 0$. For the case of a vertical initial field, the z-component of the resulting equation is

$$\frac{\partial}{\partial t}\nabla^2 w = \frac{B}{4\pi\rho_o}\frac{\partial}{\partial z}\nabla^2 b_z + \frac{g}{T}\nabla_h^2 T_1 + \frac{g}{1+X}\nabla_h^2 X_1 . \quad (38)$$

The time development of the field, for negligible resistivity, is governed by the equation

$$\frac{d\underset{\sim}{B}}{dt} + \underset{\sim}{B}\nabla\cdot\underset{\sim}{u} - (\underset{\sim}{B}\cdot\nabla)\underset{\sim}{u} = 0. \quad (39)$$

In view of Equation (21), the linearized z-component of Equation (39) is

$$\frac{\partial b_z}{\partial t} = B\frac{\partial w}{\partial z} . \quad (40)$$

Now we differentiate Equation (38) with respect to time and then substitute Equation (40) to get

$$\left(\frac{\partial^2}{\partial t^2} - v_A^2 \frac{\partial^2}{\partial z^2}\right)\nabla^2 w = \frac{g}{T}\frac{\partial}{\partial t}\nabla_h^2 T_1 + \frac{g}{1+X}\frac{\partial}{\partial t}\nabla_h^2 X_1, \quad (41)$$

where $v_A = B/(4\pi\rho_o)^{1/2}$ is the Alfvén speed.

The heat equation (16) and the ionization equation (17) are not affected by the magnetic field. However, we must add to the boundary conditions (24) the free-surface condition that

$$\frac{\partial^2 w}{\partial z^2} = 0 \quad (42)$$

at both the upper and lower boundaries.

Equations (16), (17), and (41) and the boundary conditions (24) and (42) are satisfied if the perturbation variables behave as in expression (25). Substitution of this expression reduces the three partial differential equations to a set of homogeneous algebraic equations. The secular equation is

$$a_3 n^4 + a_2 n^3 + (a_1 + v_A^2 k_z^2 a_3) n^2 + (a_0 + v_A^2 k_z^2 a_2) n + v_A^2 k_z^2 \Delta = 0, \quad (43)$$

where the a_j, as defined above, are the coefficients of n^j in Equation (26).

Equation (43) has a positive real root if the last term is negative. Therefore, inequality (34) is a sufficient condition for monotonic instability in the presence of a magnetic field. For sufficiently large magnetic fields, the growth rate of the instability is given by the appropriate solution of

$$a_3 n^2 + a_2 n + \Delta = 0. \qquad (44)$$

Since a_2 and a_3 are both positive, it follows from Equation (44) that in a strong magnetic field condition (34) is sufficient and necessary for monotonic instability whereas overstability is not possible.

4. Discussion

The physical basis of the instability treated in the preceding sections may be conveniently divided into two parts. First, one can show that a uniform partially ionized gas composed of the simple model hydrogen atoms is thermally unstable when inequality (34) is satisfied (paper II). In fact, the growth rate of the thermal instability is given by Equation (44) if it is not so large that pressure equilibrium is destroyed. Second, one can then introduce a gravitational field and examine the effect of thermal instability on buoyancy forces. In paper I I used the standard description of thermal instability (Field, 1965) to demonstrate that a thermally unstable gas in a gravitational field will be overstable if there is a sufficiently steep temperature

inversion. I also showed that monotonic instability results if the thermally unstable plasma is subject to both gravitational and magnetic fields (regardless of the temperature gradient). We have seen that these results also apply to the model chromosphere.

Evidently the conclusions of Sections 2 and 3 may be understood in terms of the findings of papers I and II. However, it would be a mistake to assume that all the results of the simplified treatment of paper I apply to the model chromosphere. In particular, whereas it was shown in paper I that a thermally unstable gas is always unstable in a gravitational field, we see from inequality (32) that field-free regions of the model chromosphere are stable for sufficiently low values of ρ and $\beta-\beta_{ad}$.

In paper I the instability of a thermally unstable gas in a gravitational field (with or without a magnetic field) was called thermal-convective instability. This term therefore applies to the instability treated above, although I have used the term convective instability since the analysis of this paper can be viewed formally as an extension of Rayleigh's original analysis of convection (Rayleigh, 1916) to include the effects of atomic structure which are important at astrophysical temperatures.

Let us now consider the applicability of the results for the model chromosphere to the chromosphere of the sun. In paper II it was shown that the thermal instability of the model hydrogen plasma results from ionization effects. Owing mainly to the temperature dependence of collisional ionization, an increase in temperature is accompanied by a reduction in the concentration of neutral atoms.

There are then fewer targets for inelastic collisions and the cooling rate declines. In addition, an increased level of ionization tends to be associated with a decrease in density (because of the tendency toward pressure equilibrium), which also leads to a reduction in the cooling rate. These ionization effects obviously can also cause thermal instability when atoms other than hydrogen are responsible for the cooling (see Section 7). However, the critical temperature for thermal instability is then an order of magnitude larger than the hydrogenic value. Evidently, the physical ingredients required for thermal-convective instability in the solar chromosphere are that (1) cooling is effected by neutral hydrogen atoms and (2) the level of hydrogen ionization is a rapidly increasing function of temperature.

According to Athay (1966b), hydrogen is responsible for most of the energy loss from the solar chromosphere above a height of 500 km. Above 1000 km, the region of most interest in this paper (see Section 6), the principal radiation loss is in the Lyman α line. Although the optical thickness in this line is often considerable, it can be shown that the chromospheric regions of interest are effectively thin in that Lyman α photons generally escape (after many scatterings) without being reconverted to thermal energy.

Pottasch (1965) showed that radiative cooling in an optically thin atmosphere is due mainly to elements other than hydrogen for $T \gtrsim 25000$ K. Athay's conclusion that hydrogen is the principal coolant in the solar chromosphere even for $T > 25000$ K is based on the importance of opacity effects in the metallic lines. However, in the chromospheric model proposed by Zirin and Dietz (1963), inter-

spicule regions with $T > 7000$ K have coronal densities, at which the opacity effects cited by Athay are negligible. It is possible, therefore, that in interspicule regions the stability analysis of this paper applies only for $T < 25000$ K. Hence subsequent numerical evaluations of the stability conditions will be restricted to $T = 20000$ K, at which hydrogen dominates even in an optically thin atmosphere.

The second requirement for thermal-convective instability, that hydrogen ionization increase rapidly with temperature, is probably fulfilled in the solar chromosphere. For example, an important ionization process may consist of collisional excitation followed by photoionization by the photospheric radiation field. Such a process would retain most of the temperature sensitivity of collisional ionization since the energy of the first excited state of hydrogen is comparable to the ionization potential. Now suppose that the dominant ionization mechanism is photo-excitation by absorption of a Lyman line photon followed by photoionization by the photospheric continuum. In this case the ionization level rises when Lyman line photons become more numerous as a result of increased collisional excitation, i.e., increased temperature. On the other hand, when the optical depth in the Lyman continuum is sufficiently large, absorption of Lyman continuum photons is an important ionization process which does not have the temperature dependence required for thermal instability. However, this process is not important in the higher, largely ionized chromospheric levels with which we are concerned (Thomas and Athay, 1961, p. 162). I therefore submit that the model chromosphere includes the essential physics (at least for $T < 25000$ K) even if the actual atomic processes

of most importance are not those included in this study.

In assessing the applicability of the model chromosphere, we must recall the assumption that the heat input per unit mass is unaffected by the perturbation. The results of the stability analysis obviously cannot be applied to the sun if the chromospheric heating mechanism happens to vary in such a manner as to produce stability. However, there is no reason to believe that this is the case. In fact, it has been suggested that the heating mechanism favors instability in the corona (Whitaker, 1963).

In addition to the assumptions made concerning the thermal properties of the chromosphere, the stability analysis employed several approximations to the dynamics. The use of the Boussinesq approximation limits the rigorous applicability of the results to an atmospheric layer with a vertical extent much less than a scale height (Spiegel and Veronis, 1960). Of possibly greater importance is the fact that the atmospheric layer analyzed in Sections 2 and 3 is isolated from its surroundings by artificial boundary conditions. Damping of oscillations by wave generation is thereby excluded. Without a non-local analysis, theoretical predictions of overstability must therefore be regarded as tentative. Finally, it should be noted that the finite-amplitude stability properties of a fluid can differ significantly from the predictions of linear stability theory (Veronis, 1965; but see also Veronis, 1968). However, thermal-convective instability is not likely to exhibit the kind of behavior found by Veronis for thermohaline convection.

5. Temperature Structure

Before interpreting observations of the solar atmosphere in terms of the thermal-convective theory, we investigate the equilibrium temperature structure which follows from the requirement that the energy input be balanced by the energy output. For simplicity we assume that the temperature structure is not affected by the dynamics. Furthermore, the effects of conduction are neglected. These assumptions are not actually valid, as will be evident, but the simple formula which results is very instructive.

As before, we consider an optically thin atmosphere in which the heat loss function depends on the density, temperature, and degree of ionization (x). For added generality (and with the poorly understood heat input in mind), we include a dependence on the vertical coordinate (z). In a steady state and in the absence of conduction, the heat loss function has the same value, namely zero, at all heights so that

$$\frac{d\mathcal{L}}{dz} = \mathcal{L}_z + \mathcal{L}_x \alpha + \mathcal{L}_\rho \frac{d\rho}{dz} + \mathcal{L}_T \beta = 0. \quad (45)$$

Substituting Equations (14) and (15) into Equation (45), we find that the temperature gradient is

$$\beta = \frac{1}{\Delta}\left[\frac{\rho^2 g}{P}(\mathcal{L}_x \mathcal{J}_\rho - \mathcal{L}_\rho \mathcal{J}_x) + \mathcal{L}_z\left(\mathcal{J}_x - \frac{\rho}{1+x}\mathcal{J}_\rho\right)\right]. \quad (46)$$

When $\mathcal{L}$ and $\mathcal{J}$ are given by Equations (2) and (8), respectively, the numerator of this expression for β is positive. In a thermally

stable region of the model chromosphere ($T < 17500$ K), we also have $\Delta > 0$ so that $\beta > 0$. As we ascend in the chromosphere toward the unstable region, T increases and Δ tends to zero with the result that β diverges at the level of marginal stability. At this point conduction is obviously important, and Equation (46) is no longer valid. In addition, the rate of heat input may change owing to the reflection of shock waves at the temperature jump. Such changes in the heat input are symbolized in Equation (46) by $\mathcal{L}_z$, although it is unlikely that such a simple representation is really adequate. We shall assume that the combined effects of conduction, $\mathcal{L}_z$, and convection cause the temperature to increase monotonically with height even in the unstable region with $\Delta < 0$. In any case, it seems reasonable to expect an unstable layer to be bounded below by a steep temperature rise.

Abrupt temperature jumps have been observed in the solar atmosphere (see Sections 6 and 7). Athay and Thomas (1956) attempted to explain these jumps in terms of thermal instability. They argued that temperatures at which the gas is unstable are, in a sense, forbidden inasmuch as any perturbation of the temperature causes it to approach a value for which there is stability. In the simple chromospheric model used in this paper, this statement does not hold for an increase in temperature, but this is simply because sources of radiative cooling other than hydrogen have been neglected. According to Athay and Thomas, the temperature structure of the chromosphere consists of thermally stable plateaus of nearly uniform temperature separated by narrow unstable layers with large temperature gradient.

We have just seen that the observed regions of large temperature gradient can be explained simply in terms of the steady-state requirement of thermal balance without invoking the concept of thermal instability. In this picture the large gradients occur in the neighborhood of levels of marginal stability rather than throughout unstable regions. Weymann (1960) has arrived at similar conclusions on the basis of a model for chromospheric heating by shock waves.

Two minor points must now be made. First, in the analysis of Equation (26) it was implicitly assumed that β and Δ are independent. Equation (46) shows that this is not the case. However, in view of the uncertainty regarding the temperature gradient in the unstable region, it seems desirable to consider β arbitrary, as mentioned in Section 2. Second, the degree of ionization is a known function of temperature in a steady state. In this section, therefore, we could have chosen $\mathcal{L}$ to be a function of ρ and T only. The denominator of the expression for β would then have been $\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho$; the usual condition for thermal instability (Field, 1965) is that this quantity be negative.

6. The Solar Chromosphere

The most conspicuous dynamical features of the "quiet" solar chromosphere are spicules, the observations of which have been reviewed recently by Beckers (1968). I suggest that spicules result from monotonic thermal-convective instability.

We have seen that, in the absence of magnetic fields, the chromosphere can be stable, overstable, or monotonically unstable

(Section 2). Later in this section it will be shown that monotonic instability is very unlikely in field-free regions of the solar chromosphere. According to Section 3, however, magnetic regions with $T > 17500$ K are monotonically unstable. The fact that spicules are observed only in the magnetic regions above the boundaries of supergranules is therefore explained.

The thermal-convective theory predicts that spicule temperatures must exceed some value, which the simple model atom approach indicates to be approximately 17500 K. In fact, spicule temperatures are usually estimated to be in or near the range 20000 to 50000 K. According to a more complete analysis of the atomic physics and radiative transfer (Thomas and Athay, 1961, p. 163), the critical temperature above which the solar chromosphere is thermally unstable is approximately 12000 K. Thus, spicule temperatures of 14000 to 17000 K, as estimated by Beckers (1968), are also consistent with the present theory.

The unstable region where spicules originate must be bounded below by an abrupt temperature jump (Section 5). In fact, there is a very rapid rise from $T \approx 8000$ K to $T \gtrsim 20000$ K at a height of about 1200 km in the chromospheric model of Thomas and Athay (1961). On the basis of his investigation of CaII line formation as well as independent studies by others, Linsky (1968) concluded that such a temperature rise is a real feature of the solar chromosphere. Since this rise presumably occurs around the level of marginal stability at $T = 17500$ K (say), its existence is itself evidence in favor of the thermal-convective theory.

Unfortunately, the height at which spicules are formed cannot be determined by direct observation. It is impossible to estimate reliably the altitudes of features seen on the disk, and limb observations of individual spicules are restricted to rather high levels (above 5000 km at the center of Hα) because of overlapping effects at lower heights. However, analysis of flash spectra has led to the conclusion that departures from spherical symmetry, presumably corresponding to spicules, commence at the level of the steep temperature rise (Thomas and Athay, 1961), just as predicted by the thermal-convective theory. Beckers (1968) emphasizes that the evidence for lack of spicule structure below the temperature rise is weak but concedes that, if spicules do exist at lower heights, their physical conditions must differ significantly from those of spicules seen above the limb.

We expect both upward and downward moving jets to be formed in an unstable chromospheric zone. At the limb, of course, only rising spicules (and, in many cases, their subsequent descent) are observed. However, limb observations refer to elevations considerably higher than the 1200 km level (say) at which spicules originate according to the thermal-convective theory. Consequently, one observes at the limb only those spicules which have been ejected upwards from the unstable region.

The overlapping effects which preclude detailed limb observations of spicules at low heights do not interfere, at least to the same degree, with observations made on the disk. The chromospheric velocity field as seen on the disk in Hα has been described by Leighton et al. (1962). At $\Delta\lambda \simeq 0.3$ Å from the line center, they see "a rather

uniform, fine-grained pattern of both upward and downward velocities," while at $\Delta\lambda \simeq 0.8$ Å "most of the disk is quiescent and the only sizable velocities present are confined to a network of narrow 'tunnels,' through which material streams predominantly downward." The observations, of course, reveal only the most prominent chromospheric velocity fields, and one expects both upward and downward motions at all heights for reasons of mass conservation. In fact, Simon and Leighton (1964) have suggested that the network of downward flow represents the slow return of spicular material to the chromosphere. However, if, with Leighton et al., we make the usual assumption that we see deeper into the atmosphere at greater distances from the line center, we obtain a simple picture in which downward motions predominate at the lowest elevations ($\Delta\lambda \simeq 0.8$ Å), both upward and downward motions are common at higher elevations ($\Delta\lambda \simeq 0.3$ Å), and upward velocities predominate at the greatest heights (limb observations). This picture would, of course, be expected if the observed motions result from an instability at the intermediate elevations, although it is not clear why a network pattern does not appear at $\Delta\lambda \simeq 0.3$ Å. It may be objected (Zirin, 1966, p. 227) that the relationship between $\Delta\lambda$ and height is complicated by the mass motions and may not conform to the assumption made here. Therefore, we now describe supporting evidence which is independent of the $\Delta\lambda$-height relation.

Both upward and downward motions were studied in Hα Doppler movies at $\Delta\lambda = 0.7$ Å by Title (1966). He showed that upward motion occurs mainly in the outer sections of rosettes, while downward flow tends to occur near rosette centers. For this reason, absorbing

regions seen on the violet side of the Hα line center surround the absorbing regions seen in the red wing, as observed also by Bhavilai (1965). This observation can be interpreted as follows. As indicated above, we expect upward motions to predominate above a certain level and downward motions to predominate below this level. We may therefore surmise that the upward moving features observed by Title are higher, on the average, than the downward moving features. It is generally believed that the magnetic field in a rosette diverges with increasing height. We would therefore expect the upward moving features, which of course follow the lines of force, to be farther, on the average, from the rosette axis than the downward moving features, and this is what is observed.

Evidently the observations made by Title (1966) extend down to the level of spicule formation. This fact points to the possibility that the growth of spicules may be depicted in the Doppler movies. In fact, the velocity history of a typical upflow event observed by Title was "a rise to peak velocity in less than thirty seconds and then a decay in velocity for the next ninety seconds." While it is not clear that the growth time of $\lesssim 30$ sec for the finite-amplitude disturbances observed by Title should be reproducible by a linear analysis (especially since the same growth time was not reported for downflow events), it seems reasonable to compare Title's result with the growth time predicted by the thermal-convective theory.

Title showed that the velocity activity he observed took place in the elements of the chromospheric network and therefore in regions of appreciable magnetic field. Let us assume that the field is sufficiently

large that the growth rate of the thermal-convective instability is given by Equation (44) and is therefore independent of the field strength. For temperatures in the range 20000 to 30000 K, the appropriate root of Equation (44) is given very nearly by $n = 10^{11}\rho$, which corresponds to an e-folding time of $6 \times 10^{12}/N_e$. A reasonable value for the electron density at the level of spicule formation is $N_e = 3 \times 10^{11}$ cm^{-3} (Thomas and Athay, 1961, p. 386), which yields a growth time of 20 sec, in excellent agreement with Title's result. From the complete secular equation (Equation [43]), we can now determine the magnetic field required for the spicule growth rate to be given by Equation (44). For $N_e = 3 \times 10^{11}$ cm^{-3} and $T = 20000$ K, it can be shown that, unless the temperature gradient is unexpectedly large, Equation (44) yields a good approximation if $B >> 2.5 \times 10^{-7} L$, where L is the vertical dimension (in cm) of the initial perturbation which produces a spicule.

It has already been mentioned that the Boussinesq approximation used in the stability analysis is not strictly valid owing to the large vertical extent of the chromospheric layer under consideration (Section 4). We must now point out that the short time scale just derived also invalidates the Boussinesq approximation since pressure equilibrium can be assumed only when the length scale is much less than the time scale multiplied by the sound speed (see paper I), i.e., much less than 500 km, and this is not the case.

Since the growth time computed from Equation (44) is independent of the disturbance scale, this parameter cannot be predicted by the linear theory. Thermal conduction introduces a wavenumber dependence

but is important only when the time required for the conductive smoothing of temperature fluctuations is less than or comparable to the growth time for the instability in the absence of conduction (paper I). For the numerical example just considered, this condition is fulfilled only if the scale of the disturbance parallel to the magnetic field is less than 5 km, which is several orders of magnitude smaller than the lengths of spicules.

A theory for the temperature structure of the thermally unstable region of the chromosphere does not exist at the present time and, as we saw in Section 5, may require a reliable calculation of the energy input. If such a theory predicts that the unstable region is very thin (say $\lesssim$ 5 km) in the absence of motions, this region will actually be stable (for vertical magnetic fields) owing to conduction. The theory of spicule formation presented here would then be untenable.

Before concluding the discussion of spicules, we note that Thomas and Athay (1961) and Kopp (1963) have also suggested that spicules may result from thermal instability. In fact, "the apparently simultaneous onset of inhomogeneity and abrupt rise in T_e" coupled with their concept of temperature plateaus (see Section 5) led Thomas and Athay (p. 383) to ask whether "spicules somehow originate in the region of abrupt rise in T_e, reflecting the radiative instability of this region." They supposed that spicules might be produced by a collapse to higher density (followed by an outward acceleration of unknown origin) since spicules observed at the limb are more dense than their surroundings and also because considerations of thermal instability tend to emphasize condensation

to the exclusion of rarefaction. In the thermal-convective interpretation, of course, spicules which are eventually observed at the limb are initially less dense than their surroundings. After accelerating upwards because of buoyancy, they lose the driving instability and simply coast into view.

We now consider chromospheric regions without magnetic fields. It was shown in Section 2 that inequality (32) is a necessary and sufficient condition for instability in field-free regions. If this condition is satisfied, the resulting instability takes the form of exponentially amplifying oscillations provided inequality (35) is also fulfilled. When $T = 20000$ K, inequality (32) is satisfied if $N_e > 4 \times 10^9$ cm^{-3} or $\beta > 4.3 \times 10^{-4}$ K cm^{-1}, and inequality (35) is satisfied if

$$\beta > 0.92 \times 10^{-4} (N_e/10^{10})^2 \text{ K cm}^{-1}. \qquad (47)$$

For $N_e = 3 \times 10^{11}$ cm^{-3}, the value assumed above for the region of spicule formation, we would expect monotonic instability since inequality (32) is definitely satisfied while inequality (47) probably is not. However, chromospheric structure in field-free regions, i.e., within the cells of the chromospheric network, differs considerably from the structure of the cell borders where spicules are located. Spectroheliograms made in the HeI λ10830 line show absorption in quiet regions only at the borders of the network cells (Zirin and Howard, 1966). Since this line is produced by regions with $T \geq 20000$ K, it is inferred that these regions either are much less dense or have much smaller vertical extent (i.e., much larger β) inside the network cells than at the cell borders. In

either case, it seems likely that inequality (47) is satisfied in the cell interiors.

A quantitative demonstration that inequality (47) holds in interspicule regions can be derived from radio observations, which place strict upper limits on the amount of hot material in the chromosphere (Zirin and Dietz, 1963). For example, the fact that the brightness temperature of the quiet sun at $\lambda = 2$ cm is 9100 ± 600 K (Buhl and Tlamicha, 1968) implies that the optical thickness (τ) at this wavelength of chromospheric regions with $T \approx 20000$ K is less than unity. If we set the linear thickness of regions with $T \approx 20000$ K equal to T/β and compute the free-free absorption coefficient at 2 cm for $T = 20000$ K, we find that the condition $\tau < 1$ takes the form

$$\beta > 2.7 \times 10^{-4} (N_e/10^{10})^2 \text{ K cm}^{-1}. \qquad (48)$$

This inequality applies to interspicule regions inasmuch as these regions occupy most of the chromospheric volume. Although some ambiguity has been introduced by the fact that inequality (48) refers to a mean value of N_e^2/β over a range of T, it seems almost certain from this inequality that inequality (47) is satisfied at $T = 20000$ K. We therefore conclude that field-free regions are probably not monotonically unstable.

In view of the data given just before inequality (47), it seems probable (but not definite) that inequality (32) is fulfilled and therefore that field-free regions are overstable. Oscillations may therefore be expected within the cells of the chromospheric network. Several

types of oscillations are known to take place in the solar atmosphere. The most famous of these are the 300 sec oscillations (Leighton et al., 1962), which form the so-called resonance peak of temporal power spectra of the solar velocity field. As one observes increasingly strong lines, these power spectra develop a high-frequency tail corresponding to oscillation periods of about 180 sec (Evans et al., 1963). On the basis of the periods alone, it is not possible to decide whether the 300 sec or 180 sec oscillations (if either) should be ascribed to thermal-convective overstability since both periods can be obtained from Equation (26) for reasonable values of the chromospheric parameters. However, the 300 sec oscillations have been observed deep in the photosphere (Edmonds et al., 1965) and may very well be generated there by the motions of granules (Evans and Michard, 1962). The 180 sec oscillations, on the other hand, are a distinctly chromospheric phenomenon. Furthermore, while several features of the 300 sec oscillations can be explained in terms of an elementary wave theory (Noyes and Leighton, 1963), the 180 sec oscillations do not appear to be amenable to such an interpretation (Evans et al., 1963). I therefore suggest that the high-frequency oscillations observed in the chromosphere result from thermal-convective overstability.

If this suggestion is correct, the 180 sec oscillations should be observed primarily inside the network cells rather than at the cell borders. Evidence that this is, in fact, the case comes from Orrall's (1966) investigation of the properties of oscillations observed in the core (K_3) of the K line as a function of the brightness of the violet K_2

emission peak. Orrall found that the high-frequency tail dominates the power spectrum of K_3 velocities only for regions in which K_2 is faint, i.e., only inside the network cells. It should be noted, however, that oscillations with periods less than 200 sec were also observed in bright K_2 regions. Additional evidence is provided by $H\alpha$ cinematograms. According to Zirin (1966, p. 287), these show that regions inside the network cells oscillate, whereas the cell borders do not. However, the meaning of this observation is not completely clear since Zirin quotes a period of 250 sec.

Although the observations appear to be consistent with a thermal-convective interpretation of short-period chromospheric oscillations, the theoretical uncertainties are considerable. There is at least one model of interspicule regions (Beckers, 1968) in which neither N_e nor β is large enough at $T = 20000$ K to satisfy inequality (32). On the other hand, β may be so large that the overstable region is too thin to be observed. Of course, the overstable layer may generate observable oscillations in neighboring regions, but we recall from Section 4 that this possibility introduces a damping mechanism not included in the calculations of this paper. Finally, conduction will stabilize the layer if it is sufficiently thin.

Somewhat related to the thermal-convective interpretation is the idea that chromospheric oscillations are Väisälä oscillations of a convectively stable atmosphere (Jensen and Orrall, 1963; Ulmschneider, 1968). The main difficulty with this idea is that one still must explain how the oscillations are excited and maintained.

7. The Corona

It has been suggested that thermal instability in the corona is responsible for the formation of solar prominences (Kiepenheuer, 1953). According to calculations of radiative cooling (Pottasch, 1965; Raju, 1968; Cox and Tucker, 1969), the solar atmosphere should be thermally unstable for $T \gtrsim 10^5$K. The radiation time scale predicted by these calculations for $T = 10^6$ K is about $6 \times 10^{12}/N_e$. This time scale is less than the time scale for conductive smoothing of temperature fluctuations only if the length scale of the fluctuations exceeds $10^{19}/N_e$. For this reason, field-free regions of the corona are probably stable. However, magnetic regions are unstable to disturbances (such as streamers) with sufficiently large length scales parallel to the field.

Although the main radiation losses at coronal temperatures are due to elements other than hydrogen, thermal instability in the corona results from the same ionization effects treated in paper II. The analysis of Section 3 is therefore applicable (although conduction invalidates the Boussinesq approximation since the dimension parallel to the field required for instability is comparable to or greater than a scale height), and we may expect monotonic thermal-convective instability in magnetic regions of the corona.

The large Doppler spreading of solar radar echoes has demonstrated the existence in the corona of both upward and downward velocities of the order of 100 km/sec (James, 1966). I suggest that these velocities be attributed to coronal spicules, that is, to jets formed by monotonic thermal-convective instability in coronal magnetic

fields. Some indication that magnetic fields do play a role is provided by the observed long-term correlation between the radar cross-section of the sun and the sunspot number. The time delay vs. frequency plots of radar echo energy display a characteristic pattern such that the fraction of energy Doppler shifted to higher frequencies decreases with increasing time delay (James, 1968). This observation may simply reflect the tendency of upward moving jets to be higher, on the average, than the downward moving jets. In addition to the radar evidence, one should recall the suggestive appearance of coronal rain.

While observations of coronal velocity fields are consistent with a thermal-convective interpretation, they provide no unambiguous confirmation of the theory. However, the theory also predicts an abrupt temperature jump at the level of marginal stability around $T \approx 10^5$ K. In fact, as is well known, the transition from the chromosphere to the corona is very rapid. The structure of this transition has been determined by Athay (1966a) and, including dielectronic recombination, by Dupree and Goldberg (1967) on the basis of ultraviolet emission line intensities. Besides verifying the extreme abruptness of the chromosphere-corona transition, these authors showed that the temperature gradient reaches a maximum near $T = 10^5$ K, as expected from the considerations of Section 5. This result indirectly provides the most convincing support available for the thermal-convective interpretation of the radar observations.

Kuperus and Athay (1967) have pointed out what at first sight appears to be an interesting dilemma associated with the chromosphere-

corona transition. Following the ideas of Athay and Thomas (1956) discussed in Section 5, they assert that the temperature rise begins when hydrogen and helium are no longer able to radiate away the mechanical energy input. The temperature jump required for efficient radiation from impurities causes a downward conduction of heat which must be absorbed by the lower part of the transition. Since the transition was presumably caused by the inability of the gas to radiate away the mechanical energy input, it is argued that the upper chromosphere is unable to dispose of the additional energy conducted down from the corona. Kuperus and Athay suggest that a Rayleigh-Taylor instability results and is responsible for the formation of spicules. On this basis, they derive a theoretical spicule growth time of 25 sec, which seems to be in accord with observation (Section 6).

While there may be some question regarding the existence of an equilibrium state, the situation described by Kuperus and Athay is actually stable in the Rayleigh-Taylor sense. These authors have overlooked the fact that a gravity field in one direction is equivalent to an acceleration field in the opposite direction; thus their value of 25 sec is, apart from a factor of 2π, an oscillation period and not a growth time. In addition, we have seen in Section 5 that, if the heat loss function $\mathcal{L}$ is considered to be a function of ρ and T only, the steep temperature rise begins when the positive quantity $\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho$ becomes small. Since $\mathcal{L}_\rho > 0$, it follows that $\mathcal{L}_T > 0$ so that the radiation rate in the lower part of the transition is still an increasing function of temperature. Indeed, the reason for the initial temperature

rise is simply that an increase in T is required to compensate for the outward decrease in ρ if the radiation power is to be maintained. Therefore, the lower part of the transition may, in fact, be able to radiate away the conductive flux by an appropriate increase in T. This view is supported by the near constancy of the conductive flux for $T > 10^5$ K found by Athay (1966a) and by Dupree and Goldberg (1967). Evidently, the conductive flux is absorbed primarily by the regions with $T < 10^5$ K and $\mathcal{L}_T > 0$.

It is interesting to note that Rayleigh-Taylor instability is actually possible if conduction is ignored. If we assume $\mathcal{L} = \mathcal{L}(\rho, T)$, it can be shown by elementary considerations analogous to those of Section 5 that an atmosphere which is isochorically stable ($\mathcal{L}_T > 0$) but isobarically unstable ($\mathcal{L}_T - \frac{\rho}{T}\mathcal{L}_\rho < 0$) is subject to a Rayleigh-Taylor instability since $d\rho/dz > 0$ (Field, private communication). In terms of the analysis of this paper, substitution of Equation (46) into Equation (15) shows that $d\rho/dz > 0$ if Δ is negative and sufficiently small in absolute value. However, it is clear from Equation (46) that conduction cannot be ignored when $|\Delta|$ is small. In any case, the solar atmosphere will not exhibit this Rayleigh-Taylor instability since $d\rho/dz < 0$ in hydrostatic equilibrium as long as $\beta > 0$.

8. Summary

The following picture of the dynamical and thermal structure of the outer solar atmosphere has been developed in this paper:

(1) There are two zones of thermal-convective instability in the outer solar atmosphere, one in the chromosphere and one in the corona.

(2) Spicules originate in regions of the unstable chromospheric zone which contain magnetic fields.

(3) Oscillations with periods of about 180 sec result from overstability of field-free regions of the chromospheric zone. This conclusion is particularly tentative.

(4) Elongated convective jets, which may be called coronal spicules, are formed in magnetic regions of the corona.

(5) Field-free regions of the corona are stabilized by thermal conduction.

(6) The lower boundaries of both zones are marked by abrupt temperature jumps.

Although a number of refinements in the theory are desirable, the most pressing need is for a theory of the temperature structure of the unstable regions in the absence of motions. If the unstable chromospheric region, for example, is found to be exceedingly thin, this region will be stabilized by conduction.

Acknowledgements

I thank Drs. P. Goldreich, E. Spiegel, and H. Zirin for their comments and suggestions. I also had a helpful discussion with Dr. G. Field, and Dr. A. Maxwell introduced me to the radar observations. Financial support from the California Institute of Technology is gratefully acknowledged

References

Athay, R. G.: 1966a, Astrophys. J. 145, 784.

Athay, R. G.: 1966b, Astrophys. J. 146, 223.

Athay, R. G. and Thomas, R. N.: 1956, Astrophys. J. 123, 299.

Beckers, J. M.: 1968, Solar Phys. 3, 367.

Bhavilai, R.: 1965, Monthly Notices Roy. Astron. Soc. 130, 411.

Buhl, D. and Tlamicha, A.: 1968, Astrophys. J. (Letters) 153, L189.

Cox, D. P. and Tucker, W. H.: 1969, Astrophys. J. 157, 1157.

Defouw, R. J.: 1970a, Astrophys. J. (in press).

Defouw, R. J.: 1970b, Astrophys. J. (in press).

Dupree, A. K. and Goldberg, L.: 1967, Solar Phys. 1, 229.

Edmonds, F. N., Jr., Michard, R., and Servajean, R.: 1965, Ann. Astrophys. 28, 534.

Evans, J. W. and Michard, R.: 1962, Astrophys. J. 136, 493.

Evans, J. W., Michard, R., and Servajean, R.: 1963, Ann. Astrophys. 26, 368.

Field, G. B.: 1965, Astrophys. J. 142, 531.

James, J. C.: 1966, Astrophys. J. 146, 356.

James, J. C.: 1968, in Radar Astronomy (ed. by J. V. Evans and T. Hagfors), McGraw-Hill, p. 369.

Jensen, E. and Orrall, F. Q.: 1963, Astrophys. J. 138, 252.

Kaplan, W.: 1962, Operational Methods for Linear Systems, Addison-Wesley.

Kiepenheuer, K. O.: 1953, in The Sun (ed. by G. P. Kuiper), University of Chicago Press, p. 430.

Kopp, R. A.: 1963, unpublished.

Kuperus, M. and Athay, R. G.: 1967, Solar Phys. 1, 361.

Leighton, R. B., Noyes, R. W., and Simon, G. W.: 1962, Astrophys. J. 135, 474.

Linsky, J. L.: 1968, Thesis, Harvard.

Noyes, R. W. and Leighton, R. B.: 1963, Astrophys. J. 138, 631.

Orrall, F. Q.: 1966, Astrophys. J. 143, 917.

Pottasch, S. R.: 1965, Bull. Astron. Inst. Neth. 18, 7.

Raju, P. K.: 1968, Monthly Notices Roy. Astron. Soc. 139, 479.

Rayleigh, Lord: 1916, Philosophical Magazine (Series 6) 32, 529.

Simon, G. W. and Leighton, R. B.: 1964, Astrophys. J. 140, 1120.

Spiegel, E. A. and Veronis, G.: 1960, Astrophys. J. 131, 442.

Thomas, R. N. and Athay, R. G.: 1961, Physics of the Solar Chromosphere, Interscience Publ., New York.

Title, A. M.: 1966, Thesis, Cal. Inst. of Technology.

Ulmschneider, P. H.: 1968, Astrophys. J. 152, 349.

Veronis, G.: 1965, J. Marine Res. 23, 1.

Veronis, G.: 1968, J. Fluid Mech. 34, 315.

Weiss, N. O.: 1964, Philosophical Trans. Roy. Soc. A 256, 99.

Weymann, R.: 1960, Astrophys. J. 132, 452.

Whitaker, W. A.: 1963, Astrophys. J. 137, 914.

Zirin, H.: 1966, The Solar Atmosphere, Blaisdell Publ.

Zirin, H. and Dietz, R. D.: 1963, Astrophys. J. 138, 664.

Zirin, H. and Howard, R.: 1966, Astrophys. J. 146, 367.