Astronomy Reports, Vol. 48, No. 11, 2004, pp. 877–899. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 963–987.
Original Russian Text Copyright © 2004 by V. Polyachenko, E. Polyachenko.

A Unified Theory for the Formation of Galactic Structures

V. L. Polyachenko and E. V. Polyachenko

Institute of Astronomy, Russian Academy of Sciences,
Pyatnitskaya ul. 48, Moscow, Russia

Received July 30, 2003; in final form, May 27, 2004

Abstract—A new theory for the formation of the main structures of galaxies is proposed: these structures are viewed as low-frequency normal modes in disks consisting of precessing stellar orbits. Mathematically, the theory is based on an integral equation in the form of a classical eigenvalue problem, with the eigenvalues being equal to the angular velocities Ω_p of the modes. Analysis of the general properties of the master integral equation (without finding concrete solutions) shows that it admits two types of solutions: barlike and spiral. The numerical algorithms are discussed and particular solutions of the integral equation are presented. If resonance interaction can be neglected, the bar mode represents a neutral perturbation of the disk. This mode can be amplified by the effect of the long-range gravitational field of the mode on stars located in the vicinity of the corotation and outer-Lindblad resonances. Spiral perturbations are waves with zero total angular momentum, and spiral modes are excited at the inner-Lindblad resonance. The approach proposed is compared to currently accepted mechanisms for the formation of galactic structures. In particular, Toomre's application of the swing amplification mechanism to explain the formation of global modes is critically discussed. In addition, we show that it is not correct to simulate the real stellar velocity dispersion in a galaxy using softened gravity. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

In this paper, we assume, as has been assumed in many previous studies, that the observed spiral and barlike structures in galaxies represent normal modes in a flat, axisymmetric disk. Strictly speaking, the computation of normal modes is a rather difficult problem. General integral equations for normal modes were derived long ago by Kalnajs [1] and Shu [2]. However, these equations are quite complicated, and have rarely been used. For the same reason, neither the physical mechanisms nor the instabilities responsible for the formation of the modes could be understood by analyzing the general equations. At present, analyses of stellar disks are usually carried out via numerical N-body simulations, which are difficult to interpret. In addition, most analytical studies have replaced the stellar disk with an "effective" gaseous disk.

The situation becomes much simpler if we are interested in *low-frequency* rather than arbitrary modes. Let us explain why. Let an axisymmetric disk be characterized by the unperturbed potential $\Phi_0(r)$ and the corresponding angular velocity $\Omega(r)$ of its circular rotation. Consider an orbit precessing at an angular velocity Ω_{pr} and subject to a perturbing potential rotating at the angular velocity Ω_p. Lynden-Bell [3] pointed out that the orbit varies little during one orbital period of the star if

$$\epsilon = \frac{\delta\Omega}{\Omega} = \frac{|\Omega_p - \Omega_{pr}|}{\Omega} \ll 1. \qquad (1)$$

We can therefore assume that the orbit as a whole (and not just individual stars moving in the orbit) should respond to the perturbation. We thus come to a model with the disk viewed as a set of precessing orbits. In this model, galactic structures are compression and rarefaction waves in the orbit density traveling in the azimuthal direction at some angular velocity Ω_p, which is of the same order of magnitude as Ω_{pr}. The velocity of the orbital precession is small compared to the angular rotational velocity of the central regions of the disk, so that, in this sense, bars and spirals are low-frequency structures.

The resulting description of stellar systems is similar to the drift approximation in the physics of a magnetized plasma (see, e.g., [4]), but for orbits of a much more general form: Larmor circles in a plasma correspond to orbits with, strictly speaking, an arbitrary degree of oblateness. It is important that, in plasma physics, low-frequency (drift) waves are most likely to become unstable [5].

Let us hypothesize that observed galactic structures can be described as low-frequency modes. At first glance, it appears that this cannot be plausible, since bars and spirals usually reach the corotation

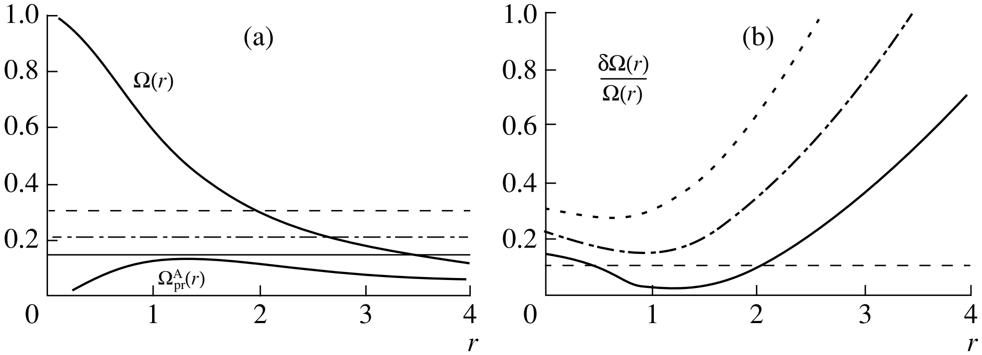

Fig. 1. (a) Curves $\Omega(r)$ and $\Omega_{pr}(r)$ for a Plummer potential, which AS used to compute bar modes. Horizontal lines correspond to various possible angular velocities of the wave, Ω_p. (b) The $\delta\Omega(r)/\Omega(r)$ curves corresponding to the left-hand plot.

radius, when the pattern speed is equal to the angular rotational velocity of the disk. Note, however, that the disk mode may be determined by the central part of the disk, due to the strong central concentration of the mass and the rapid decrease in the density with radius.

We now show that published computations and estimates of the angular velocities of galactic bars and spirals justify the use of a model that treats the disk as a set of precessing orbits; i.e., that justify use of inequality (1).

Athanassoula and Sellwood [6] (hereafter AS) used N-body simulations to analyze the development of normal modes in more than 30 model stellar disks with various distribution functions in a Plummer potential $\Phi_0(r) = -1/\sqrt{r^2+1}$. Figure 1a shows the dependences of the disk rotational angular speed $\Omega(r)$ and the precessional speed $\Omega_{pr}(r)$ in the epicycle approximation adopted by AS, with several horizontal lines corresponding to the rotational velocities of bar modes. Recall that $\Omega_{pr}(r) = \Omega(r) - \kappa(r)/2$ for nearly circular orbits, where $\kappa(r)$ is the epicyclic frequency and $\kappa^2 = 4\Omega^2 + r d\Omega^2/dr$. The thin solid, dash–dotted, and dashed horizontal lines correspond to the minimum ($\Omega_p^{\min} = 0.14$), average ($\Omega_p^{\rm mid} = 0.21$), and maximum ($\Omega_p^{\max} = 0.3$) angular velocities of the bar modes, according to the list given by AS in their Table 1. Figure 1b shows the ratios $\delta\Omega/\Omega$ for the three modes mentioned above. These ratios are smaller than 0.1 for the first of these modes, which is localized at $r < 2$ according to AS (thin dashed line in Fig. 1b). However, $\delta\Omega/\Omega < 0.3$ even for the highest-frequency bar mode ($\Omega_p = \Omega_p^{\max} = 0.3$), if we take into account the fact that this mode is more concentrated toward the center. We show below that, even if some of the orbits obey an appreciably weaker inequality than (1), $\delta\Omega/\Omega < 1$ (e.g., by a factor of a few but not by an order of magnitude), the accuracy of the final results remains high.

Figures 2a and 2b show plots for the Milky Way similar to those shown in Figs. 1a and 1b. Here, we have adopted the data of the classic paper of Lin *et al.* [7]. The horizontal line in Fig. 2a corresponds to the mean value $\Omega_p = 12$ km s^{-1} kpc^{-1} for the pattern speed interval $\Omega_p = 11$–13 km s^{-1} kpc^{-1} recommended in [7]. It follows from Fig. 2b that $\delta\Omega/\Omega \ll 1$ for this Ω_p. Here, we must make the reservation that no consensus has been reached about the angular speed of the spiral pattern in our Galaxy. Some authors have reported Ω_p values twice as high as the one we use above.[1] However, even in this case, inequality (1) is satisfied, and the eigenmodes of the Galactic disk can be analyzed in terms of a low-frequency-mode approximation.

Thus, the aim of this paper is to develop the theory of low-frequency modes of stellar disks. In Section 2, we use perturbation theory in the small parameter ϵ to derive a master integral equation for the low-frequency normal modes. Section 3 gives a general qualitative analysis of this equation. Note that an initial integral equation for the low-frequency modes of a stellar disk has already been derived by one of us (VLP) in [9]. Since then, a simplified version of this equation has been used only once [10], to compute the anomalously slow bar modes found by AS in their N-body simulations. Our return to and enhanced interest in this equation was due first and foremost to the transformation of the initial integral equation proposed by one of us (EVP). Although this transformation is very simple (we simply exchange one of the unknown functions for a different function), it results in an integral equation in the form of a classical eigenvalue problem, where the eigenvalues are equal to the pattern speeds Ω_p themselves. Standard

[1] Blitz [8] gives approximately the same Ω_p values for the four-armed spirals beyond the solar circle. In this case, we must be dealing with a separate peripheral tier of the spiral pattern of our Galaxy.

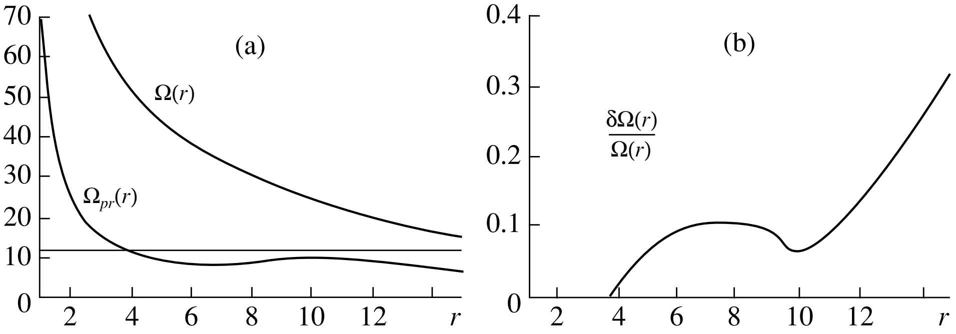

Fig. 2. Same as Fig. 1 for the model of the Galaxy of Lin *et al.* [7].

mathematical program packages can be used to solve this equation numerically and identify all of its eigenmodes, making it possible to appreciably enhance the efficiency of the computations. More importantly, the integral equation in the transformed form can be easily analyzed in general, enabling us to derive the main properties of its solutions. In particular, we can demonstrate the crucial role of the derivative of the distribution function $\mathcal{F}_0$ with respect to the angular momentum L for a fixed value of the Lynden-Bell invariant J_f [3], $\mathcal{F}_0' \equiv \partial\mathcal{F}_0(J_f, L)/\partial L$. Depending on the behavior of $\mathcal{F}_0'$, a general (unified) theory of barlike and spiral modes can be developed (Section 3).

In Sections 4–6, we analyze specific numerical solutions of our master integral equation for bisymmetric modes (with azimuthal wavenumber $m = 2$). To demonstrate the potential of the proposed theory, we begin in Section 4 by analyzing test models studied earlier by AS using an N-body approach, and show that the results of our theory agree well with the N-body simulations of AS. Bearing in mind the somewhat artificial nature of the models considered by AS, we analyze first and foremost in Sections 5 and 6 other, much more realistic and general models, based on a general Schwartzschild distribution function [2] to obtain a more adequate representation of spiral modes.

In the Conclusion (Section 7), we make a detailed comparison of our approximation with existing theories. The most important among these are based on the swing-amplification mechanism [11]. A detailed criticism of the swing mechanism, which has enchanted most specialists in stellar dynamics for more than 20 years,[2] is the second important aim of this paper. We show that the current unsatisfactory situation in the theory of galaxy dynamics is due, first, to the lack of an adequate theory of global modes (resulting in attempts to analyze these modes in the language of local density-wave theory, which is ill suited for this) and, second, to the application of a beautiful but not entirely successful analogy with lasers (testified to by the adoption of terms that speak for themselves, such as the "waser mechanism").

Appendix I gives some additional arguments in support of our approach, based on the matrix equation derived in this Appendix, which is a natural generalization of the master integral equation of Section 2. Appendix II describes the numerical algorithms used to solve the master integral equation. Appendix III analyzes the very popular (but fundamentally flawed) method of simulating stellar-velocity dispersions using softened gravity. This issue is very important, because, among other things, such simulations have been used by the authors of certain classic books, and have then migrated into textbooks on stellar dynamics (see, e.g., [13]).

2. DERIVATION OF THE MASTER INTEGRAL EQUATION FOR THE LOW-FREQUENCY MODES OF A STELLAR DISK

The unperturbed motion of a star in a flat, axisymmetric field with the potential $\Phi_0(r)$ is described by the Hamiltonian

$$H_0 = \frac{\mathbf{v}^2}{2} + \Phi_0(r), \qquad (2)$$

where $\mathbf{v}$ is the velocity of the star. Let us now introduce in the standard way the action and angle variables $(\mathbf{I}, \mathbf{w})$. Actions are defined as integrals over the complete period for the variation in the coordinates of the star:

$$I_1 = \frac{1}{2\pi}\oint p_r dr, \quad I_2 = \frac{1}{2\pi}\oint p_\varphi d\varphi = L, \qquad (3)$$

where $p_r = [2(H_0 - \Phi_0(r)) - L^2/r^2]^{1/2}$ and $p_\varphi = L$ are the generalized momenta (see, e.g., [14]) and L is the angular momentum.

[2] Acceptance of the swing ideology unified even competing (with regard to other issues) teams of theoreticians dealing with the formation of galactic structures. An excellent review of the modern history of achievements in the theory of spiral structure is given by Pasha [12].

The angular variables $\mathbf{w} = (w_1, w_2)$ are canonically conjugate to $\mathbf{I} = (I_1, I_2)$. The unperturbed Hamiltonian depends solely on the action variables: $H_0 = H_0(\mathbf{I})$.

Consider now a time-dependent perturbation of the initial system:

$$H = H_0 + \Phi(r, \varphi; t), \tag{4}$$

where Φ is the perturbation of the potential, which can be expanded into a Fourier series in the angular variables $\mathbf{w}$:

$$\begin{aligned} \Phi(r, \varphi; t) &= \Psi(\mathbf{I}, w_1) e^{i(mw_2 - \omega t)} \qquad (5) \\ &= e^{i(mw_2 - \omega t)} \sum_l \Psi_l(\mathbf{I}) e^{ilw_1}, \end{aligned}$$

where ω is the frequency. The expansion (5) is a single-variable series in l, since we are interested only in configurations with a certain azimuthal wavenumber m. Below, we analyze the bisymmetric modes, $m = 2$.

Since $\Omega_i(\mathbf{I}) = \partial H_0/\partial I_i$, we can write the dynamical equations in the form

$$\dot{I}_i = -\frac{\partial \Phi}{\partial w_i}, \qquad \dot{w}_i = \Omega_i(\mathbf{I}) + \frac{\partial \Phi}{\partial I_i}. \tag{6}$$

Recall that $\Omega_1 = \kappa(r)$ and $\Omega_2 = \Omega(r)$ for nearly circular orbits.

In the first-order perturbation theory (in Φ), the forces in (6) can be calculated by substituting the unperturbed orbits. We then have for the first order corrections to I_i [15]

$$\Delta_1 I_i = \partial \chi / \partial w_i, \tag{7}$$

where

$$\chi = \frac{e^{i(2w_2 - \omega t)}}{i} \sum_l \frac{\Psi_l(\mathbf{I}) e^{ilw_1}}{l\Omega_1 + 2\Omega_2 - \omega}. \tag{8}$$

This expansion contains one dominant term, which corresponds to l values for which the sum $(2\Omega_2 + l\Omega_1)$ in the denominator is small and equal to $2\Omega_{pr}$ ($\Omega_{pr}(\mathbf{I}) = \Omega_2(\mathbf{I}) - \Omega_1(\mathbf{I})/2$ is the precessional velocity of the orbit); i.e., $l = -1$. It follows that the denominator of the dominant term is equal to $\omega - 2\Omega_{pr} = 2(\Omega_p - \Omega_{pr})$; it is natural to call this term the inner-Lindblad-resonance (ILR) term. Similarly, the corotation (CR) and outer-Lindblad-resonance (OLR) terms correspond to $l = 0$ and $l = 1$; compared to the ILR term, they are small in the Lynden-Bell parameter ϵ from (1). Of course, this is true only for the central regions of the disk, which are sufficiently far from the corotation radius (where the CR term becomes important). At first glance, this means that the corotation term must be taken into account when describing bars, since bars extend approximately out to the corotation radius. However, in reality, the amplitude of the bar is large only in the central region and decreases rapidly with galactocentric distance. Therefore, the corotation term does not play a significant role compared to the ILR term. We confirm this statement via detailed numerical computations in Appendix I.

We thus have in the first-order perturbation theory

$$\chi \approx \frac{e^{-i\omega t}}{i} \Psi_{-1}(\mathbf{I}) \frac{e^{2i(w_2 - w_1/2)}}{2(\Omega_{pr} - \Omega_p)}. \tag{9}$$

We obtain from (9)

$$\Delta I_1 = \frac{\partial \chi}{\partial w_1} \approx -i\chi, \qquad \Delta I_2 = \frac{\partial \chi}{\partial w_2} \approx 2i\chi. \tag{10}$$

Consequently,

$$\Delta I_1 + \Delta I_2/2 \equiv \Delta J_f = 0,$$

i.e., the quantity

$$J_f = I_1 + I_2/2$$

is an invariant. Note that J_f was first introduced by Lynden-Bell [3].

It is now natural to change to new action variables, $\mathbf{I} = (I_1, I_2) \to \mathbf{J} = (J_f, L)$. In addition, Eq. (9) suggests the convenient change of angular variables $\mathbf{w} = (w_1, w_2) \to \bar{\mathbf{w}} = (\bar{w}_1, \bar{w}_2)$:

$$\bar{w}_1 = w_1, \qquad \bar{w}_2 = w_2 - w_1/2. \tag{11}$$

The new variable $\bar{w}_2$ is slow compared to $\bar{w}_1$. Indeed, $\bar{w}_2(t) = \Omega_{pr} t$ for unperturbed orbits, whereas $\bar{w}_1(t) = \Omega_1 t$. Note that the variables $\mathbf{J}$ and $\bar{\mathbf{w}}$ introduced in this way are canonically conjugate.

If we now write the perturbed potential as a function of the new variables in the form

$$\begin{aligned} \Phi(r, \varphi; t) &= \Phi(\mathbf{J}, \bar{w}_1) e^{i(m\bar{w}_2 - \omega t)} \qquad (12) \\ &= e^{i(m\bar{w}_2 - \omega t)} \sum_l \Phi_l(\mathbf{J}) e^{il\bar{w}_1}, \end{aligned}$$

we can easily see that the quantity Ψ_{-1} in (9) is equal to the function $\Phi(\mathbf{J}, \bar{w}_1)$ averaged over $\bar{w}_1$:

$$\Psi_{-1} = \bar{\Phi} = \frac{1}{2\pi} \int_0^{2\pi} d\bar{w}_1 \Phi(\mathbf{J}, \bar{w}_1). \tag{13}$$

We then use (9) and (10) to obtain (dropping the dependence $e^{2i\bar{w}_2 - i\omega t}$)

$$\Delta L = -\frac{\bar{\Phi}(\mathbf{J})}{\Omega_{pr}(\mathbf{J}) - \Omega_p}. \tag{14}$$

Let us suppose that the distribution of the orbits is described by the function $\mathcal{F}_0(J_f, L)$ at the initial time t_0. Let ΔL be a small variation in the angular momentum arising due to a small perturbing potential that acts from time t_0 until the current time t. Since the flow of phase fluid is incompressible, a phase-space element with density $\mathcal{F}_0(J_f, L - \Delta L)$ arrives at point (J_f, L) at time t. The perturbation of the distribution function can therefore be computed as the Euler difference

$$\bar{\mathcal{F}} = \mathcal{F}_0(J_f, L - \Delta L) - \mathcal{F}_0(J_f, L) \quad (15)$$
$$\approx -\Delta L \frac{\partial \mathcal{F}_0}{\partial L} = \frac{\partial \mathcal{F}_0}{\partial L} \frac{\bar{\Phi}}{\Omega_{pr} - \Omega_p}.$$

This expression relates the distribution function $\bar{\mathcal{F}}$ and the slowly varying component of the perturbed potential. This relation is proportional to the derivative

$$\mathcal{F}_0' \equiv \left. \frac{\partial \mathcal{F}_0}{\partial L} \right|_{J_f} = -\frac{1}{2} \frac{\partial \mathcal{F}_0(\mathbf{I})}{\partial I_1} + \frac{\partial \mathcal{F}_0(\mathbf{I})}{\partial I_2}. \quad (16)$$

We show below that precisely this derivative plays a crucial role in our theory of low-frequency modes; we will refer to it as the Lynden-Bell derivative of the distribution function.[3]

We now compute the perturbed surface density

$$\Sigma = \int d\mathbf{v} \bar{\mathcal{F}} \quad (17)$$

and use the formula for the potential of a simple layer to obtain

$$\Phi(\mathbf{r}) = -G \int d\mathbf{r}' \frac{\Sigma(\mathbf{r}')}{r_{12}} = -G \int d\mathbf{r}' d\mathbf{v}' \frac{\bar{\mathcal{F}}}{r_{12}}, \quad (18)$$

where G is the gravitational constant

$$r_{12} = [r^2 + r'^2 - 2rr' \cos(\varphi' - \varphi)]^{1/2}.$$

Changing in formula (18) from the variables $\mathbf{r}'$, $\mathbf{v}'$ to $\mathbf{J}'$, $\bar{\mathbf{w}}'$, using the fact that $d\mathbf{r}'d\mathbf{v}' = d\mathbf{J}'d\bar{\mathbf{w}}'$, we obtain

$$\Phi(\mathbf{J}, \bar{w}_1) = -G \int d\mathbf{J}' d\bar{\mathbf{w}}' \frac{\bar{\mathcal{F}}(\mathbf{J}') \exp[im\delta\bar{w}_2]}{r_{12}}, \quad (19)$$

where $\delta\bar{w}_2 \equiv \bar{w}_2' - \bar{w}_2$. Finally, we average the potential Φ over $\bar{w}_1$ to obtain the integral equation

$$\bar{\Phi}(\mathbf{J}) = \frac{G}{2\pi} \int d\mathbf{J}' \Pi(\mathbf{J}, \mathbf{J}') \frac{\mathcal{F}_0'(\mathbf{J}')}{\Omega_p - \Omega_{pr}(\mathbf{J}')} \bar{\Phi}(\mathbf{J}'), \quad (20)$$

where

$$\Pi(\mathbf{J}, \mathbf{J}') = \int d\bar{w}_1 d\bar{w}_1' d\delta\bar{w}_2 \frac{\exp(im\delta\bar{w}_2)}{r_{12}}. \quad (21)$$

The coordinates r and φ of the star in (21) must be written in terms of $\mathbf{J}$, $\bar{\mathbf{w}}$. The radius $r = r(\mathbf{J}, \bar{w}_1)$ is determined by the solution of the equation

$$\bar{w}_1(r, \mathbf{J}) \quad (22)$$

$$= \Omega_1 \int\limits_{r_{\min}(\mathbf{J})}^{r} \frac{dr'}{\sqrt{2[E(\mathbf{J}) - \Phi_0(r')] - L^2/r'^2}},$$
$$\text{if} \quad 0 \le w_1 \le \pi,$$
$$2\pi - \bar{w}_1(r, \mathbf{J})$$
$$= \Omega_1 \int\limits_{r_{\min}(\mathbf{J})}^{r} \frac{dr'}{\sqrt{2[E(\mathbf{J}) - \Phi_0(r')] - L^2/r'^2}},$$
$$\text{if} \quad \pi < w_1 \le 2\pi.$$

The slow variable $\bar{w}_2$ is related to the azimuth φ as

$$\bar{w}_2 = \varphi + \varphi_1(\mathbf{J}, \bar{w}_1), \quad (23)$$

where

$$\varphi_1(\mathbf{J}, \bar{w}_1) \quad (24)$$
$$= \Omega_{pr}(\mathbf{J}) \int\limits_{r_{\min}(\mathbf{J})}^{r} \frac{dr'}{\sqrt{2E(\mathbf{J}) - 2\Phi_0(r') - L^2/r'^2}}$$
$$- L \int\limits_{r_{\min}(\mathbf{J})}^{r} \frac{dr'}{r'^2\sqrt{2E(\mathbf{J}) - 2\Phi_0(r') - L^2/r'^2}}.$$

The primed quantities r' and φ' are expressed by similar formulas.

We now use the obvious symmetry properties of the orbit,

$$r(2\pi - \bar{w}_1) = r(\bar{w}_1), \quad (25)$$
$$\varphi_1(2\pi - \bar{w}_1) = \pi - \varphi_1(\bar{w}_1),$$

to show that the function Π can be written in the form

$$\Pi(\mathbf{J}, \mathbf{J}') = 8 \int\limits_0^{\pi} d\bar{w}_1 \cos m\varphi_1 \int\limits_0^{\pi} d\bar{w}_1' \cos m\varphi_1' \psi(r, r'), \quad (26)$$

where

$$\psi(r, r') = \int\limits_0^{\pi} d\alpha \frac{\cos m\alpha}{\sqrt{r^2 + r'^2 - 2rr' \cos\alpha}}. \quad (27)$$

It is obvious from (26) that Π is real.

It is remarkable that the integral equation (20) written in terms of the function $\bar{\mathcal{F}}$ has the form of a classical eigenvalue problem, where the eigenvalues are equal to the angular velocities Ω_p of the modes:

$$\Omega_p \bar{\mathcal{F}}(\mathbf{J}) = \int d\mathbf{J}' K(\mathbf{J}, \mathbf{J}') \bar{\mathcal{F}}(\mathbf{J}'), \quad (28)$$

where the kernel is

$$K(\mathbf{J}, \mathbf{J}') = \frac{G}{2\pi} \mathcal{F}_0'(\mathbf{J}) \Pi(\mathbf{J}, \mathbf{J}') + \Omega_{pr}(\mathbf{J}) \delta[\mathbf{J} - \mathbf{J}']. \quad (29)$$

[3] Lynden-Bell [3] used a similar derivative of the precessional velocity of stellar orbits, $\partial\Omega_{pr}/\partial L$.

Equation (28) is the master integral equation of our theory. Since the integration on the right-hand side is performed over the variables J_f and L, only unstable ($\mathrm{Im}\ \omega > 0$) or neutral modes can be correctly computed with (28), provided that $\Omega_p > 0$ lies outside the variation interval for Ω_{pr}.

Like the general integral equations of Kalnajs [1] and Shu [2], Eq. (28) can be used to compute the eigenfrequency and the corresponding form of the mode. The most important advantage of our equation is that it gives insight into the physical processes leading to the formation of structures in galaxy disks. These physical mechanisms would be difficult to identify using Kalnajs's and Shu's integral equations or N-body simulations. This is quite clear from the current situation in the theory of the formation of galactic structures (see the Conclusion).

The physical picture becomes more transparent if we derive an integral equation equivalent to (28) in a different way, using explicitly the fact that each orbit as a whole participates in the slow perturbations in which we are interested. Consider now a distribution function for the closed precessing orbits,

$$f_{tot}(\mathbf{J}, \alpha, t) = f_0(\mathbf{J}) + f(\mathbf{J}, \alpha, t), \qquad (30)$$
$$|f| \ll |f_0|,$$

such that $d\mathcal{M} = f dJ_f dL d\alpha$ is the perturbed mass of stars in orbits in the interval $d\alpha$, where α denotes the azimuth of the minor axis, so that its precessional velocity is $\Omega_{pr}(\mathbf{J}) = \dot{\alpha}$. We will assume that this mass is distributed in proportion to the time spent by the particle at a given point of the orbit[4]; i.e., the mass element located in an orbit $\mathbf{J}$ with minor-axis azimuth α is

$$d\mu(\mathbf{r}) = d\mathcal{M} d\gamma, \qquad (31)$$
$$d\gamma = \frac{dt}{T_1} = \frac{\Omega_1(\mathbf{J})}{2\pi} dt = \frac{d\bar{w}_1}{2\pi}.$$

Let $\bar{\Phi}$ be the potential produced by all other orbits of the system and averaged along the chosen orbit:

$$\bar{\Phi}(\mathbf{J}, \alpha, t) = \int d\gamma \Phi(\mathbf{r}) \qquad (32)$$
$$= -G \int d\gamma \frac{d\mu(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}$$
$$= G \int \frac{d\bar{w}_1}{2\pi} dJ'_f dL' d\alpha' \frac{d\bar{w}'_1}{2\pi} \frac{f(\mathbf{J}, \alpha, t)}{|\mathbf{r} - \mathbf{r}'|}.$$

Formula (32) uses the same notation $\bar{\Phi}$ for the potential as does (20), since it is the same quantity in both cases. To make this clear, we must take into account the fact that $f, \bar{\Phi} \propto \exp(-i\omega t + im\alpha)$, and prove that α coincides with $\bar{w}_2$. To this end, we transform (23) into the form

$$\bar{w}_2 = \varphi - \Delta\varphi + \Omega_{pr}\Delta t, \qquad (33)$$

where Δt is the time required for the star to turn through an angle $\Delta\varphi$ between the minor-axis azimuth and the current azimuth, φ. The identity $\bar{w}_2 = \alpha$ then follows from Fig. 3.

The collisionless kinetic equation for such a distribution function is

$$\frac{df_{tot}}{dt} = \frac{\partial f_{tot}}{\partial t} + \Omega_{pr}\frac{\partial f_{tot}}{\partial \alpha} + M\frac{\partial f_{tot}}{\partial L} = 0, \qquad (34)$$

where we have used the facts that $(\partial f_{tot}/\partial J_f)\dot{J}_f = 0$ for the slow modes in which we are interested and $\dot{L} = M$, where M is the torque acting on an orbit with a given J_f, L, α:

$$M_1 = \int d\gamma [r \times dF] \qquad (35)$$
$$= G \int d\gamma d\mu(\mathbf{r}') \frac{rr' \sin(\delta\varphi)}{|\mathbf{r} - \mathbf{r}'|^3}$$
$$= -G \int d\gamma \frac{\partial}{\partial \delta\varphi} \frac{d\mu(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} = -\frac{\partial \bar{\Phi}}{\partial \alpha},$$

where $\delta\varphi = \varphi - \varphi'$ is the difference of the azimuths of the vectors $\mathbf{r}$ and $\mathbf{r}'$. Note that the distribution function (30) and kinetic equation (34) were proposed earlier by Polyachenko [9]. We now linearize (34) to obtain an equation analogous to (15):

$$-i(\omega - m\Omega_{pr})f = -M\frac{\partial f_0}{\partial L}. \qquad (36)$$

We finally use (32) for the two-dimensional potential of the disk to obtain an integral equation that coincides with (28).

3. GENERAL ANALYSIS OF THE MASTER INTEGRAL EQUATION

It follows from (26) that the function Π is real and symmetric: $\Pi(\mathbf{J}, \mathbf{J}')^* = \Pi(\mathbf{J}, \mathbf{J}'), \Pi(\mathbf{J}, \mathbf{J}') = \Pi(\mathbf{J}', \mathbf{J})$. It follows from (15) that $\bar{\mathcal{F}} \propto \mathcal{F}'_0$, and we can therefore divide both sides of (28) by $\mathcal{F}'_0$, multiply them by $\bar{\mathcal{F}}^*$, and integrate over $\mathbf{J}$. We then compute the imaginary part of the resulting equation to obtain

$$\mathrm{Im}\,\Omega_p \int d\mathbf{J} \frac{|\bar{\mathcal{F}}|^2}{\mathcal{F}'_0} = 0. \qquad (37)$$

Formula (15) can be used to reduce (37) to the form

$$\mathrm{Im}\,\Omega_p L_m = 0, \qquad (38)$$

[4] Arnold [16] has pointed out that Gauss suggested long ago to spread the mass of each planet along its orbit in proportion to the time spent at each point and substitute the attraction of such rings for that of the planets.

where L_m denotes the angular momentum of the mode (or, to be more precise, the quasi-momentum—see, e.g., [17]), which can be obtained from the general formula of Lynden-Bell and Kalnajs [15] by dropping all terms except the one that dominates for the low-frequency modes:

$$L_m = -\int d\mathbf{J}\frac{|\bar{\mathcal{F}}|^2}{\mathcal{F}_0'} = -\int d\mathbf{J}\mathcal{F}_0'\frac{|\bar{\Phi}|^2}{|\Omega_p - \Omega_{pr}|^2}. \quad (39)$$

The type of solution of the master integral equation depends crucially on the behavior of the derivative $\mathcal{F}_0'$. We consider two cases below.

(1) Let us suppose that $\mathcal{F}_0'$ is strictly positive everywhere in the phase space of the system. In this case, L_m is negative, like the energy E_m of the mode, since $E_m = \Omega_p L_m$ [15]. We therefore find from (38) that $\mathrm{Im}\Omega_p = 0$. The corresponding real eigenfunctions $\bar{\mathcal{F}}$ describe nonspiral solutions. These solutions represent the central parts of bar modes; i.e., the bars proper.

It is obvious that, in this case, when $\mathcal{F}_0' > 0$, the integral equation (28) determines only the angular velocity $\mathrm{Re}\Omega_p$ of the mode. The amplification of this mode is due to the exchange of angular momentum with resonance stars at the corotation radius and the OLR.[5]

We now estimate the corresponding rate of the mode amplification[6] using the formula

$$\gamma = \dot{L}_m/2L_m, \quad (40)$$

where $\dot{L}_m = \dot{L}_m^{(1)} + \dot{L}_m^{(2)}$, and we derive formulas for the rate of exchange of angular momentum at the corotation radius ($\dot{L}_m^{(1)}$) and the OLR ($\dot{L}_m^{(2)}$) from the general formulas of Lynden-Bell and Kalnajs [15] by slightly transforming these relations:

$$\dot{L}_m^{(l)} = -\frac{1}{4\pi}\int\left(\frac{l}{2}\frac{\partial\mathcal{F}_0}{\partial J_f} + \frac{\partial\mathcal{F}_0}{\partial L}\right)|\Phi_l|^2 \quad (41)$$
$$\times\, \delta[\Omega^{(l)}(\mathbf{J}) - \Omega_p]d\mathbf{J},$$

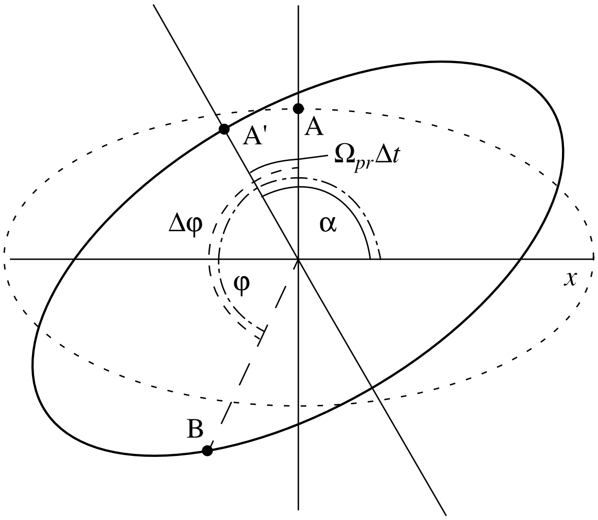

Fig. 3. Illustration of Eq. (33) illustrating the identity $\bar{w}_2 = \alpha$. The dashed and solid ovals show the orientations of the orbit at $t = 0$ and $t = \Delta t$, respectively; during time Δt, the orbit turns through the angle $\Omega_{pr}\Delta t$. The initial azimuth of the star (at $t = 0$) coincides with the azimuth of the minor axis (point A). Point B shows the position of the star at time $t = \Delta t$; the azimuth of this point is $\varphi = \widehat{xB}$ (dot—dashed arc), and the change in the azimuth is $\Delta\varphi = \widehat{AB}$ (dashed arc).

where $\Omega^{(l)}(\mathbf{J}) \equiv \Omega_2(\mathbf{J}) + (l-1)\Omega_1(\mathbf{J})/2$ and Φ_l are the Fourier coefficients corresponding to the CR ($l = 1$) and OLR ($l = 2$) terms in the expansion of the perturbed potential (12) in a series in $e^{il\bar{w}_1}$.

The interaction of stars in resonance regions with the gravitational potential of the mode results in spiral responses (see, e.g., [20, 21]).

(2) Let us now suppose that $\mathcal{F}_0'$ becomes negative in some regions of the phase space. We show below, using a generalized Schwartzschild model as an example that, for realistic distribution functions, such regions, if they exist at all, can occupy only a small fraction of the total volume of the phase space (see Section 5 for details). In this case, unstable spiral solutions may develop in addition to the bars considered above. In contrast to the bars, these new modes grow as a result of the inherent "internal" instability of the mode itself. It follows from (38) that, for an unstable mode ($\mathrm{Im}\Omega_p > 0$), the angular momentum $L_m = 0$; i.e., the contributions to L_m provided by regions with opposite signs of the Lynden-Bell derivative $\mathcal{F}_0'$, exactly cancel each other: $L_m = L^+ + L^- = 0$. The instability criterion $\mathcal{F}_0' < 0$ exactly coincides with the condition for the excitation of waves at the ILR as derived by Lynden-Bell and Kalnajs [15]. Note, however, that these authors assumed that $\mathcal{F}_0' > 0$ throughout the entire phase space. Therefore, negative-energy waves at the ILR must have decayed.

[5] Strictly speaking, if the disk is immersed in real (and not passive, as is most often assumed) nonplanar components (a halo), we must also take into account the resonance exchange of angular momentum between the bar mode and stars of the components considered. Polyachenko and Shukhman [18] were the first to analyze dynamical friction due to resonance interactions of stars of the spherical system with a wave.

[6] The estimate given below is only a rather crude approximation. For example, strictly speaking, Φ_l in (41) is the Fourier transform of the perturbed potential, which includes, in addition to the bar potential, the potential of the spiral density wave excited at the resonance (see, e.g., [19]). Moreover, the components of the potential not included in the first-order perturbation theory may prove to be important for computing the increment. See Appendix I for a more detailed discussion of these issues.

The increment of the unstable mode can be estimated as

$$\gamma = \frac{\dot{L}^+}{2L^+} = \frac{\dot{L}^-}{2L^-}, \tag{42}$$

where

$$\dot{L}_\pm = -\frac{1}{4\pi} \int\limits_{\Gamma_\pm} \mathcal{F}_0' |\bar{\Phi}|^2 \delta[\Omega_{pr}(\mathbf{J}) - \Omega_p] d\mathbf{J}, \tag{43}$$

and Γ_+ and Γ_- denote the phase-space domains with positive and negative values of the Lynden-Bell derivative $\mathcal{F}_0'$.

4. BARLIKE AND SPIRAL SOLUTIONS FOR TEST MODELS

We solved the master integral equation (28) numerically. We considered both the unknown function $\mathcal{F}(\mathbf{J})$ and the kernel $K(\mathbf{J}, \mathbf{J}')$ on a 31×31 network in the phase space (E, L). The resulting matrix equation can then be solved using standard methods of linear algebra. See Appendix II for a more detailed description of the specific features of the numerical algorithms employed.

We will now demonstrate the potential of the theory using test models as examples. We analyzed about ten models investigated earlier by AS using N-body simulations. The results of the computations agreed well in all cases: the accuracy of the computed angular velocities of the modes was better than 10%. We report below the results of our mode computations for two typical models displaying different behavior of the Lynden-Bell derivative of the distribution function.

The unperturbed potential for all the models described by AS [6] has the form of a Plummer potential:

$$\Phi_0(r) = -GM(1 + r^2/b^2)^{-1/2}, \tag{44}$$

where M and b are the mass of the galaxy and the scale length, respectively. The total potential is equal to the sum of the potentials of the disk and the passive spherical component. The surface density of the disk is

$$\sigma_0(r) = \frac{Mq}{2\pi b^2}(1 + r^2/b^2)^{-3/2}, \tag{45}$$

where q is the ratio of the disk mass to the total mass M. If $q < 1$, the system has a passive halo with volume density

$$\rho_0(r) = \frac{3M(1-q)}{4\pi b^3}(1 + r^2/b^2)^{-5/2}, \tag{46}$$

and the same scale length b as in the disk. Such a halo produces a Plummer-type potential like the potential produced by the disk.

The series for the distribution functions for the models we consider here can be written in the form of the series used by AS, and depends on two parameters: a positive integer[7] m and a real parameter β. In the case of a disk at rest, this series has the form

$$f_0(e_1, x) = \frac{\exp\beta}{2\pi^{3/2}} e_1^{m-1} \sum_{k=0}^{\infty} \frac{(-\beta)^k}{k!} e_1^{2k} \tag{47}$$
$$\times \sum_{l=0}^{\infty} \left[\sum_{j=0}^{\infty} \frac{\beta^j}{j!} \begin{pmatrix} \frac{3}{2} - \frac{m}{2} \\ l - j \end{pmatrix} \right] \left(-\frac{1}{4}\right)^l$$
$$\times \frac{\Gamma(2k + 2l + m + 1)}{\Gamma\left(l + \frac{1}{2}\right) \Gamma(2k + 2m + l)} x^{2l},$$

where $e_1 = E/\Phi_0(0)$, $x = -(-2E)^{1/2} L / r_* \Phi_0(0)$, E and L are the energy and angular momentum of the star, respectively, $r_* = \lim\limits_{r \to \infty} r\Phi_0(r)/\Phi_0(0)$, and $\begin{pmatrix} a \\ b \end{pmatrix} = C_b^a$ is the number of permutations of b of a elements. The distribution functions (47) are normalized so that $\int d\mathbf{r} d\mathbf{v} f_0 = 1$. When $\beta = 0$, they become the distribution functions of Kalnajs [22]. Rotation of the disk is produced by introducing another parameter—the angular momentum of the cutoff of retrograde stars, J_c. Thus, the distribution functions under study are

$$f_0^{(J_c)} = \begin{cases} 0, \quad L < -J_c, \\ f_0 p_-, \quad -J_c < L < 0, \\ f_0(1 - p_-), \quad 0 < L < J_c, \\ f_0, \quad L > J_c, \end{cases} \tag{48}$$

where $p_- = (1 - |L|/J_c)^3/2$. Further details (in particular, plots of the rotational velocity, velocity dispersion, etc.) can be found in [6]. Below, like AS, we use units such that $G = 1$, $M = 1$, and $b = 1$; when $q \neq 1$, the gravitational constant G in our master integral equation must be replaced by the product qG.

All Kalnajs's models describe disks with relatively low central radial stellar-velocity dispersions. Kalnajs's model with $m = 6$, $\beta = 0$, $q = 1$, and $J_c = 0.25$ is representative of such disks. From the viewpoint of our theory, the most important property of this model is that its Lynden-Bell derivative is positive everywhere in the phase space.

[7] To facilitate the comparison of the results, we adhere to the notation adopted by AS. That is why m in this section is a parameter of the distribution function and not the azimuthal number, which is fixed and equal to 2.

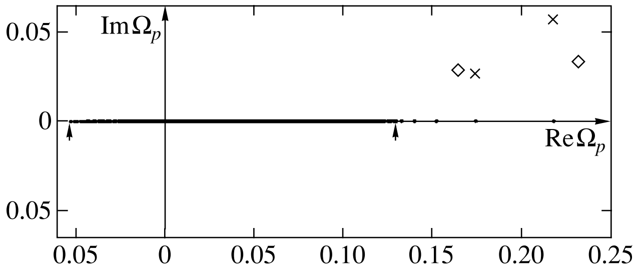

Fig. 4. Spectrum of the angular velocities of the bar (bold dots) computed as eigenvalues of (28) for Kalnajs's model ($m = 6$, $\beta = 0$, $q = 1$, and $J_c = 0.25$). The crosses show the "theoretical" complex velocities of the bar, whose imaginary parts are estimated by (40) and (41). The diamonds denote the "experimental" complex bar velocities according to AS. The arrows indicate the minimum and maximum precessional velocities of the stars.

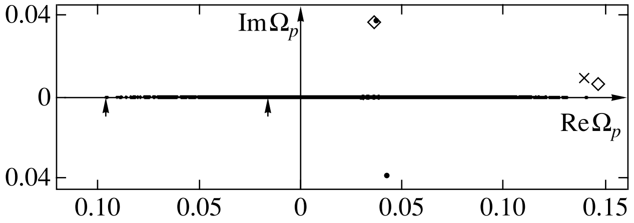

Fig. 5. Spectrum of the complex angular velocities of the bar (bold dots) computed as eigenvalues of (28) for the model with $m = 6$, $\beta = 3$, $q = 1$, and $J_c = 0.6$. The cross shows the "theoretical" complex velocity of the bar with the increment computed using (40) and (41). The diamonds show the "experimental" complex velocities of the mode according to AS. The arrows show the minimum and maximum precessional velocities of the stars.

Figure 4 shows the computed eigenvalue spectrum. It is obvious that the full spectrum includes both discrete and continuous components. All the frequencies are real, due to the positiveness of the Lynden-Bell derivative. The set of frequencies between the minimum and maximum precessional velocities (-0.051, 0.126) approximates a continuous spectrum. These frequencies correspond to modes narrowly localized in the phase space. Five discrete modes can be discerned to the right of the continuous spectrum. When analyzing galactic structures, we will be interested only in these modes. For example, the barlike and spiral structures observed in the N-body simulations of AS developed from discrete modes. The various eigenfunctions of the discrete spectrum correspond to perturbed profiles of the surface brightness and potential with different numbers of radial nodes, and the nodeless mode is always the one with the maximum bar velocity.

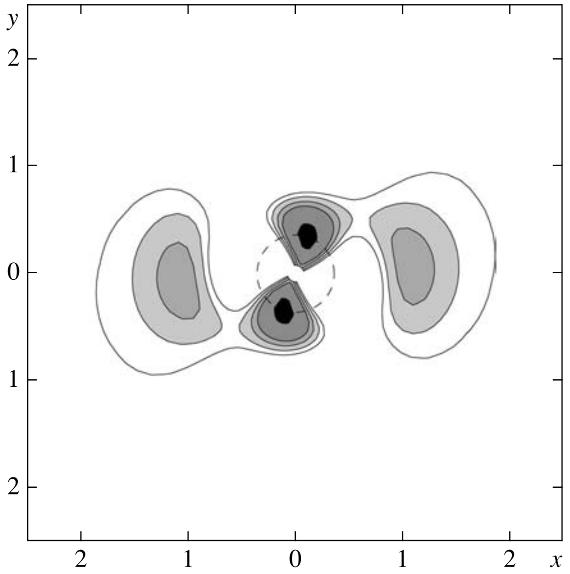

Fig. 6. Pattern of spirals for the unstable mode in the model with $m = 6$, $\beta = 3$, $q = 1$, and $J_c = 0.6$ of [6]. The dashed circle indicates the position of the inner ILR (the outer ILR for this model is located at $r = 5$).

The bar velocities for the first and second modes (counting from the right) derived by solving (28) agree with the velocities obtained in the N-body simulations to within 6%. The other three modes have substantially lower increments, and could not be observed in the numerical simulations.

We roughly estimated the increments for the fastest two modes using (40), assuming that the stars move in nearly circular orbits. The total amplification increments $\gamma_1 = 0.1167$ and $\gamma_2 = 0.0544$ are sums of two components, corresponding to corotation and the OLR: $\gamma_{1\mathrm{CR}} = 0.0199$, $\gamma_{1\mathrm{OLR}} = 0.0968$; $\gamma_{2\mathrm{CR}} = 0.0107$, $\gamma_{2\mathrm{OLR}} = 0.0437$. It is clear that the OLR makes the dominant contribution to the amplification. We show in Appendix I that this is quite natural for disks with nearly circular orbits. The mode with the maximum bar velocity also has the highest increment. This is obviously due to the fact that this mode has the smallest radii for corotation and the OLR, so that the disk surface brightness is still fairly high at these radii. The estimate of the amplification increment for the fastest mode obtained above is about twice the value derived by AS from their N-body simulations, whereas the "experimental" and "theoretical" increments for the second mode are fairly close to each other.

In the second model, whose parameters in the notation of AS are $m = 6$, $\beta = 3$, $q = 1$, and $J_c = 0.6$, the unperturbed distribution function describes a stellar disk with a fairly high central radial-velocity dispersion. From the viewpoint of our theory, the most important difference between this model and the one

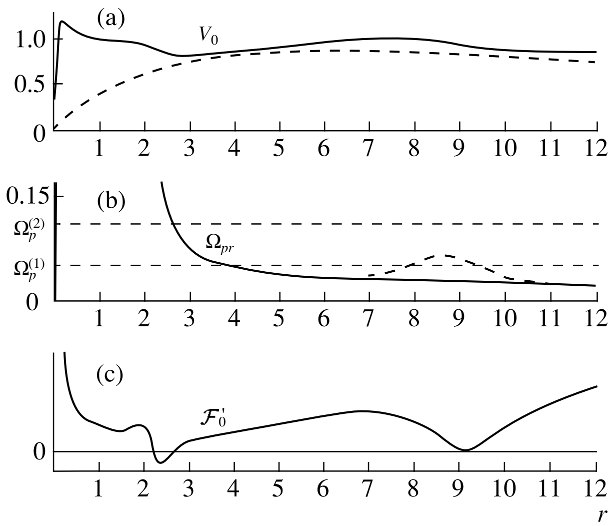

Fig. 7. (a) Rotation curve of the type displayed by our Galaxy (solid); the dashed curve corresponds to the contribution of the exponential disk. (b) Precessional velocities Ω_{pr} of circular orbits for the standard model (solid) and a model with a small "peak" near $r = 8{-}10$ (dashed). The horizontal lines correspond to various velocities of the spiral pattern (pattern speeds). (c) Lynden-Bell derivative $\mathcal{F}_0'^{(m_0)}$ for circular orbits.

considered above is that its Lynden-Bell derivative is not everywhere positive in the phase space.

Figure 5 shows the computed eigenfrequency spectrum for this model. As in the previous example, the complete spectrum has both a continuous and a discrete part.

Since the Lynden-Bell derivative does not have the same sign throughout the phase space of the model, the discrete spectrum contains complex frequencies, which correspond to unstable spiral modes. The existence of these modes is associated with the fact that the Lynden-Bell distribution function has different signs in different regions of phase space. The interpretations suggested earlier (see [6, 10]) were inaccurate.

We see in Fig. 5 one unstable, complex and one real frequency. The complex frequency obtained by solving the master integral equation is very close to the "experimental" frequency determined by AS. Figure 6 shows the spiral pattern corresponding to this mode. The poor resemblance between the mode obtained and the observed spirals can be explained by our use of the distribution functions (47), which are not fully adequate for describing the spiral pattern (as was already pointed out by AS). It is also evident that the presence of such short and open spirals is due to the fact that the instability increment is comparable to the real part of the frequency (this relation, likewise, has no particular practical meaning). Both a spiral mode and a bar are excited in the numerical experiment described by AS, and it is the latter that survives in the competition between the two, while the spiral mode gradually decays and disappears. Note, however, that this scenario is by no means inevitable. In particular, there may well exist situations where the spiral mode is the only mode possible (e.g., in the case of disks with sufficiently low masses).

The velocity of the pattern of the real mode agrees with the velocity of the bar obtained in the N-body simulations to within 3%. A crude estimate of the mode increment in the approximation of nearly circular orbits yields $\gamma = 0.018$ and, as in the first model, the OLR is the main contributor. This estimate is only 30% higher than the γ given by AS.

Despite the somewhat artificial nature of the models considered above, our analysis points toward a general mechanism for bar formation. Moreover, scenarios for bar formation can be derived proceeding exclusively from the general properties of our master integral equation (see Section 3 and the Conclusions).

5. ANALYSIS OF THE LYNDEN-BELL DERIVATIVE FOR A TYPICAL MODEL OF THE GALACTIC DISK

Consider the generalized Schwartzschild distribution function [2]

$$f_0(E, r_0) = \frac{2\Omega(r_0)}{\kappa(r_0)} \frac{\sigma_0(r_0)}{2\pi c_0^2(r_0)} \exp\left(-\frac{E - E_c(r_0)}{c_0^2(r_0)}\right), \tag{49}$$

where E is the energy of the star, r_0 the radius of the guiding center $(L = r_0^2\Omega(r_0))$, $\kappa = (4\Omega^2 + r, d\Omega^2/dr)^{1/2}$ is the epicyclic frequency, $E_c(r_0) = v_0^2(r_0)/2 + \Phi_0(r_0)$ is the energy of the star in a circular orbit, $v_0(r_0) = r_0\Omega(r_0)$ is the circular velocity, and $\Phi_0(r_0)$ is the equilibrium potential. The specific model is specified by the functions $\sigma_0(r_0)$ and $c_0(r_0)$. In the epicyclic limit, when $v_0/c_0 \gg 1$, $\sigma_0(r_0) = \Sigma_0(r_0)$ and $c_0(r_0) = c_r(r_0)$, where $\Sigma_0(r_0)$ and $c_r(r_0)$ are the disk surface density and the radial-velocity dispersion, respectively. In the general case, $\Sigma_0(r_0)$ and $c_r(r_0)$ can be expressed in terms of $\sigma_0(r_0)$ and $c_0(r_0)$ in a more complex way.

The Lynden-Bell derivative of the distribution function (49) is

$$\frac{\partial \mathcal{F}_0}{\partial L} = \frac{2\Omega(r_0)}{r_0^2\kappa^2(r_0)} \mathcal{F}_0 \biggl\{ r_0 \frac{\Omega'(r_0)}{\Omega(r_0)} - r_0 \frac{\kappa'(r_0)}{\kappa(r_0)} \tag{50}$$

$$+ r_0 \frac{\sigma_0'(r_0)}{\sigma_0(r_0)} - r_0 \frac{2c_0'(r_0)}{c_0(r_0)} + \frac{r_0^2\kappa^3}{2m\Omega(r_0)c_0^2(r_0)}$$

$$+ r_0 \frac{2c_0'(r_0)}{c_0^3(r_0)} (E - E_c(r_0)) \biggr\},$$

where m again denotes the azimuthal number, $\mathcal{F}_0(\mathbf{J}) = f_0(E(\mathbf{J}), r_0(L))$, and a prime indicates differentiation with respect to r_0. As a rule, $\mathcal{F}_0' > 0$ in all or nearly all of the phase space. This property is ensured by the term $r_0^3\kappa^3/2m\Omega c_0^2$ in (50). For example, in the case of a flat rotation curve ($v_0 =$ const), this term is equal to $(\sqrt{2}/m)(v_0/c_0)^2$. It is the dominant term, $v_0/c_0 \gg 1$, almost everywhere in disk galaxies.

However, certain narrow domains where this term does not dominate may exist. This is possible in regions where the function $\kappa(r)$ is sufficiently small. For example, for a rotation curve resembling that of our Galaxy, such regions are located at $r \approx 2.5$ and 8–10 kpc. Likely origins of such features in rotation curves include the transition from the potential of the spherical subsystem to that of the disk or a sharp edge of one of the disk components. Second, the function $c_0(r_0)$ increases rapidly toward the galactic center, so that the rotational velocity v_0 in the central regions becomes comparable to c_0.

Let us summarize the main factors affecting the properties of the solutions of the master integral equation. First, the solutions can change substantially due to small variations in the rotation curve. Second, the mass of the disk is also an important factor. This is due to the fact that the self-gravity of the disk ensures that the angular velocity of the mode exceeds the maximum precessional velocity of the orbits. Therefore the solutions for sufficiently massive disks have barlike forms, whereas only spiral modes can be obtained for disks with relatively low masses. A third factor is the importance of the velocity dispersion. It is obvious from (50) that an increase in c_0 decreases the dominance of the positive term $r_0^3\kappa^3/2m\Omega c_0^2$. Finally, the number of highly elongated orbits is of considerable importance. We show in the next section that an excess of the number of such orbits over the number predicted by the generalized Schwartzschild distribution function (49) strongly increases the increments of spiral modes. This comes about because, in this case, (50) implies an increased influence of the negative term $2r_0c_0'(r_0)(E - E_c(r_0))/c_0^3(r_0)$, resulting in an increase in $|\mathcal{F}_0'|$ ($\mathcal{F}_0' < 0$).

6. SPIRAL SOLUTIONS OF THE MASTER INTEGRAL EQUATION

When the Lynden-Bell derivative is negative in some regions of the phase space, a great variety of spiral modes appear—both leading and trailing. In this section, we briefly discuss these solutions, deferring a more detailed analysis to a separate paper.

We adopt the generalized Schwartzschild distribution function described in Section 5 as a basis distribution function for analyzing spiral solutions. By the standard model, we mean a model with a rotation curve similar to that of our Galaxy (Fig. 7a) and exponents $\sigma_0(r_0) = \sigma_0 e^{-r_0/r_d}$, $c_0(r_0) = c_0 e^{-r_0/r_c}$ specifying a particular Schwartzschild function. In our units, 1 kpc corresponds to unity (1), $r_d = 3$, $r_c = 2r_d$, the gravitational constant $G = 1$, and $\sigma_0 = (\pi r_d)^{-1}$.

The disk provides the dominating contribution to the rotation curve at $r_0 > 5$–6. The maximum rotational velocity is $(v_0)_{\max} = 1.2$ and $c_0 \simeq 0.83$, so that $v_0(r_0)/c_0(r_0) \simeq 4$ for $r_0 = 8$. With these parameters, the galactic disk is exponential almost everywhere except in its innermost part. With this rotation curve, the equilibrium potential can be computed as $\Phi_0(r) = \int^r v_0^2(r')/r'dr'$. We thus obtain a model galaxy that resembles our Galaxy (but, of course, is not identical to it[8]).

Figure 7c shows the Lynden-Bell derivative $\mathcal{F}_0'$ for the standard model computed using (50) on the line of circular orbits ($E = E_c(r_0)$). There is only one narrow region with $\mathcal{F}_0' < 0$, located near the center. Figure 8 demonstrates how deeply negative $\mathcal{F}_0'$ values extend into the phase domain of the system.

The frequency spectrum for the standard model has both continuous and discrete parts. The continuous spectrum occupies a wide band of real frequencies, $\Omega_{pr}^{\min} < \Omega_p < \Omega_{pr}^{\max}$. However, we will be interested only in the spiral modes of the discrete spectrum. For the standard model, the growth rates of these modes are low, corresponding to an amplitude increase by a factor of e in about $(3–5) \times 10^9$ yr (given that $(v_0)_{\max} = 1.2$ corresponds to a rotational period of $T = 2.5 \times 10^8$ yr). The main reason for these low growth rates becomes clear from Fig. 8, which shows that the curve $\Omega_{pr}(E, L) = \mathrm{Re}\Omega_p$ (for a typical unstable mode) crosses the band of negative values of the Lynden-Bell derivative, $\mathcal{F}_0'$, deeply inside the phase space of the system considered—where (E, L) corresponds to highly elongated orbits. However, the number of such orbits is small for a Schwartzschild distribution function, which decreases exponentially with $(E - E_c(r_0))/c_0^2(r_0)$.

The growth rates of spiral modes can be made substantially higher if we take into account the fact that there are many more such orbits in real galaxies.

[8] Here, we have in mind structures similar to the spirals of the galaxy NGC 2997, whose image is shown on the cover of the well-known textbook by Binney and Tremaine [13]. Note that the existence of a "grand design" pattern in our Galaxy remains an open question.

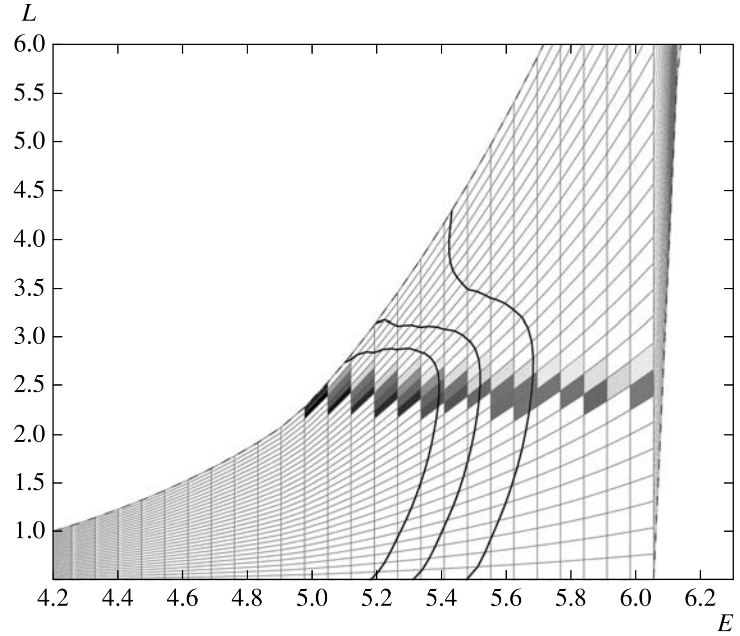

Fig. 8. The band of negative values of the Lynden-Bell derivative $\mathcal{F}_0'$ in the (E, L) phase plane. The phase-space domain of the systems is bounded by the line of circular orbits (above) and the line corresponding to $0.02(\mathcal{F}_0)_{\max}$, where $(\mathcal{F}_0)_{\max}$ is the maximum of the distribution function $\mathcal{F}_0(E, L)$ at each fixed E (right). The solid curves $\Omega_{pr}(E, L) = \Omega_p$ show the positions of resonance orbits in the phase plane for several values of Ω_p.

Of great importance for the modes considered are the central regions of galaxies, which are characterized by a substantial contribution from nonplanar components with radially elongated velocity diagrams. This also indicates that the spherical component may play an important active role, not just the passive role of an unperturbed halo that is usually ascribed to it. Computations made using an appropriately modified standard model yield a tenfold increase in the growth rate of spiral modes. The modification consists of the substitution $c_0(r_0) \to 2c_0(r_0)$ under the condition that $(E - E_c(r_0)) > c_0^2(r_0)$. This modification has virtually no effect on the observed surface brightness and the radial-velocity dispersion of the stars. It is also important that this modification results in the dominance of trailing spiral modes. Observed galactic spirals probably correspond to such modes (or to similar modes in other modifications of the standard model). These modes usually have the form of half-turn spirals, as in most real galaxies. Figure 9 shows the eigenfrequency spectrum for the modified model.

One interesting modification of the standard model is the formation of a small "peak"[9] in the Ω_{pr} curve at fairly large r, where the Lynden-Bell derivative can be either positive or negative. We do not describe all possibilities here, and note only that, in this case, the patterns differ substantially for the spiral modes with angular velocities $\Omega_p^{(1)}$ and $\Omega_p^{(2)}$ shown in Fig. 7b: in the first case, we have one ILR, whereas there are three such resonances in the second case. Figure 10 shows the typical pattern of a spiral mode for $\Omega_p = \Omega_p^{(1)}$.

7. CONCLUSIONS: SUMMARY AND COMPARISON WITH PREVIOUS WORKS

In our theory, galactic structures are viewed as normal modes of the stellar disk. In this respect, our theory has its roots in the pioneering work of Lindblad (see, e.g., [23]). A similar approach was adopted in the classic studies of Lin and Shu [24], Lin *et al.* [7], Kalnajs [25], and many others. The most advanced version of this theory has been developed in the recent work of Bertin and Lin [26]. However, these authors compute modes of an "effective" gaseous disk, not those of the original stellar disk. As a result, most of the specific features that distinguish the dynamics of the stellar systems are lost, so that the results obtained in these papers have only a limited applicability to real galaxies.

The main distinctive feature of our approach is that we suppose that low-frequency modes whose angular velocities are of the order of the precessional velocity of the stellar orbits are of greatest importance for describing galactic structures. This assumption, which we have tried to substantiate in detail above

[9] In the case of our Galaxy, such a peak was first pointed out in the well-known paper of Lin *et al.* [7].

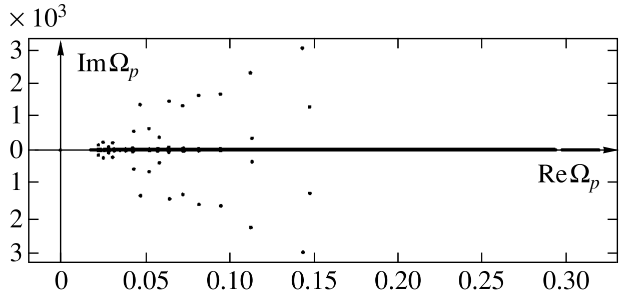

Fig. 9. Spectrum of complex angular velocities for a modified Schwartzschild model. Also shown are the extraneous roots with Im $\Omega_p < 0$.

(Section 1), is crucial for all the simplifications in our theory. As a result, we were able to derive a simple integral equation for the low-frequency modes of the stellar disk. The properties of the solutions of this equation are determined by the behavior of the Lynden-Bell derivative of the distribution function $\mathcal{F}_0'$ in the phase space of the disk. The solutions have barlike or spiral shapes depending on whether $\mathcal{F}_0'$ is positive everywhere in the phase domain or whether there are narrow regions with negative values of this derivative.

Below, we give a detailed critical analysis of various other approaches used to explain the formation of galactic structures. We analyze bars and spirals separately, considering bars in more detail. We restrict our analysis of spirals to a few introductory comments, deferring a more detailed study to other papers.

7.1. Formation of Galactic Bars

In our theory, a bar mode develops as a result of azimuthal tuning of the orbits. The allowed frequency (angular velocity) of the bar is equal to one of the eigenvalues of the master integral equation. This means that the corresponding waves remain unchanged over many galactic rotations, despite differences in the precessional velocities of the orbits. The mode grows due to the exchange of angular momentum with stars at the corotation resonance and OLR. Half-turn spirals adjacent to the bar also develop as a result of resonance interactions.

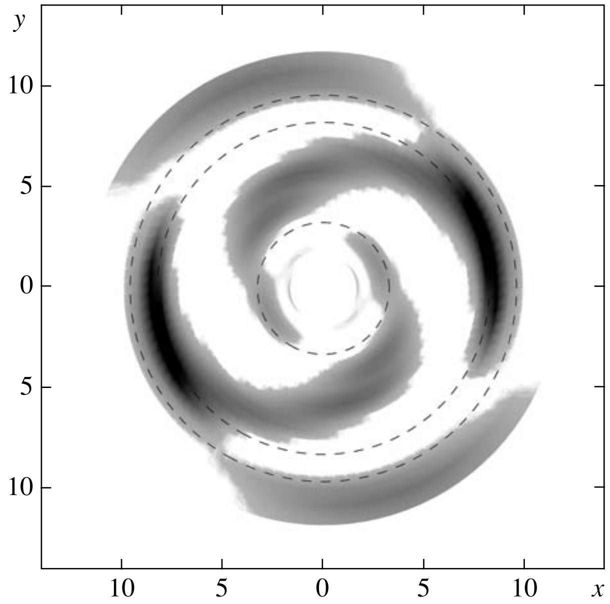

Fig. 10. Pattern of one of the typical spiral modes whose angular velocity corresponds to $\Omega_p^{(1)}$ in Fig. 7b. The dashed circles indicate the positions of the three ILRs.

The mechanism of bar formation considered here imposes a natural constraint on the angular velocity of the corresponding modes: $\Omega_p > (\Omega_{pr})_{\max}$. Otherwise, inner Lindblad resonances appear, where the mode should decay (at $\mathcal{F}_0' > 0$). The excess of Ω_p over $(\Omega_{pr})_{\max}$ is provided by the self-gravitation of the system, so that bars can form only in galaxies with sufficiently massive disks. If the mass of the disk is small, specific numerical computations imply that only a continuous spectrum exists at $\mathcal{F}_0' > 0$, while an unstable spiral mode develops if the phase space contains a region with negative $\mathcal{F}_0'$. Note that the inequality $\Omega_p > (\Omega_{pr})_{\max}$ becomes difficult to satisfy for disks with low masses and high central mass concentrations.

After this brief summary of our results, we now turn to a critical analysis of various approaches to studying bar formation that have been discussed earlier in the literature. A comparatively recent review of possible bar-formation mechanisms is given by Lynden-Bell [27], who lists six different approaches:

(1) Attempts to associate observed stellar bars with classical figures of a gravitating, incompressible Jacoby and Riemann fluid.

(2) Representing the bar as a pair of oppositely propagating density waves with almost equal intensities, which are amplified via Toomre's swing mechanism.

(3) Tidal excitation of the bar as a result of close encounters with other galaxies.

(4) Formation of bars as a result of the deformation of circular orbits, which turn into rotating ovals (Contopoulos's approach).

(5) Instability related to the well-known instability of radial orbits, which results in the formation of slow Lynden-Bell bars.

(6) An unconventional (and so far poorly developed) statistical approach that considers the formation of barlike structures from rotating initial configurations when their energy is lower than some critical limit.

Leaving the last of these possibilities without comment, we give here brief explanations of the first and third approaches as a prelude to discussing the remaining approaches in more detail.

Although the direct use of incompressible liquid figures to represent collisionless (stellar) systems is clearly not an optimal approach, it was believed for a long time that bars form when the galactic disks rotate sufficiently rapidly, as happens with classical, incompressible MacLaurin spheroids that are strongly oblate (and, consequently, have very high rotational velocities). However, Toomre [11] noted that real galactic bar modes have little in common with incompressible "edge" modes, and therefore must have a different formation mechanism.

Tidal interaction of close galaxies may result in the formation of bars. However, Lynden-Bell [27] pointed out that barred galaxies are much more numerous than is predicted by this mechanism. It follows that we must look for internal mechanisms for the formation of barlike structures in galaxies.

7.1.1. Slow Lynden-Bell bars. Our theory is closest in spirit to Lynden-Bell's [3] theory of slow bars, which also has to do with slowly precessing orbits. According to Lynden-Bell, the alignment of orbits is possible if $\partial\Omega_{pr}(\mathbf{J})/\partial L > 0$. Polyachenko [9] showed that this inequality is actually a necessary condition for instability similar to the instability of radial orbits (instability develops only if the dispersion of the precessional velocities of the orbits is not too high).

Note that Kalnajs [28] has already pointed out the possibility of the alignment of freely precessing orbits (and, in particular, of the formation of a bar via this process) in his theory of kinematic waves. In the theory of Kalnajs, who considered the case of nearly circular orbits, it is very important that the precessional velocity $\Omega_{pr}(r)$ be independent of the radius. Lynden-Bell [3] likewise requires that $\Omega_{pr}(r)$ be approximately constant in his more general treatment of the problem.

According to Lynden-Bell, the bar must rotate at some average precessional velocity for its constituent orbits. In contrast, we assume that bars (like spirals) can move faster than the medium, fully analogous to numerous other wave phenomena. Numerical computations of bar-rotation velocities via N-body simulations yield results that agree with the predictions of our mechanism (and not that of Lynden-Bell).

We consider the disk to consist of a large number of slowly deforming and rotating quasi-elliptical orbits. Compared to the motion of the orbits, the motions of individual stars are so rapid that they can be considered to be "spread" along the orbit. Small variations in the shape and orientation of the orbits lead to the development of regions of high surface density. In the absence of self-gravitation, any region with an excess surface density winds up, since the part of the pattern located at radius r should precess at a velocity $\Omega_{pr}(r)$, which in reality varies strongly with r. Individual orbits change their mutual positions under the action of gravity. The slow variations of the angular momenta and the radial actions of orbits due to these interactions result in deformations of the orbits themselves and changes in the surface-density distribution. Such variations can display wavelike behavior and generate a pattern that rotates with an angular velocity that exceeds the maximum precessional velocity of the orbits in the disk.

7.1.2. Contopoulos's theory. The properties of families of periodic orbits in rotating barred potentials found by Contopoulos [29] are of fundamental importance for the theory of galactic bars. The so-called x_1 family of orbits, which are elongated along the bar inside the corotation circle, plays a central role. The theory assumes (and this appears to be quite reasonable) that these periodic orbits and nonperiodic orbits close to them are the building blocks of the figures of galactic bars. However, these statements refer only to already formed bars and, strictly speaking, are not directly related to the bar-formation mechanism. This is particularly true of the linear stage of bar-forming instabilities. It is clear that the orbits considered by Contopoulos are captured in the potential of the bar, but the capture process is obviously a nonlinear phenomenon. However, it is common practice to analyze the pattern of a bar's constituent orbits described above to draw conclusions about mechanisms for bar formation that seem natural at first glance. It is believed that a growing bar forces circular orbits to change their shapes to adjust to the increasingly stronger and thinner bar [30, 31] (see also [27] and the end of [3]). The possible importance of this effect even in the linear stage is justified in the theory of weak bars (see, e.g., [32]), where a linear perturbation theory is used to show that the bar potential changes initially circular orbits into weakly oblate or prolate ovals whose orientations relative to the bar are the same as those of the general orbits of the x_1 family.

However, in reality, circular orbits cannot be typical representatives of the orbits of an unperturbed disk (except in "cool" disk models, which are of little interest for the problem we are considering), before the velocity perturbations due to the action of the bar exceed the velocity dispersion in the initial axisymmetric disk. It is clear that this is true, first and foremost, of the initial stage of the development of the instability (starting from a fairly low level).

The eccentricity of the orbits participating in the perturbation is important for the bar modes considered here: it is the deviation of an orbit from circularity that produces the irregularities that are "hooked" by the axisymmetric gravitational field of the mode, thereby producing a torque. In turn, the eccentricity of the orbits is determined primarily by the radial-velocity dispersion of the stars. Here, we can clearly see the flaws of the common approach of simulating the velocity dispersion using softened gravity [11, 33, 34]; i.e., using an effective decrease of the gravitational constant G while preserving purely circular motion of the stars. Since circular orbits have no irregularities (these appear only when the orbits are perturbed), it is clear that such simulations cannot be valid[10]: in this case, all modes of the type considered here would be lost. The only phenomenon that can be correctly described qualitatively in this way is the stabilization of the Jean's instability of the disk by the stellar velocity dispersion. We give a detailed quantitative analysis of these problems in Appendix III.

7.1.3. Swing amplification mechanism. The pattern of global modes considered here leaves virtually no room for the currently popular swing amplification mechanism of Toomre [11] or related models, such as over-reflection (the waser mechanism of Mark *et al.* [36]).[11] Recall that a swing description of global modes in galaxies can be divided into two parts. The first is the mechanism of swing amplification, which has been known since the classic works by Goldreich and Lynden-Bell [37] and Julian and Toomre [38]: an initial perturbation in the form of a leading (preferably tightly wound) spiral ultimately becomes a tightly wound trailing spiral with an amplitude higher than its initial amplitude. Quantitatively, the amplification depends on many factors, most importantly, how "hot" the galactic disk is (as measured by the extent to which Toomre's parameter Q exceeds unity). However, for typical conditions ($Q \approx 1.5$), the amplitude is enhanced by a factor of thirty or so.

However, swing amplification alone is not sufficient for a global mode to form, because the leading and trailing wave packets considered in the swing approach propagate radially with a group velocity that is usually estimated from the local dispersion equation of Kalnajs [1] and Lin and Shu [24]. Leading waves move from the center toward the corotation radius, while trailing waves move in the opposite direction. Swing amplification occurs in the corotation region when leading waves are transformed into trailing waves, and is thus a single act for each initial wave packet. Therefore, in order for a global mode to be established, the trailing wave must somehow be reflected from the central region of the galaxy, so that it is transformed into a leading wave and travels back toward the corotation region, where it is again amplified and transformed into a trailing wave, etc.

It follows that the swing approach replaces the direct resonance interaction of the bar mode with the disk stars due to the long-range nature of the mode's gravitation with an interaction with traveling waves [11, 13] (which are transformed at the corotation radius).[12]

Representing a standing wave (bar mode) using two waves traveling in opposite directions is quite justified from a formal mathematical viewpoint. However, from a physical viewpoint, we must consider an intrinsic standing wave due to the angular regrouping of the stellar orbits. Since the bar mode is concentrated in the central region of the galaxy and its amplitude at the corotation radius is very small, the general swing ideology must be supplemented by an explanation of why the waves traveling away from the center decay in a well-defined way (so as to exactly match the radial dependence of the amplitude of the bar mode), while the waves moving toward the center grow in a similar manner. Where can we find such an explanation without returning to the language of global standing waves?

Moreover, in reality, the long-range gravitational field of the bar mode results in a resonance exchange of angular momentum not only at the corotation radius (which is the only active region in the swing mechanism), but also at other resonances (the OLR in the disk and, strictly speaking, at resonances with stars of the nonplanar components of the galaxy). Note that, for example, the effect of the OLR plays the dominant role in some cases (including the specific models considered here—see Section 4).

We also believe that it is not correct to shift the focus onto the problem of wave amplification. We focus our attention not on the mechanism for the amplification of an initially small perturbation, but on how possible mode velocities $\mathrm{Re}\Omega_p$ are selected. The problem considers the degree to which the rotations

[10] If, of course, the system considered is not so "cool" that the velocity dispersion is lower than the perturbed velocities for any perturbations of interest. This may be the case in planetary rings when they are perturbed by satellites, which trigger so-called "wave trains" (see, e.g., [35]).

[11] The essence of the brief critical commentary about this approach given by Lynden-Bell [27] (item 2 in his list of bar-formation mechanisms) is that this mechanism cannot explain the formation of barred, early-type spirals with extended and strong bars and a very small external spiral wave.

[12] The bar modes investigated by Toomre [11] and then discussed by Binney and Tremaine [13] are actually edge modes due to the sharp edges of the Gaussian disks these authors used in their simulations. The rotational velocities of these modes can be computed using the technique we used to analyze bar modes.

of the orbits are differential (or, to be more precise, the degree to which the precessional velocities of the orbits differ). This problem is virtually absent in the case of lasers, which swing analyses consider to be almost direct analogs of galaxies. In the case of lasers, we are dealing with an essentially predetermined frequency (a narrow frequency interval within the bandwidth of the operating transition: only these frequencies are amplified by the active medium). The role of resonators (mirrors) is only to return the wave many times toward the active medium, thereby providing subsequent amplification and adjustment of the frequency of the wave (monochromatization). In a "differentially" rotating disk of orbits, modes (i.e., angular perturbations that are not destroyed by the differential rotation) can develop only at individual frequencies corresponding to certain specific compressions and rarefacations of the initial distribution of orbits in the phase space of the system (their mutual "tuning"), primarily in regions where there are many orbits. It stands to reason that perturbations of the density of the stellar orbits do not need to travel from the center to the corotation radius and back. Thus, in the case of a bar mode, the monochromatic wave (mode) acquires the required amplification at the locations of resonances (in particular, at the corotation radius) due to the long-range nature of the gravitational force; the wave need not travel directly to these resonance regions. It is obvious that the mode with the highest growth increment will be singled out. This is usually the fastest mode, so that its corotation radius and other resonances are closer to the center than are those of other modes.

The unfortunate analogy with lasers is also misleading, since it focuses attention on a "correct" mutual phasing of waves, which, in the swing theory, propagate radially from the corotation radius to the center and back. It is obvious that the real problem instead consists in determining the azimuthal redistribution of the precessing orbits, which results in the formation of a mode that rotates rigidly at some angular velocity Ω_p. The swing explanation of barred-spiral structures could be justified earlier by the lack of an adequate theory of global modes. Now, when we believe we have developed such a theory, the entire complex architecture of the swing pattern, with its wave packets traveling in the galaxy, the need for a special mechanism to reflect the waves, a feedback loop, and phase correspondences, etc., becomes unnecessary.

Athanossoula and Sellwood [6] performed the most detailed attempt to interpret the results of N-body computations of bar modes in terms of Toomre's swing theory, and we accordingly now use their paper to carry out a more detailed critical analysis of the main points of this approach.

In Toomre's theory (as in all cases involving over-reflection), the increment is determined by two factors. The first is the amplification in a single act of the swing transformation of a leading wave into a trailing wave; AS call this quantity the NGF (net growth factor). The second factor is the time τ required for the wave to travel toward the center and return to the initial point, thereby closing the feedback loop. Given the NGF and τ, the mode-growth increment can be calculated using the obvious formula $\gamma = \log \mathrm{NGF}/\tau$.

An (obviously rough) estimate of the gain factor can be obtained in terms of the simplest local theory, which Toomre [11] calls the GLB+LSK theory, corresponding to the first letters of the names of its principal authors—Goldreich, Lynden-Bell, Lin, Shu, and Kalnajs. We must first choose the radius (which, strictly speaking, will be different for different modes) at which to compute the characteristic "shear velocity" Ω' (currently available local theories use the approximation of a constant shear velocity), some typical value of Toomre's parameter Q, and the wavelength of the perturbation. We now simply quote the corresponding passage from AS, since it is impossible to rephrase this passage in any logically consistent way: "It is clear from Toomre's (1981) [11] Fig. 8 that the outgoing wave turns back before reaching corotation, and that its amplitude and pitch angle change continuously with radius. [Therefore] No single radius can be identified as the radius on the cycle, where all amplification occurs. Thus any radius we choose has to be a compromise where the NFG is reasonably representative of the amplitude around the cycle. After experimenting with several choices, we found that the radius where $m(\Omega - \Omega_p) = -\kappa/2$ seems appropriate." This quotation clearly shows the (forced!) level of argumentation.

The arguments used to choose the formula for estimating the NGF appear equally arbitrary: "The maximum growth factor (MGF) obtainable requires an optimal initial phase for the leading wave. Since all other phases produce less amplification, and could even result in a reduced amplitude, Toomre (private communication) now favors (!) $\mathrm{NGF} = (\mathrm{MGF} + 1/\mathrm{MGF})/2$ as a more representative estimate of the actual amplification." It is evident from this passage that the authors found nothing more convincing than to say that Toomre "favors" this estimate.

Estimating the group velocity involves finding the interval of radii where there exist solutions of the Lin–Shu–Kalnajs [7, 1] dispersion equation (which was derived from the WKB theory for short-wavelength perturbations) for each of the bar modes analyzed by AS (which, on the contrary, have maximum wavelengths). This is again a forced procedure, necessitated by the fact that these authors had no other suitable theory at hand. It is therefore quite natural that some of the modes had no appropriate WKB solutions (so that no estimate can be found for these modes). However, AS nevertheless could coarsely estimate (in accordance with Toomre [39]) the group velocity and, consequently, the time τ. AS combine all these results to obtain local estimates for the growth velocities of modes, which are typically about half of the "experimental" increments.

We computed the angular velocities and growth rates of the modes obtained earlier in the N-body simulations of AS in our approach without swing amplification, and found our results to agree well with those of the above authors. The bar mode grows due to the direct effect of its gravitation on the stars located in the vicinity of the resonances, resulting in an exchange of angular momentum between these stars and the bar mode.

AS overlooked this simple possibility, although they computed a large number of bar modes and could have easily estimated their growth rates as a resonance interaction effect. They instead turn to the language of swing amplification and the concepts of local amplification, waves propagating inward and outward with group velocities that depend on the radius, etc. This was essentially an attempt to use an obviously inadequate language to describe global modes (this is especially true of the bar modes). AS were therefore forced to adopt a large number of unjustified assumptions when estimating the growth rates of the modes (as we could see above). However, they choose the source of the instability of the bar modes to be a "global" counterpart of the local mechanism, which can quite realistically operate (especially if we remain within the limits prescribed by the approximation used). This may explain why the estimates obtained by AS are correct to an order of magnitude. That is also why we believe that, first and foremost, the swing explanation for bar formation does not provide a satisfactory language for this problem. However, the fact that both the general theory (Section 3) and specific computations (Section 4 and Appendix I) show that both the corotation and other resonances (most importantly, the OLR) can play important (and sometimes even dominant) roles casts doubt on whether the swing approach can be applied in any way to galactic bars. The entire swing pattern is fundamentally tied exclusively to the corotation region.

To avoid confusion, we emphasize that we analyze here only standard "fast" bars. The slow bars considered by Lynden-Bell [3] are probably just the central parts of unstable spiral modes (and have a secondary nature in this sense). We do not consider here the possible formation of slow bars in the very hot centers of galactic disks (fully analogous to the ellipsoidal deformation of spherical systems in the case of instability of radial orbits). In contrast, fast bars are primary objects and the spirals adjacent to them are secondary features, since they form as a response of the disk to the gravitation of the bar in the resonance region.

We conclude this section by briefly mentioning our preprint [40] and the related paper [41], where we give some arguments supporting the possibility of developing a unified theory for galactic bars. The more detailed analysis we have undertaken here shows that these arguments and the hopes to which they gave rise have been fulfilled only partially. On the one hand, we have succeeded in achieving a deeper generalization that unites spiral and bar modes. On the other hand, we have not been able to develop a theory for fast bars based on the same scheme as the Lynden-Bell theory for slow bars (i.e., instability of elongated orbits or an appropriate generalization).

7.2. Formation of Spirals

The commonly adopted approach to studies of the formation of structures in normal (SA) galaxies consists of applying either the swing mechanism [11] or some type of over-reflection, such as the waser mechanism (see, e.g., [26]). These concepts (and, of course, the idea of swing amplification) gave a great push to the development of theories of galactic dynamics and associated observational studies. However, we believe that these ideas are outdated. Let us note some of the problems faced by the swing mechanism as applied to normal spirals. First, a swing is not a true instability, and a global spiral mode can develop only if there is a closed feedback loop. Moreover, waves must be able to cross the galaxy many times (during its lifetime). The analogy between galaxies and lasers, which seems so attractive at first glance, in part because it is so unexpected, proves to be flawed, due to the very different nature of the two objects (which is not a formal, but a fundamental difference). Specific estimates are required to justify this analogy, and they fail to support it. One crucial factor is the time τ required for the wave packet to cross the galaxy. Toomre [39] himself estimated $\tau \approx 10^9$ yr for our Galaxy. With such values of τ, we can hardly imagine global spiral modes reaching a steady state as a result of swing amplification.

The damping of waves at the ILR poses an even more serious problem. If the Lynden-Bell derivative $\mathcal{F}_0'$ is everywhere positive, the waves in low-mass galaxies must inevitably meet the ILR and decay. If $\mathcal{F}_0' > 0$ but the mass of the disk is sufficiently high, bar modes should form. Finally, if $\mathcal{F}_0' < 0$ somewhere in the disk, we face the new situation considered above (Section 3) and a spiral mode should form.

The swing amplification mechanism may sometimes be the only way for an initial perturbation to grow. Such galaxies probably possess no regularly organized spiral structure with a modal nature. Note that it is global modes that are unlikely to be due to the swing mechanism. It stands to reason that this mechanism is capable of amplifying transient wave perturbations.

The corotation region plays the central role in traditional mechanisms. In our approach, spiral perturbations are unstable waves with zero total angular momentum. Such waves can exist only in the presence of regions with a negative Lynden-Bell derivative: $\mathcal{F}_0' < 0$. We already pointed out at the end of Section 3 that $\mathcal{F}_0' < 0$ is the condition for the growth of a negative-energy wave at the ILR.

In conclusion, we would like to make another general comment. Our theory (like most earlier theories) considers the formation of structures in a "ready made" stellar disk. Of course, such an approach cannot be considered to be fully satisfactory in view of both the long evolution of the galaxy prior to the formation of the disk and the complex composition of the material of which the galaxy is made. In particular, the gaseous component may play an important role (during both late and earlier stages of the evolution). A detailed theory of processes involving galactic gaseous disks can be found in the works of Fridman and his coauthors (see, e.g., [42]).

ACKNOWLEDGMENTS

We are grateful to O.V. Khoruzhiĭ for his numerous valuable comments made while refereeing the paper. This work was supported by the Foundation for the Support of Russian Science, the Russian Foundation for Basic Research (project no. 02-02-16878) and The Ministry of Industry and Science of the Russian Federation (grant NSh-925.2003.2 and state contracts 40.022.1.1.1101 and 40.020.1.1.1167.IA.1).

Appendix I

SOME ADDITIONAL ARGUMENTS IN SUPPORT OF OUR APPROACH

(1) Let us begin by generalizing our master integral equation so that it includes all the resonances (at fixed azimuthal number m). We can write the Fourier expansions of the perturbed potential and distribution function:

$$\Phi(\mathbf{I}, w_1) = \sum_l \Phi_l(\mathbf{I}) e^{ilw_1}, \qquad \text{(AI.1)}$$

$$f(\mathbf{I}, w_1) = \sum_l f_l(\mathbf{I}) e^{ilw_1},$$

where $\mathbf{I} = (I_1, I_2)$ are the common actions and w_1 is the radial angle variable. We then substitute the full potential and full distribution function

$$\Phi_0 + \Phi(\mathbf{I}, w_1) e^{imw_2}, \quad f_0 + f(\mathbf{I}, w_1) e^{imw_2} \qquad \text{(AI.2)}$$

into the linearized collisionless Boltzmann equation to obtain the following relations between $\Phi_l(\mathbf{I})$ and $f_l(\mathbf{I})$:

$$f_l(l\Omega_1 + m\Omega_2 - \omega) = \Phi_l f_{0,l}', \qquad \text{(AI.3)}$$

where we have used the notation

$$f_{0,l}'(\mathbf{I}) = l\frac{\partial f_0(\mathbf{I})}{\partial I_1} + m\frac{\partial f_0(\mathbf{I})}{\partial I_2}. \qquad \text{(AI.4)}$$

We now derive a relation between the amplitudes of the Fourier expansions (AI.1), which follows from the Poisson equation:

$$\Phi(\mathbf{I}, w_1) e^{imw_2} = -G \int d\mathbf{I}' d\mathbf{w}' \frac{f(\mathbf{I}', w_1') e^{imw_2'}}{r_{12}}. \qquad \text{(AI.5)}$$

Multiplying both sides of this equation by $e^{imw_1/2}$ and introducing the function

$$\varphi_1(w_1) = w_2 - w_1/2 - \varphi,, \qquad \text{(AI.6)}$$

we can rewrite (AI.5) in the form

$$\sum_l \Phi_l(\mathbf{I}) e^{i(l+m/2)w_1} = -G \qquad \text{(AI.7)}$$

$$\times \int d\mathbf{I}' dw_1 \sum_{l'} f_{l'}(\mathbf{I}') e^{i(l'+m/2)w_1'} e^{im\delta\varphi_1} \psi(r, r'),$$

where $\delta\varphi_1 = \varphi_1' - \varphi_1$. We now denote

$$\Pi_{l,l'}(\mathbf{I}, \mathbf{I}') \qquad \text{(AI.8)}$$

$$= \int dw_1 dw_1' \psi(r, r') e^{i(l'+m/2)w_1' - i(l+m/2)w_1} e^{im\delta\varphi_1},$$

to find the desired second relation between the amplitudes of the Fourier expansions of the potential and the distribution function:

$$\Phi_l = -\frac{G}{2\pi} \int d\mathbf{I}' \Pi_{l,l'}(\mathbf{I}, \mathbf{I}') f_{l'}(\mathbf{I}') \qquad \text{(AI.9)}$$

(here and below, summation over repeated subscript l' is implied). We finally use (AI.3) to express f_l in terms of Φ_l in order to derive from (AI.9) a set of equations that is a natural generalization of (20):

$$\Phi_l(\mathbf{I}) = -\frac{G}{2\pi} \qquad \text{(AI.10)}$$

$$\times \int d\mathbf{I}' \Pi_{l,l'}(\mathbf{I}, \mathbf{I}') f_{0,l'}'(\mathbf{I}') \frac{\Phi_{l'}(\mathbf{I}')}{l\Omega_1 + m\Omega_2 - \omega}.$$

For numerical computations it is more convenient to rewrite (AI.10) in the form of a classical eigenvalue problem. Formula (AI.3) enables us to obtain the following set of equations for the amplitudes f_l in place of (AI.10):

$$f_l(\mathbf{I})(l\Omega_1 + m\Omega_2 - \omega) = -\frac{G}{2\pi} f'_{0,l}(\mathbf{I}) \int d\mathbf{I}' \Pi_{l,l'}(\mathbf{I}, \mathbf{I}') f_{l'}(\mathbf{I}'). \tag{AI.11}$$

Note that our master integral equation (28) is a special case of the set of equations (AI.11) for $l = l' = -1$.

We now apply a discretization of the integrals, $\int d\mathbf{I}' \to \sum \Delta\mathbf{I}'$, to obtain the following characteristic equation for the mode eigenfrequencies:

$$\det \left| \frac{G}{2\pi} f'_{0,l}(\mathbf{I}) \Pi_{l,l'}(\mathbf{I}, \mathbf{I}') \Delta\mathbf{I}' + \mathbf{E}(l\Omega_1(\mathbf{I}) + m\Omega_2(\mathbf{I}) - \omega) \right| = 0, \tag{AI.12}$$

where $\mathbf{E}$ is the identity matrix.

Equations (AI.11, AI.12) represent a new formulation of the general eigenvalue problem for the stellar disk. This can be viewed as an alternative to the well-known matrix approach of Kalnajs [43].

(2) We now use (AI.12) to perform a more detailed analysis of the eigenfrequencies of the bisymmetric modes, $m = 2$. For definiteness, we will use Kalnajs's model (6, 0, 1, 0.25). We showed in Section 4 that the Lynden-Bell derivative of the distribution function for this model is positive everywhere in the phase space, so that bar modes are the only modes possible. The rotational frequencies of these modes can be determined using the master integral equation (28). The results are given in the table (see version 1—computation including only the ILR term). For comparison, the first row of the table gives the "experimental" values of the eigenfrequencies obtained by AS in their N-body simulations.

Leaving only the three most important terms in (AI.11)—those with $l = -1$, $l = 0$, and $l = 1$ (which we call the ILR, CR, and OLR terms, respectively)—we can separately analyze the effect of each of the three resonances on the growth rate of the modes (it is obvious from the table that the real part of the frequency remains virtually unchanged in the various versions). We first found the unstable bar modes when all three terms are included (version 2). Note that the eigenfrequency is close to the "experimental" value for the main mode 1. We then analyzed the effect of each of the terms by dropping various combinations of them from (AI.11). The pattern of the unstable modes in the complex ω plane was more or less the same in all cases considered, provided that the ILR term was present, but changed drastically when the ILR term was dropped. In this case, we obtained numerous unstable modes with increments of the order of 10^{-3} and obviously incorrect frequencies that had nothing in common with the experimental frequencies. This is indicative of the dominant role of the ILR term.

The table gives the eigenfrequencies of (AI.11) computed including various combinations of the terms of the expansion (AI.1) with $|l| \leq 1$.

It is clear from the table that the contribution of the OLR term to the growth increment of the most unstable mode, 1, exceeds that of the corotation term. A comparison of the eigenfrequencies of mode 2 computed in versions 2 and 4 shows that the growth of this mode is due entirely to the interaction of the gravitational potential of the mode with resonance stars at the OLR.

In principle, any number of expansion terms can be included in (AI.11). However, intuition suggests that higher-order harmonics (with $|l| \geq 2$) cannot have a significant effect on the large-scale modes that are of interest for us. Numerical computations confirm this hypothesis. Thus, including terms with $l = \pm 2$ does not lead to the appearance of any new modes. The growth rates for the fastest modes increase somewhat: $\gamma_1 = 0.08$, $\gamma_2 = 0.032$.

Recall that the rough estimates of the growth rates of the modes yielded $\gamma_1 = 0.117$ and $\gamma_2 = 0.054$, with the OLR providing the dominant contribution in both cases.

(3) Let us now explain why the corotation term is less important than both the ILR and OLR terms (despite the fact that corotation is the closest resonance to the ILR). The ratio of the denominators of the ILR and CR terms,

$$A = \frac{|\omega - m\Omega_{pr}|}{|\omega - m\Omega_2|} = \frac{\delta\Omega}{|\Omega - \Omega_p|}, \tag{AI.13}$$

is equal to the small parameter ϵ from (1) in Section 1 for $|\Omega| \gg |\Omega_p|$; i.e., in the central regions, which are fairly far from corotation. However, it is important to take into account the fact that, according to (AI.3), the coefficient f_l of the Fourier expansion is proportional to a linear combination of the derivatives of the unperturbed distribution function $f_0(\mathbf{I})$:

$$f_l \sim f'_{0,lm}(\mathbf{I}), \tag{AI.14}$$

Eigenfrequencies of equation (AI.11)

Version	Resonances	Mode 1	Mode 2
AS	N body	$0.465 + 0.066i$	$0.33 + 0.058i$
1	ILR	0.44	0.33
2	ILR, CR, OLR	$0.48 + 0.058i$	$0.38 + 0.024i$
3	ILR, CR	$0.43 + 0.015i$	0.38
4	ILR, OLR	$0.49 + 0.036i$	$0.38 + 0.024i$

where $f'_{0,lm}$ is given by (AI.14). Therefore, the relative importance of the corotation and ILR terms is determined by the ratio

$$A' = \frac{f'_{0,0}(\mathbf{I})}{|\omega - m\Omega_2|} : \frac{f'_{0,-1}(\mathbf{I})}{|\omega - m\Omega_{pr}|}, \qquad \text{(AI.15)}$$

and not by (AI.13). In an epicycle approximation, which is usually valid for the galactic disk,

$$\left|\frac{\partial f_0}{\partial I_1}\right| \gg \left|\frac{\partial f_0}{\partial I_2}\right|. \qquad \text{(AI.16)}$$

However, $l = 0$ for the corotation term, so that the ratio (AI.15) contains a small factor in addition to A. This factor evidently extends the domain of applicability of the master integral equation (28), allowing it to be used even for the fastest modes.

Appendix II

SOME SPECIFIC FEATURES OF THE NUMERICAL ALGORITHMS FOR THE SOLUTION OF THE MASTER INTEGRAL EQUATION

Let us now consider some specific features of the homogeneous Fredholm equation of the second kind

$$\lambda f(x) = \int\limits_a^b K(x,z) f(z) dz, \qquad \text{(AII.1)}$$

of which our master integral equation is a special case. In (AII.1), $K(x,z)$ is the kernel, $f(x)$ is an unknown function, and λ is the parameter of the equation. Solutions $f(x)$ that are not identically equal to zero exist only at the eigenvalues λ, which form the spectrum of the integral equation. Solving the integral equation (AII.1) consists in finding these eigenvalues and the corresponding eigenfunctions.

The number of eigenvalues depends on the form of the kernel $K(x,z)$. If $K(x,z) = g(x)\delta(x-z)$, where $\delta(x)$ is the Dirac delta function, the eigenvalues and corresponding eigenfunctions are equal to $\lambda = g(x_0)$ and $f(x) = \delta(x - x_0)$. Thus, in this example, we have an uncountable number (continuum) of eigenvalues. On the contrary, if the kernel of the integral equation obeys the Hilbert–Schmidt condition,

$$\int\limits_a^b \int\limits_a^b |K(x,z)|^2 dx dz < \infty, \qquad \text{(AII.2)}$$

the number of eigenvalues is at most countable; i.e., the spectrum is discrete (see, e.g., [44]).

One characteristic of the kernel $K(\mathbf{J},\mathbf{J}')$ of our integral equation is that it contains the term $\Omega_{pr}(\mathbf{J}) \times \delta(\mathbf{J} - \mathbf{J}')$, so that the condition (AII.2) is not satisfied. Therefore, strictly speaking, the spectrum of this integral equation contains both a continuous and a discrete component.

Nearly all methods for the numerical solution of integral equations use quadrature formulas. These formulas make it possible to associate linear integral equations that relate functions in an infinite-dimensional space with ordinary linear algebraic equations for vectors in a finite-dimensional vector space. Let us suppose that we must solve an integral equation of the form (AII.1). We first choose some quadrature formula

$$\int\limits_a^b f(z) dz = \sum_{j=1}^N w_j f(z_j).$$

Here, $\{w_j\}$ are the weights of the quadrature formula and $\{z_j\}$ are some points in the interval $[a,b]$. Thus, for example, to integrate using the trapezoid method,[13] which we used to solve the master integral equation (16), $w_1 = w_N = 1/2$, $w_j = 1$, $(j = 2, \ldots, N-1)$, $z_j = a + (b-a)(j-1)/(N-1)$, $j = 1, \ldots, N$. We next compute (AII.1) at the quadrature points to obtain

$$\lambda f(x_i) = \sum_{j=1}^N w_j K(x_i, z_j) f(z_j). \qquad \text{(AII.3)}$$

We now denote $f_i \equiv f(x_i)$, $K_{ij} \equiv K(x_i, z_j)$ and define

$$\tilde{K}_{ij} = w_j K_{ij}$$

to obtain a finite-dimensional analog of (AII.1),

$$\sum_{j=1}^N (\tilde{K}_{ij} - \lambda\delta_{ij}) f_j = 0, \qquad \text{(AII.4)}$$

[13] Since the computation of eigenfunctions and eigenvalues requires $O(N^3)$ operations, the most efficient way to do this is to use high-order quadrature formulas (see, e.g., [45]). Here, however, we restrict our analysis to applying the trapezoid formula.

which can be solved using standard methods of linear algebra.

The state of the system we analyze here is determined by the unknown distribution function $\mathcal{F}$, which depends on two independent variables; i.e., the phase space of the system is two-dimensional. In our numerical computations, the variables in question were the energy E and angular momentum L of the star. The quadrature grid $\{E_i\}$ in energy was always uniform, $E_1 = E_{\min} \equiv \Phi_0(r_{\min})$, $E_{N_E} = E_{\max} \equiv \Phi_0(r_{\max}) + v^2(r_{\max})/2$, where $\Phi_0(r)$ is the equilibrium potential, $r_{\min}$ and $r_{\max}$ are the inner and outer cutoff radii of the disk, and $v(r_{\max})$ is the circular velocity at $r = r_{\max}$. We used various grids $\{L_j\}$ in angular momentum, depending on the degree to which the equilibrium distribution function $f_0(E, L)$ was concentrated toward the line of circular orbits, $L = \pm L_c(E)$. The upper boundary of the grid for this energy always coincided with the line of prograde circular orbits: $L^i_{N_L} = L_c(E_i)$. If the equilibrium distribution function was more or less uniformly spread throughout the phase space, the lower boundary coincided with the line of retrograde circular orbits, $L^i_1 = -L_c(E_i)$ (or $L^i_1 = 0$, if there are no retrograde stars). The advantage of such a grid is that it depends only on the potential $\Phi_0(r)$ and is independent of the particular distribution function f_0 of the model. It follows that different models with the same equilibrium potential can be analyzed without recomputing the function $\Pi(\mathbf{J}, \mathbf{J}')$. If, on the other hand, $f_0(E, L)$ is concentrated toward the line of circular orbits $L_c(E)$, this is not a practical way to define the lower boundary. In this case, we define L^i_1 so that the distribution function $f_0(E_i, L)$ at $-L(E_i) \le L \le L^i_1$ (i.e., in the part of the phase plane that we discarded) is less than 2% of its maximum value $f_0(E_i, L)$.

The integral equation on a two-dimensional phase space can be easily reduced to the one-dimensional integral equation considered above via a simple renumbering of the subscripts: $i, j \to \nu$. We used the simplest method for such renumbering in accordance with the rule

$$\nu = (N_E - 1)i + j.$$

In most of our computations, each of the grids in E and L consisted of $N_E = N_L = 31$ points; i.e., the full grid of the phase space consisted of $N \equiv N_E N_L = 961$ points. Consequently, to approximately determine the eigenvalues and eigenfunctions for our master integral equation, we must solve the eigenvalue problem for the $N \times N$ matrix $M_{\nu\nu'} \equiv K_{\nu\nu'} w_{\nu'}$:

$$\sum_{\nu'=1}^{N} M_{\nu\nu'} F_{\nu'} = \Omega_p F_\nu, \qquad \text{(AII.5)}$$

where

$$K_{\nu\nu'} = \frac{G}{2\pi} \mathcal{F}'_{0\nu} \Pi_{\nu\nu'} + (\Omega_{pr})_\nu \delta_{\nu\nu'} \qquad \text{(AII.6)}$$

is the finite-dimensional analog of the kernel of the integral equation, $F_\nu = \mathcal{F}(\mathbf{J}_\nu)$ is the unknown grid function, w_ν is the weight function introduced above, Ω_p is the unknown eigenfrequency, $\mathcal{F}'_{0\nu} = \mathcal{F}'_0(\mathbf{J}_\nu)$, $\Pi_{\nu\nu'} = \Pi(\mathbf{J}_\nu, \mathbf{J}'_{\nu'})$, $(\Omega_{pr})_\nu = \Omega_{pr}(\mathbf{J}_\nu)$, and $\delta_{\nu\nu'}$ is a Kronecher delta function.

The 31×31 grid can adequately be calculated using the computer we employed. All the eigenvalues and eigenvectors could be computed in several minutes. A twofold densification of the grid in each of the phase-space coordinates increases the computation time by a factor of 64, while yielding no qualitatively new results.

Appendix III

IMITATION OF THE STELLAR VELOCITY DISPERSION USING SOFTENED GRAVITY

In Section 7.1.2, we pointed out the flaws of imitating the effects of the real dispersion of the particle velocities using softened gravity[14]). Toomre, among others, used this approach in his paper [11], where he also reports the results computations of the stability of a Gaussian disk (which is cool but has softened gravity). Toomre believed that a comparison of these results with the classical bar modes in liquid incompressible MacLaurin spheroids should demonstrate how the “true” bar modes in stellar systems differ from classical bar modes (which are obviously due to the edge instability). Toomre [11] then considers the results of these computations to provide the main argument in favor of the swing mechanism for bar formation, which he proposed in the same paper [11]. Binney and Tremaine [13] used the same example in their famous textbook on stellar dynamics. Following Toomre [11], they tried to give a physical interpretation for the bar instability based on swing amplification; they call Toomre’s example “one striking clue,” which allegedly supports this viewpoint. The persistent popularity of this idea is testified to, for example, by the recent paper by Tremaine [34], who used the

[14] To prevent misunderstanding, we note that the procedure of softening gravity that has been used in many N-body simulations (including those of AS cited above) has a different goal—ensuring reasonable accuracy of computations of the forces of gravitational attraction between stars during close encounters. The initial velocities of the stars are modeled in accordance with the equilibrium distribution function studied (and are not assumed to be purely circular as in the imitation approach discussed here).

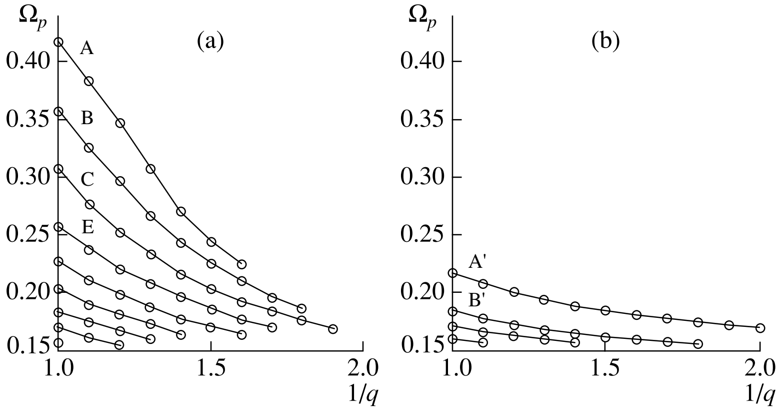

Fig. 11. Angular velocities of various ($m = 2$) modes of a Gaussian disk (see text for details) as a function of $1/q$, where q is the mass fraction that participates in the perturbations, (a) in a softened-gravity approximation [11] and (b) with the actual velocity dispersion corresponding to the approximation shown in (a).

same approach to analyze slow, single-armed modes ($m = 1$) in nearly Keplerian disks.

The arguments presented above justify the need for a more detailed analysis of this imitation. This idea was first suggested by Miller [46, 47], who noted that substituting the softened potential $\Phi = -GM/(b^2 + d^2)^{1/2}$ (where b is the smoothing scale length) for the exact potential $\Phi = -GM/d$ due to the mass M at a distance d results in the following dispersion equation for short-wavelength axisymmetric perturbations of a cool disk:

$$\omega^2 = \kappa^2(r) - 2\pi G\sigma_0(r)|k|e^{-|k|b}, \qquad \text{(AIII.1)}$$

which differs from the well-known standard dispersion equation in an exponential reduction factor. We can easily show using (AIII.1) that all the axisymmetric perturbations described by this equation are stable if

$$b > b_{\min} = \max\left[\frac{2\pi G\sigma_0(r)}{e\kappa^2(r)}\right]. \qquad \text{(AIII.2)}$$

In particular, for a Gaussian disk with surface density $\sigma_0(r) = e^{-r^2/2}/2\pi$ (so that the mass of the disk is $M = 1$), we obtain from (AIII.2)

$$b_{\min} = \max[B(r)], \qquad B(r) = e^{-r^2/2-1}/\kappa^2(r), \qquad \text{(AIII.3)}$$

where we set $G = 1$, as did Toomre [11]. The subsequent computation of $b_{\min}$ must be performed numerically. The potential produced by a disk with surface density $\sigma_0(r)$ can most easily be computed using the formula

$$\Phi_0(r) = -4Gr\int\limits_0^{\pi/2} d\psi \sin\psi \int\limits_0^{\infty} \mathrm{ch}\varphi\sigma_0(r\sin\psi\mathrm{ch}\varphi)d\varphi. \qquad \text{(AIII.4)}$$

We can now compute the square of the angular velocity of rotation, $\Omega^2(r) = \Phi_0'/r$, and $\kappa^2(r) = 4\Omega^2 + r(\Omega^2)'$. This yields the function $B(r)$, whose maximum is equal to $B_{\max} = b_{\min} \approx 0.2$, in accordance with Toomre [11].

Having thus determined the $b = b_{\min}$ at which all radial instabilities are suppressed, Toomre [11] sets $b = 0.25$ (which is somewhat higher than $b_{\min}$), then analyzes possible nonaxisymmetric instabilities—primarily bisymmetric instabilities ($m = 2$). Figure 11a shows the velocities of Ω_p modes A, B, C, E ... as a function of $1/q$, adopted from Toomre [11]. Here, q is the mass fraction of the active disk (the fraction of mass that participates in the perturbations); our parameter q is related to Toomre's parameter f as $q = 1/(1 + f)$. These modes differ in their radial wavelength, which progressively decreases from A to E such that mode A corresponds to the most open spirals.

However, as we noted in Section 7.1.2, this type of imitation is itself inadequate. Moreover, we can directly compute the velocities Ω_p for the linear modes that had to be imitated using softened gravity and compare our results with those of Toomre. It is clear from an analysis of the limit of the dispersion equation (AIII.1) as $kb \ll 1$ (or from dimensional considerations) that the velocity dispersion c is related to b as $c^2 = 2\pi G\sigma_0(r)b$. Note that, when $b = 0.25$, the Toomre parameter for a Gaussian disk is $Q(r) \approx 0.93\kappa(r)e^{r^2/4}$; numerical computations show that $Q(r) \approx 1.5 = \text{const}$ in most of the disk. The disk considered is therefore rather "cool." Such a disk can be adequately represented by a generalized Schwartzschild model with known $\sigma_0(r)$ and $c(r)$ (see Sections 3 and 4). We then solve our master integral equation for this model to find the eigenfrequencies Ω_p of the bisymmetric modes.

Figure 11b shows these eigenfrequencies as functions of q. Here, A$'$, B$'$, C$'$, and E$'$ correspond to various radial wavelengths (similar to the modes A, B, C, and E in Fig. 11a). A comparison of Figs. 11a and 11b shows that imitation using softened gravity yields very overestimated mode velocities Ω_p (by about a factor of two in the example considered). To obtain frequencies close to those shown in Fig. 11a, the velocity dispersion must be decreased by a factor of 2.5 over the value we used in the computations illustrated in Fig. 11b, yielding totally unrealistic values for Toomre's parameter, $Q < 1$.

REFERENCES

1. A. J. Kalnajs, Ph.D. Thesis (Harvard Univ, 1965).
2. F. H. Shu, Astrophys. J. **160**, 89 (1970).
3. D. Lynden-Bell, Mon. Not. R. Astron. Soc. **187**, 101 (1979).
4. G. Chew, M. Goldberger, and F. Low, Proc. R. Soc. London **236**, 112 (1956).
5. A. B. Mikhailovskii, *Theory of Plasma Instabilities* (Atomizdat, Moscow, 1970), Vols. 1, 2.
6. E. Athanassoula and J. Sellwood, Mon. Not. R. Astron. Soc. **221**, 213 (1986).
7. C. C. Lin, C. Yuan, and F. H. Shu, Astrophys. J. **155**, 721 (1969).
8. L. Blitz, M. Fich, and S. Kulkarni, Science **220**, 1233 (1983).
9. V. L. Polyachenko, Zh. Éksp. Teor. Fiz. **59**, 228 (1992).
10. V. L. Polaychenko, Astron. Zh. **69**, 10 (1992) [Sov. Astron. **36**, 5 (1992)].
11. A. Toomre, in *Structure and Evolution of Normal Galaxies,* Ed. by S. M. Fall and D. Lynden-Bell (Cambridge Univ. Press, 1981), p. 111.
12. I. I. Pasha, Istor.-Astron. Issled. **27**, 102 (2002); **29**, 8 (2004).
13. J. Binney and S. Tremaine, *Galactic Dynamics* (Princeton Univ. Press, Princeton, 1987).
14. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 1: *Mechanics* (Nauka, Moscow, 1982, 1988; Pergamon Press, New York, 1988).
15. D. Lynden-Bell and A. J. Kalnajs, Mon. Not. R. Astron. Soc. **157**, 1 (1972).
16. V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Nauka, Moscow, 1979).
17. M. McIntyre, J. Fluid Mech. **106** (1981), *A Special Issue Celebrating the 25th Anniversary of the Journal*, Ed. by G. Batchelor and H. Moffat (Cambridge Univ. Press, Cambridge, 1981; Mir, Moscow, 1984), p. 454.
18. V. L. Polyachenko and I. G. Shukhman, Astron. Zh. **59**, 228 (1982) [Sov. Astron. **26**, 140 (1982)].
19. P. Goldreich and S. Tremaine, Astrophys. J. **233**, 857 (1979).
20. E. V. Polyachenko, Mon. Not. R. Astron. Soc. **330**, 105 (2002).
21. E. V. Polyachenko, Mon. Not. R. Astron. Soc. **331**, 394 (2002).
22. A. J. Kalnajs, Astrophys. J. **205**, 751 (1976).
23. B. Lindblad and R. Langebartel, Stockholm Observ. Ann. **17** (6) (1953).
24. C. C. Lin and F. H. Shu, Proc. Nat. Acad. Sci. USA **55**, 229 (1966).
25. A. J. Kalnajs, in *IAU Symposium No. 38: The Spiral Structure of our Galaxy*, Ed. by W. Becker and G. I. Kontopoulos (Reidel, Dordrecht, 1970), p. 318.
26. G. Bertin and C. C. Lin, *Spiral Structure in Galaxies. A Density Wave Theory* (MIT Press, Cambridge, 1996).
27. D. Lynden-Bell, Lect. Notes Phys. **474**, 7 (1996).
28. A. Kalnajs, Proc. Astron. Soc. Austral. **2**, 174 (1973).
29. G. Contopoulos, Astron. Astrophys. **81**, 198 (1980).
30. G. Contopoulos, Astrophys. J. **201**, 566 (1975).
31. G. Contopoulos and C. Mertzanides, Astron. Astrophys. **61**, 477 (1977).
32. J. A. Sellwood and A. Wilkinson, Rep. Prog. Phys. **56**, 173 (1993).
33. S. A. Erickson, Ph.D. Thesis (MIT, Cambridge, USA, 1974).
34. S. Tremaine, Astron. J. **121**, 1776 (2001).
35. N. Meyer-Vernet and B. Sicardy, Icarus **69**, 157 (1987).
36. J. W-K. Mark, Astrophys. J. **205**, 363 (1976).
37. P. Goldreich and D. Lynden-Bell, Mon. Not. R. Astron. Soc. **130**, 125 (1965).
38. W. H. Julian and A. Toomre, Astrophys. J. **146**, 810 (1966).
39. A. Toomre, Astrophys. J. **158**, 899 (1969).
40. E. V. Polyachenko and V. L. Polyachenko, astro-ph/0212553 (2002).
41. V. L. Polyachenko and E. V. Polyachenko, Pis'ma Astron. Zh. **29**, 508 (2003) [Astron. Lett. **29**, 447 (2003)].
42. A. M. Fridman and O. V. Khoruzhii, Space Sci. Rev. **105**, 1 (2003).
43. A. Kalnajs, Astrophys. J. **212**, 637 (1977).
44. A. N. Kolmogorov and S. V. Fomin, *Elements of Function Theory and Functional Analysis* (Nauka, Moscow, 1989) [in Russian].
45. A. H. Stroud and D. Secrest, *Gaussian Quadrature Formulas* (Prentice-Hall, Englewood Cliffs, 1966).
46. R. H. Miller, Astrophys. Space Sci. **14**, 73 (1971).
47. R. H. Miller, Astrophys. J. **190**, 539 (1974).

Translated by A. Dambis

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 900–908. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 988–997.
Original Russian Text Copyright © 2004 by Pushkarev, Kovalev, Molotov, Nechaeva, Gorshenkov, Tuccari, Stanghellini, Hong, Quick, Dougherty, Liu.

Quasi-Simultaneous VLBI and RATAN-600 Observations of Active Galactic Nuclei

A. B. Pushkarev[1], Yu. Yu. Kovalev[2,3], I. E. Molotov[1], M. B. Nechaeva[4], Yu. N. Gorshenkov[5], G. Tuccari[6], C. Stanghellini[6], X. Hong[7], J. Quick[8], S. Dougherty[9], and X. Liu[10]

[1]*Main Astronomical Observatory, Russian Academy of Sciences, Pulkovo, St. Petersburg, 196140 Russia*
[2]*National Radio Astronomy Observatory, P.O. Box 2, Rt. 28/92, Green Bank, WV 24944-0002, USA*
[3]*Astro Space Center of the Lebedev Institute of Physics, Russian Academy of Sciences, Profsoyuznaya ul. 84/32, Moscow, 117997 Russia*
[4]*Radiophysical Research Institute, ul. Bol'shaya Pecherskaya 25/14, Nizhni Novgorod, 603600 Russia*
[5]*Special Design Bureau, Power Engineering Institute, Krasnokazarmennaya ul. 14, Moscow, 111250 Russia*
[6]*Istituto di Radioastronomia del C.N.R., Stazione VLBI di Noto, C. da Renna Bassa—Loc. Case di Mezzo, C.P. 141, 96017 Noto (SR), Italy*
[7]*Shanghai Astronomical Observatory, 80 Nandan Road, Shanghai 200030, China*
[8]*Hartebeesthoek Radio Astronomy Observatory, P.O. Box 443, Krugersdorp 1740, South Africa*
[9]*Dominion Radio Astrophysical Observatory, P.O. Box 248, Penticton, B.C. V2A 6K3, Canada*
[10]*Urumqi Astronomical Observatory, 40-5 South Beijing Road, Urumqi, Xinjiang 830011, China*

Received March 10, 2004; in final form, May 27, 2004

Abstract—VLBI observations of several quasars and BL Lacertae objects were carried out at 1.66 GHz in November–December 1999 using six antennas (Medvezh'i Ozera, Svetloe, Pushchino, Noto, HartRAO, and Shanghai). Maps of six sources (0420+022, 0420–014, 1308+326, 1345+125, 1803+784, and DA 193) obtained with milliarcsecond resolution are presented and discussed, together with their broad-band (1–22 GHz) spectra obtained on the RATAN-600 radio telescope at epochs close to those of the VLBI observations. Comparison of the VLBI maps with maps of these sources obtained on standard VLBI networks and with the RATAN-600 quasisimultaneous total-flux measurements indicates the reliability of the results obtained on this Low Frequency VLBI Network and the good efficiency of this network. *© 2004 MAIK "Nauka/Interperiodica".*

1. INTRODUCTION

The Low Frequency VLBI Network (LFVN) project has been in operation since 1996 [1]. Its main goal is to help organize international VLBI experiments at low frequencies with the participation of Russian radio telescopes. During this time, 13 antennas have been equipped with the necessary radio astronomy receivers and data acquisition instrumentation: the Medvezh'i Ozera (64 m), Pushchino (22 m), Zimenki (15 m), and Staraya Pustyn' (14 m) telescopes in Russia, the Evpatoria (70 m) and Simeiz (22 m) telescopes in the Ukraine, as well as the Ventspils (32 m, Latvia), Noto (32 m, Italy), Toruń (14 m, Poland), Pune (45 m, India), Urumqi (25 m, China) and Shanghai (25 m, China) antennas and the Ooty 500×30 parabolic cylinder (India). Eighteen VLBI experiments using various combinations of radio telescopes located in England, India, Italy, Canada, China, Latvia, Poland, Russia, USA, Ukraine, South Africa, and Japan were organized, as well as correlation centers in Canada, Russia, and the USA.

At present, three aspects of the LFVN project are being developed: (1) a subsystem based on a Mk-2 data acquisition terminal and the NIRFI-3 correlator in Nizhni Novgorod for studies of the solar wind and solar microflares (spikes) at 327 and 610 MHz; (2) an international network based on the S2 broadband Canadian recording terminal [2, 3] and the Dominion Radio Astronophysical Observatory (DRAO) correlator at Penticton (Canada) [4] for studies of active galactic nuclei, maser sources, and active stars at 1.66 and 4.82 GHz; and (3) VLBI radar at 5010 MHz with retransmission of the received echo signals to the Noto processing center [5] via the Internet for measurements of the motions of the terrestrial planets, asteroids approaching the Earth, and so-called "space garbage."

The international network of radio telescopes equipped with S2 recorders included the Medvezh'i Ozera, Svetloe (32 m), Pushchino telescopes in Russia, the Green Bank (43 m) and Arecibo (300 m) telescopes in the USA, and the Noto (Italy), HartRAO (26 m, South Africa), and Shanghai (China) telescopes. Since 1998, the experiments INTAS98.2, INTAS98.5, INTAS99.4, INTAS00.3, and LFVN03.1 have been performed at 18 cm. The first four of these have been successfully correlated on the Penticton correlator; our results for INTAS99.4 are presented and discussed in this paper. The results for the other experiments will appear in subsequent publications.

In this paper, we will use the values of the Hubble constant $H_0 = 70h$ km s^{-1} Mpc^{-1} and the deceleration parameter $q_0 = 0.5$.

2. OBSERVATIONS AND REDUCTION

The LFVN observations were carried out from November 30 to December 3, 1999 (epoch 1999.91), at a frequency of 1.66 GHz. The total duration of the experiment was 43 h, with a mean duration for each scan of about 30 min. Each source was observed in 5–8 scans. Figure 1 shows the resulting coverage of the (u, v) plane for DA 193 as an example.

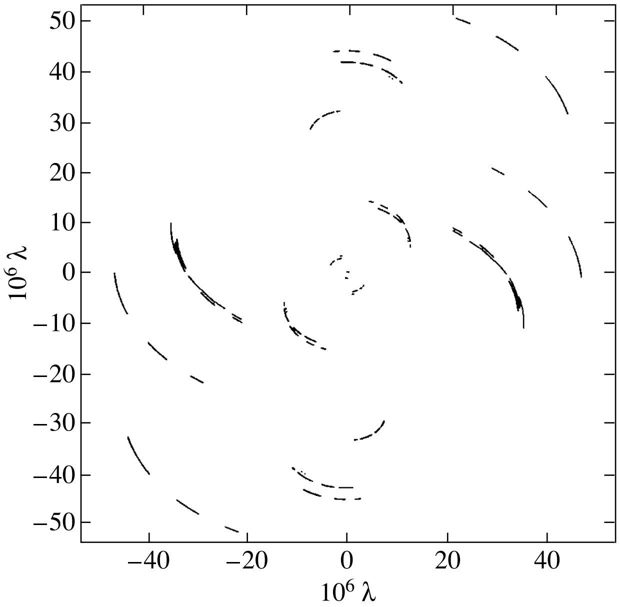

Fig. 1. Coverage of the (u, v) plane for DA 193 for LFVN observations INTAS99.4 at 18 cm.

The Medvezh'i Ozera, Pushchino, Svetloe, Noto, Shanghai, and HartRAO antennas participated in the experiment. Some parameters of the radio telescopes communicated to us by the staff of the observatories are listed in Table 1 (diameter, system temperature, and system equivalent flux density, SEFD). The participation of the HartRAO antenna (South Africa) considerably improved the angular resolution in the north–south direction. The maximum projected baseline between Shanghai and HartRAO reached 10 170 km. The Canadian S2 data acquisition system was used. The bandwidth was 4 MHz (256 spectral channels, each 15.625 kHz). Left-circular polarization was recorded with one-bit signal sampling. The correlation of the data was performed on the DRAO correlator in Penticton with an averaging time of 2 s.

The data analysis, editing, calibration, and imaging (for more details, see [6]) were done using standard procedures in the AIPS package (NRAO). For the amplitude calibration of the data, we used gain curves and system temperatures measured for each of the antennas involved in the observations. The primary phase calibration was done using the AIPS task FRING with a coherent integration time of 120 s, with subsequent phase corrections for the residual delays being found for the entire duration of the experiment, with the Medvezh'i Ozera telescope used as the reference antenna. A point source at the phase center was used for the initial models in the hybrid mapping.

The observations of the broadband spectra of the sources were carried out as part of an ongoing program of monitoring of compact extragalactic objects on the largest Russian radio telescope—the RATAN-600 (Special Astrophysical Observatory, Russian Academy of Sciences). A description of this program, the procedure used for the observations, and the data processing are given by Kovalev *et al.* [7].

3. DISCUSSION

Our results for six extragalactic objects are presented below. Figures 2–7 show the LFVN maps, together with the broadband spectra of the sources obtained on the RATAN-600 at epochs close to the date of the VLBI experiment. We have modeled the

Table 1. Antennas and their parameters at 1.66 GHz

Antenna	Diameter, m	T_{sys}, K	SEFD, Jy
Svetloe (Russia)	32	71	394
Medvezh'i Ozera (Russia)	64	95	156
Pushchino (Russia)	22	111	1586
HartRAO (South Africa)	26	50	500
Noto (Italy)	32	107	1070
Shanghai (China)	25	100	1250

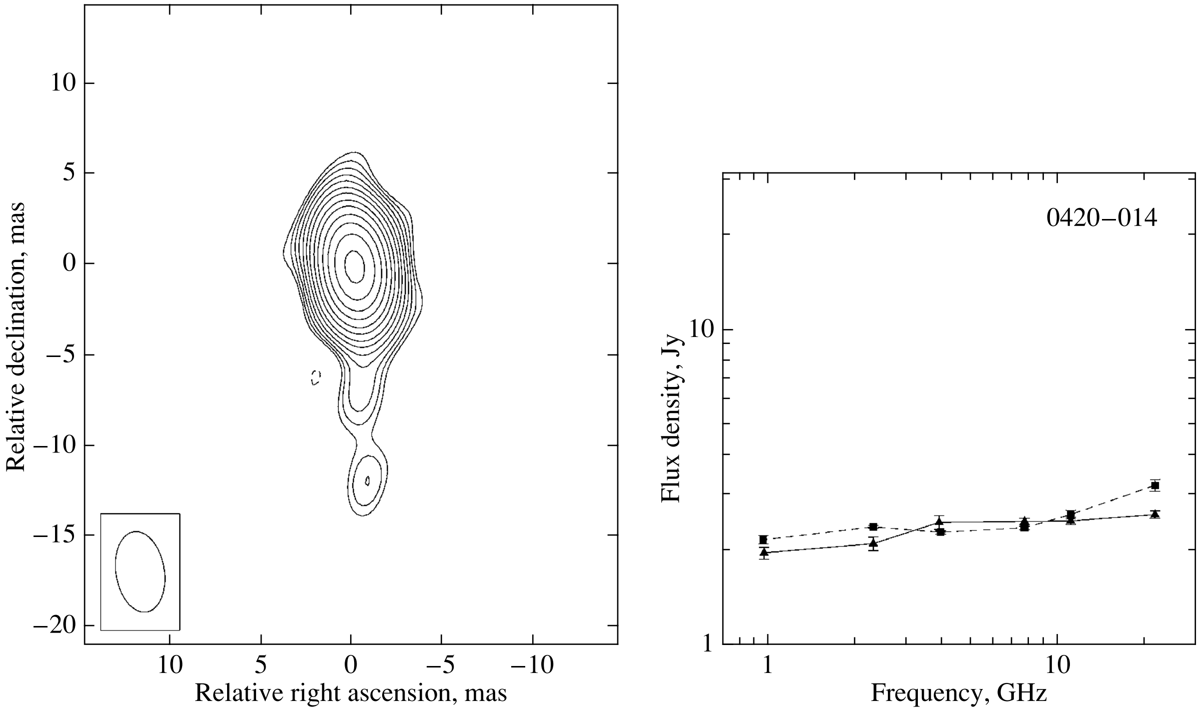

Fig. 2. Left: 1.66-GHz LFVN map of 0420−014. The lowest contour is drawn at a level of 1.4% of the peak value of 1370 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 3.6 × 2.2 mas in position angle −8°. Right: the broadband spectrum measured on the RATAN-600. Individual measurements are shown with $\pm1\sigma$ errors and are connected with lines. The filled triangles and solid line segments show the measurements for September 1999, and the filled squares and dashed line segments show those for April 2000.

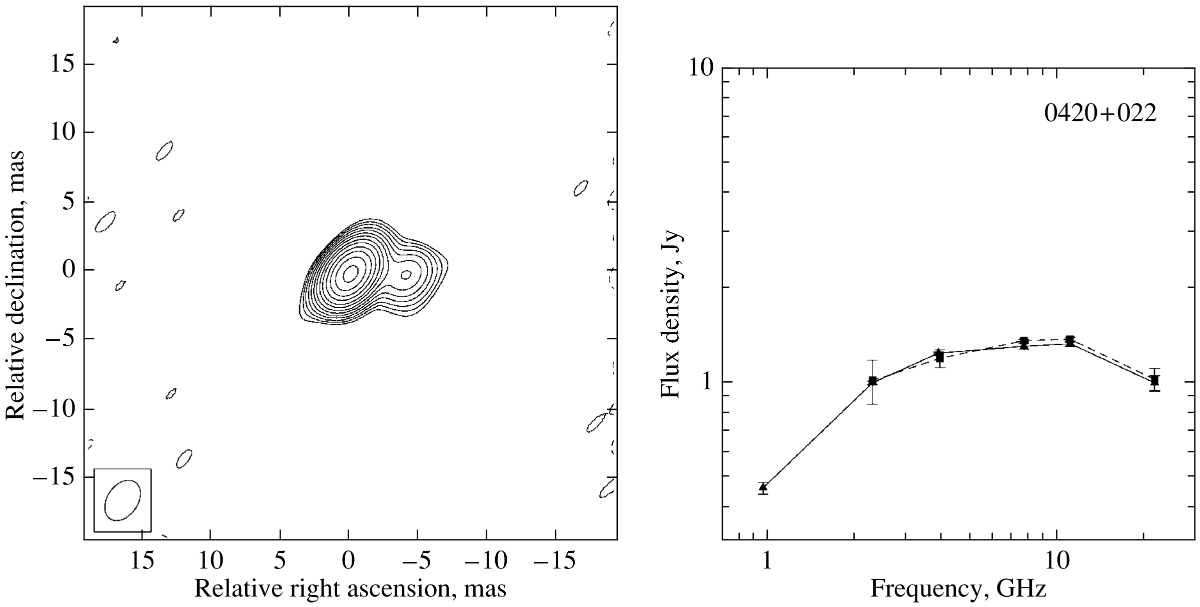

Fig. 3. Left: 1.66-GHz LFVN map of 0420+022. The lowest contour is drawn at a level of 1.4% of the peak value of 775 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 3.2 × 2.2 mas in position angle −35°. Right: same as Fig. 2b for 0420+022.

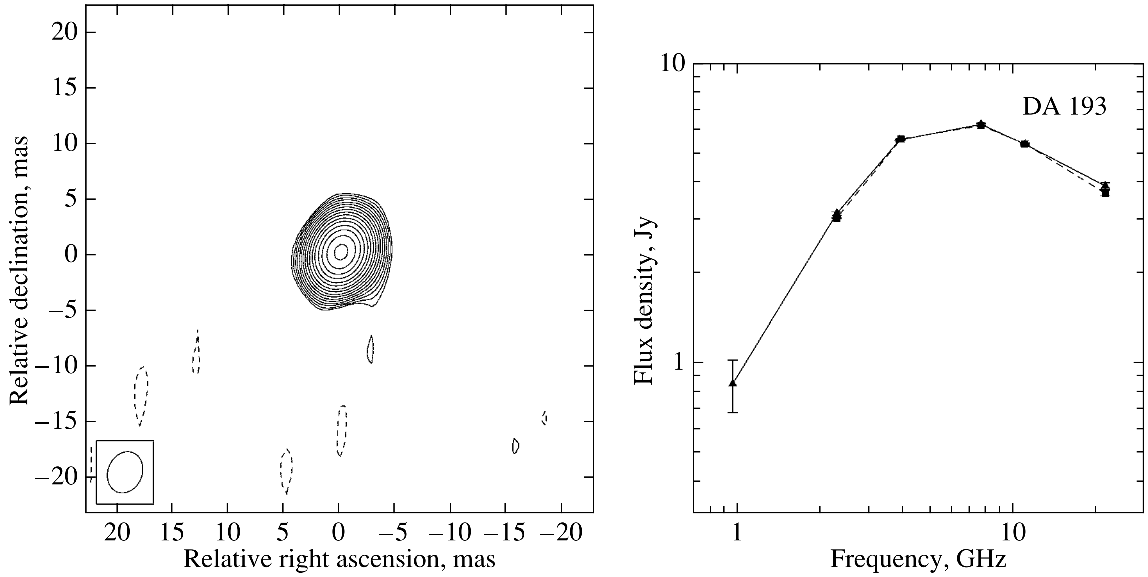

Fig. 4. Left: 1.66-GHz LFVN map of DA 193 (0552+398). The lowest contour is drawn at a level of 0.5% of the peak value of 1833 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 3.7 × 3.0 mas in position angle −48°. Right: same as Fig. 2b for DA 193.

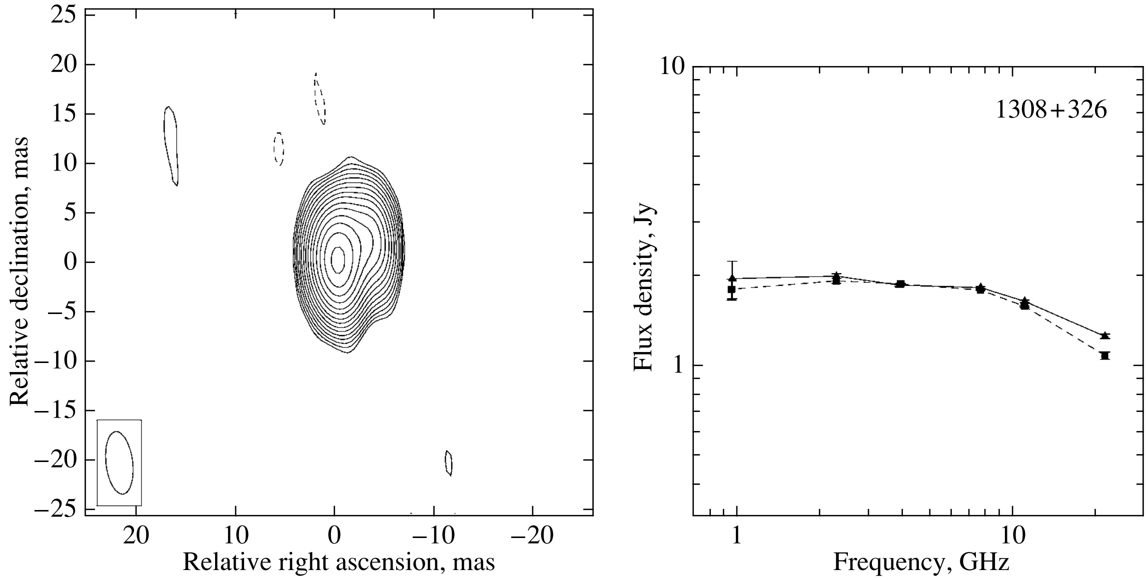

Fig. 5. Left: 1.66-GHz LFVN map of 1308+326. The lowest contour is drawn at a level of 0.7% of the peak value of 1565 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 6.2 × 2.6 mas in position angle +8°. Right: same as Fig. 2b for 1308+326.

VLBI structures of the sources with circular Gaussian components by fitting the models to the fully calibrated observational data in the visibility (u, v) plane using the Brandeis VLBI package [8]. The source models are listed in Table 2, which gives the (1) object name, (2) total flux density of the model component, (3), (4) component position on the map in polar coordinates r and φ relative to the brightest component, (5) the FWHM of the Gaussian component. The formal errors are given at the 1σ level; this corresponds to an increase in the value of χ^2 for the obtained model by unity (for details see [9]). For the object 1345+125, which has a composite structure, it was not possible to derive an adequate model of the source due to the sparse coverage of the (u, v) plane.

Let us proceed to a discussion of each of the studied objects.

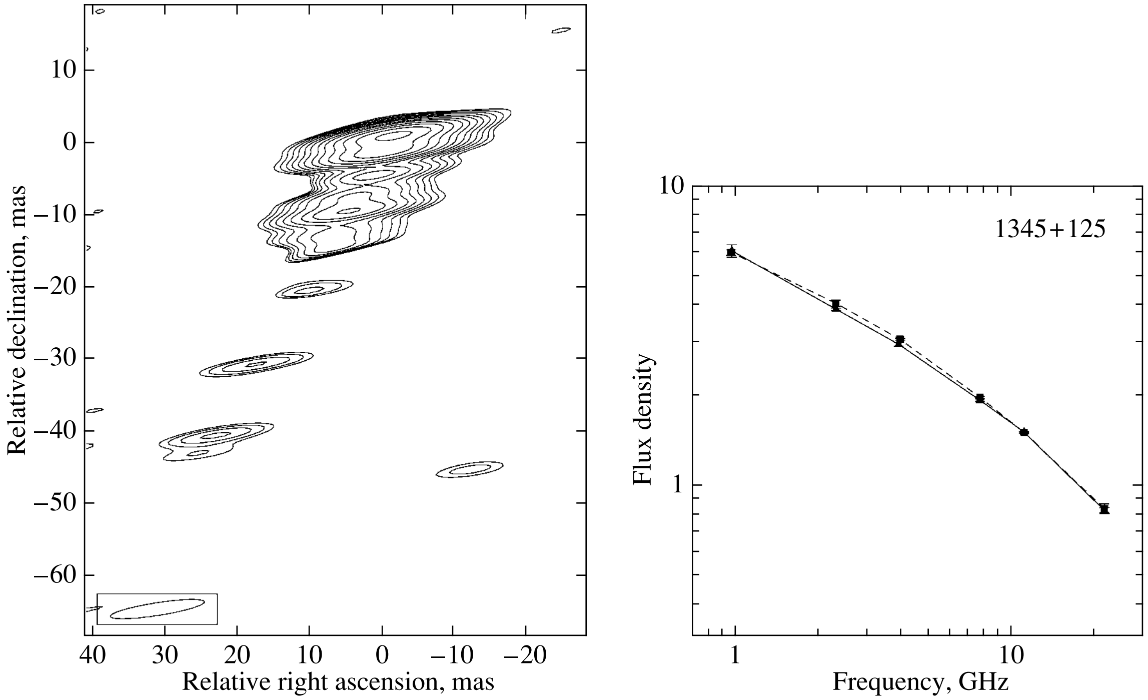

Fig. 6. Left: 1.66-GHz LFVN map of 1345+125. The lowest contour is drawn at a level of 5.6% of the peak value of 515 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 12.1 × 1.8 mas in position angle −82°. Right: same as Fig. 2b for 1345+125.

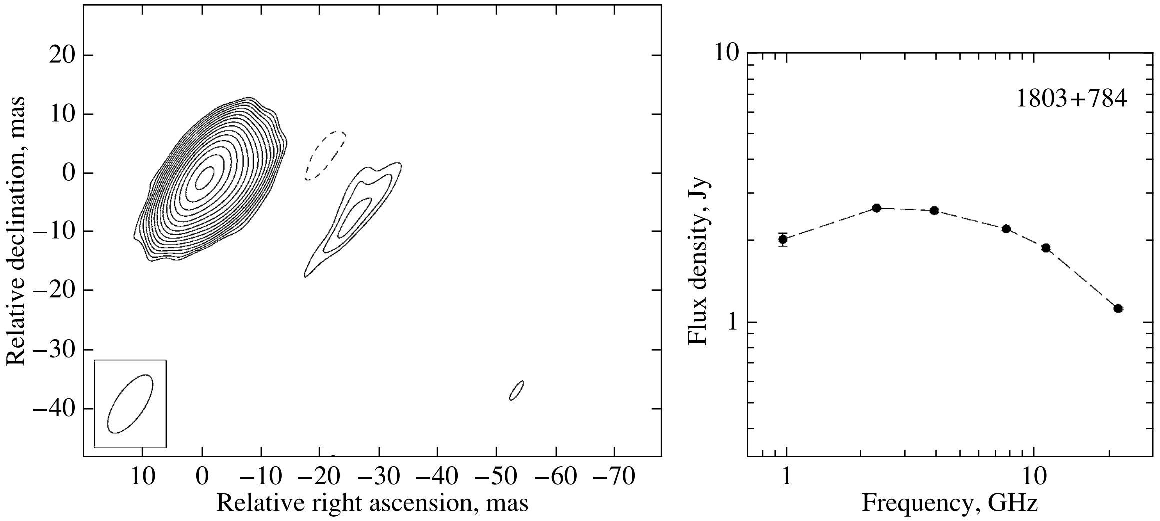

Fig. 7. Left: 1.66-GHz LFVN map of 1803+784. The lowest contour is drawn at a level of 0.5% of the peak value of 1480 mJy/beam, and the contours increase in steps of $\sqrt{2}$. The restoring beam is 11.5 × 4.8 mas in position angle −34°. Right: same as Fig. 2b for 1803+784, but for measurements made in September 1998.

0420−014. This source is a strongly variable and highly polarized quasar (redshift $z = 0.915$) with a flat radio spectrum (Fig. 2). 0420−014 was identified with a gamma-ray source based on analysis of Compton Gamma-Ray Observatory EGRET data [10, 11]. In 1992, simultaneous optical and gamma-ray flares were observed together with a considerable increase of the radio emission, which was subsequently identified with the appearance of a new component in the jet. Analysis of the data for the 1992 flare suggest the presence of a binary black hole in this source [12].

Our map of 0420−014 demonstrates a dominant VLBI core and several weak components in the southward jet (Fig. 2). The morphology of the

Table 2. Models of the sources

Source	$I \pm \sigma_I$, mJy	$r \pm \sigma_r$, mas	$\varphi \pm \sigma_\varphi$, deg	$\theta \pm \sigma_\theta$, mas
0420−014	1309 ± 42	...	...	<0.5
	110 ± 23	1.78 ± 0.12	159 ± 5.5	<0.5
	30 ± 4	7.06 ± 0.22	-176 ± 0.7	<0.5
	20 ± 5	12.59 ± 0.38	-176 ± 0.8	<0.5
0420+022	772 ± 28	...	...	<0.5
	52 ± 9	1.78 ± 0.14	-41 ± 4.6	<0.5
	61 ± 8	2.31 ± 0.21	-86 ± 4.8	0.63 ± 0.25
	73 ± 8	4.53 ± 0.25	-93 ± 3.3	<0.64
DA 193	1872 ± 8	...	...	0.55 ± 0.05
1308+326	1389 ± 44	...	...	1.11 ± 0.10
	925 ± 68	2.33 ± 0.13	35 ± 3.1	1.97 ± 0.23
	299 ± 60	4.99 ± 0.15	54 ± 1.8	1.45 ± 0.61
1803+784	1459 ± 87	...	...	1.10 ± 0.26
	80 ± 16	2.80 ± 0.23	-100 ± 4.1	0.89 ± 0.31
	322 ± 31	4.83 ± 0.46	-104 ± 5.2	1.99 ± 0.88
	36 ± 9	26.24 ± 0.87	-103 ± 1.2	1.18 ± 0.35

source on kiloparsec scales shows a similar southward core–jet structure [13]. The jet component closest to the VLBI core, at an angular distance of 1.78 mas from the core (Table 2), can be identified with a feature observed in October 1995 (1995.83) at 5 GHz [14]. The proper motion of the component is ~0.035 mas year^{-1}, which corresponds to an apparent projected linear speed of $\beta_{\text{app}} = 1.3h^{-1}$. The speed of this component between 1992 and 1995 was higher, $\beta_{\text{app}} = 4.1h^{-1}$ [14]; this is consistent with the possibility that the jet is decelerated, with the speed of the VLBI component decreasing with distance from the core.

0420+022. 0420+022, which has a flat radio spectrum (Fig. 3), was originally classified as a probable BL Lacertae object [15]. However, its redshift was soon found to be $z = 2.28$ [16]. Such a large redshift is not typical of BL Lacertae objects, whose redshifts, as a rule, do not exceed unity [17]. Following [16], we will consider this source to be a quasar.

RATAN-600 monitoring of the broadband spectrum of this object revealed unusual variability. In some time intervals (on scales of months), strong variability of the total flux density at frequencies below 10 GHz was observed together with weak variability above this frequency, with the variability amplitude increasing with decreasing frequency. VLBI observations were carried out to identify the mechanism responsible for the observed atypical variability. A detailed analysis of this behavior incorporating the results presented here and other VLBI and RATAN-600 observations will appear in a forthcoming paper by Kovalev *et al.*

The LFVN map of the source (Fig. 3) reveals a core—jet structure on milliarcsecond scales. The jet extends westward to a projected distance of about 5 mas (~28 pc) at a level of about 10 mJy.

DA 193. The variable quasar 0552+398 (DA 193, $z = 2.36$) is classified as a GPS source [18], reflecting the fact that its broadband radio spectrum peaks at decimeter-centimeter waves (Fig. 4). DA 193 was observed as a calibrator in our VLBI observations; it is one of the most compact radio sources currently known. Figure 1 shows the (u, v) plane coverage for this source obtained in INTAS99.4. Our 1.66-GHz observations do not detect any jet emission—only a “naked” 0.55-mas VLBI core (Fig. 4). The westward VLBI jet becomes visible at 5 GHz [14] and higher frequencies. The jet components’ speeds measured using 43 GHz VLBI maps turn out to be superluminal, which is not typical of GPS sources [19].

1308+326. This variable, flat-spectrum source ($z = 0.996$; Fig. 5) belongs to a complete sample of radio-bright northern BL Lacertae objects [20], although there are reasons to suppose that the source should, in fact, be classified as a quasar [21]. 1308+326 is a candidate microlensed object. The structure of 1308+326 on kiloparsec scales represents a bright core, a component ∼11″ to the north, and a fainter component ∼6″ to the east [22].

In our experiment, this object was observed to link the emission detected on arcsecond (kiloparsec) and milliarcsecond (parsec) scales. The image (Fig. 5) shows that the source structure at 18 cm is compact, consisting of a VLBI core and two jet components, with the outer component lying 5 mas (30 pc) to the northwest of the core ($\varphi = 54°$). Thus, the jet we have detected is too short to trace the outflow direction on intermediate scales; it may be possible to image this structure using an array with a combination of relatively long and small VLBI baselines (for example, the EVN+MERLIN or the VLBA+NMA).

1345+125. This object ($z = 0.122$) is one of the nearest bright GPS sources (Fig. 6). The host galaxy contains a western and eastern component. The radio source 1345+125 is identified with the western component, which is an elliptical galaxy [23]. We may be observing the merger of two galaxies, which stimulates the activity at radio frequencies [24].

The observations of this object were of interest from the point of view of its classification: it is probably a compact symmetrical object (CSO). Sources of this class are powerful and compact objects with a total size not larger than one kiloparsec. As a rule, the emission is dominated by regions of the jet and hot spots on either sides of the "central engine," and these sources are probably not subject to considerable Doppler brightening [25]. The small sizes of these sources may be a consequence of their youth ($< 10^4$ years). This hypothesis was confirmed after the detection of motion of the hot spots in CSO sources and estimation of their speeds [26]. Most identified CSO sources are also classified as GPS objects based on their broadband radio spectra. Due to the high compactness of these objects, their structure can be resolved only by VLBI observations.

Our 18-cm map of 1345+125 displays a very rich structure. The VLBI jet is detected to distances of up to 50 mas (103 pc) to the southeast of the core in position angle ∼150°. According to Fey *et al.* [27], the core is not detected at 1.6 GHz because of strong self-absorption in the circumnuclear region. Our observations confirm the results of (EVN+GEO) observations of 1345+125 at 8.4, 2.3, and 1.6 GHz [28], and the conclusion that this object is a CSO source. Unfortunately, the lack of data for this source on the short Medvezh'i Ozera–Pushchino baseline has hindered the detection of emission from the most extended regions of the source [29]; in turn, this has resulted in a considerable underestimation of its integrated flux (Table 3).

1803+784. This flat-spectrum source (Fig. 7) is a BL Lacertae object, and is included in the 1-Jy catalog of such objects [20] ($z = 0.68$ [30]).

The purpose of our observations of 1803+784 was to study the structure of this BL Lacertae object and to link the source morphology on parsec and kiloparsec scales. Maps on kiloparsec scales [31, 32] demonstrate the presence of two extended components, one located 2″ to the southwest of the core and the other, weaker, component located approximately 45″ in position angle ∼−165°; there is also very faint emission between these features [33, 34]. The total size of the observed radio structure is ∼180 kpc. On the other hand, VLBI observations with the HALCA orbiting radio telescope [35] made it possible to study the subparsec structure of the source and the direction of the VLBI jet at ∼0.5 mas, close to the central engine. The jet extends first to the northwest in position angle ∼−65° and then turns to the southwest [36].

Figure 7 shows our 18-cm LFVN map of the object. The source has a number of jet components to the southwest of the core in position angle ∼−100°. The most distant component detected on our map is 26 mas (143 pc) from the core. The jet direction and the location of this component are consistent with the results obtained on a more sensitive network (VLBA, VLA, Goldstone) at epoch 1998.55 [37].

Thus, 1803+784 displays a considerable difference in the projected jet directions on parsec and kiloparsec scales (∼100° in the plane of the sky). This can be explained by the effect of projection or interaction of the jet with the surrounding medium.

Modeling of the broadband spectra presented in this paper (see details in [38]) and literature data for VLA observations of these sources indicate the presence of extended radio structures in half of these objects. Table 3 lists the total fluxes of the sources from the LFVN maps and the total fluxes obtained by interpolating the RATAN-600 data. Based on the accuracy of the calibration curves and system temperatures used, the total uncertainties in the integrated fluxes on VLBI scales are ≈10%. We estimate the uncertainties in the total flux densities for the RATAN-600 data to be not larger than 5% (allowing for uncertainties in the interpolation and the measurements themselves; Figs. 2–7).

The excess of the integrated fluxes from the VLBI maps (decaparsec scales) above the RATAN-600 flux densities for 0420+022 and 1308+326 could be due to inaccuracy of the LFVN amplitude calibration and by the fact that the VLBI and RATAN-600 observations were not strictly simultaneous (the minimum interval between the two sets of observations was two months). This may also play some role for 0420+022,

Table 3. Comparison of the total flux densities of the sources measured on the LFVN and the RATAN-600

Instrument	Total flux density at 18 cm, Jy					
	0420−014	0420+022	DA 193	1308+326	1345+125	1803+784
RATAN-600*	2.0	0.76	1.9	2.0	4.7	<2.4
LFVN	1.5[1]	0.96	1.9	2.6	2.0	1.9[2]

* Values obtained by interpolating the 31- and 13-cm data.

[1] The source is significantly resolved on kiloparsec scales.

[2] The epochs of the LFVN and RATAN-600 observations differ by 1 year.

which is variable at low frequencies (this will be described in more detail in a forthcoming paper by Kovalev *et al.*).

4. CONCLUSION

We have presented the results of observations of six extragalactic radio sources on the Low Frequency VLBI Network, involving three Russian and three foreign radio telescopes. We have restored the intensity distributions of the objects with millisecond angular resolution by processing the data using the standard method and with the standard software package. We have discussed our results in the context of our the broadband RATAN-600 spectral observations and previously published EVN and VLBA maps. Comparison of our maps with those from other studies indicates the reliability of the LFVN images and the efficiency of the LFVN. It is desirable to refine the calibration curves of some of the LFVN telescopes to improve the accuracy of amplitude calibration of the data.

We have also obtained positive experience in connection with planning and realizing VLBI experiments. Using the available fully steerable Russian radio telescopes in these experiments helped to maintain them in working condition and to equip them with new radio astronomical instrumentation. The collaboration between the Russian and foreign observatories and the correlation centers allows the Low Frequency VLBI Network to carry out yearly observing sessions aimed at acquiring data for the solution of a broad variety of scientific problems, including observations of active galactic nuclei.

ACKNOWLEDGMENTS

This work was supported by INTAS (grants 96-0183, 2001-0669), the NSFC (grant 19973014), and the Russian Foundation for Basic Research (project no. 03-02-27173). The RATAN-600 measurements were partially supported by the Federal Science and Technology Program in Astronomy (grant 1.2.5.1), the Ministry of Industry, Science, and Technology of the Russian Federation under subcontract to the Special Astrophysical Observatory with the Astro Space Center of the Lebedev Physical Institute, and the Russian Foundation for Basic Research (project no. 02-02-16305). Yu.Yu. Kovalev participated in this work as an NRAO Karl Jansky Postdoctoral Fellow. The authors thank the staff of the observatories and of the Penticton correlation center who made this experiment possible.

REFERENCES

1. I. E. Molotov, S. F. Likhachev, A. A. Chuprikov, *et al.*, *The Universe at Low Radio Frequencies*, Ed. by A. Pramesh Rao, G. Swarup, and Gopal-Krishna, IAU Publ. **199**, 492 (2002).
2. R. D. Wietfeldt, D. Baer, W. H. Cannon, *et al.*, IEEE Trans. Instrum. Meas. **45**, 923 (1996).
3. W. H. Cannon, D. Baer, G. Feil, *et al.*, Vistas Astron. **41**, 297 (1997).
4. B. R. Carlson, P. E. Dewdney, T. A. Burgess, *et al.*, Publ. Astron. Soc. Pac. **111**, 1025 (1999).
5. G. Tuccari, I. Molotov, S. Buttaccio, *et al.*, *Booklet of the 3rd IVS General Meeting, Ottawa, 2004* (Geodetic Survey Div., Natural Resources Canada, 2004), p. 40.
6. A. R. Thompson, J. M. Moran, and G. W. Swenson, Jr., *Interferometry and Synthesis in Radio Astronomy* (Wiley, 2001; Fizmatlit, Moscow, 2003).
7. Yu. Yu. Kovalev, N. A. Nizhelsky, Yu. A. Kovalev, *et al.*, Astron. Astrophys., Suppl. Ser. **139**, 545 (1999).
8. D. C. Gabuzda, T. V. Cawthorne, D. H. Roberts, and J. F. C. Wardle, Astrophys. J. **347**, 701 (1989).
9. D. H. Roberts, J. F. C. Wardle, and I. F. Brown, Astrophys. J. **427**, 718 (1994).
10. J. R. Mattox, R. C. Hartman, and O. Reimer, Astrophys. J., Suppl. Ser. **135**, 155 (2001).
11. D. Sowards-Emmerd, R. W. Romani, and P. F. Michelson, Astrophys. J. **590**, 109 (2003).
12. S. Britzen, A. Witzel, T. P. Krichbaum, *et al.*, Astron. Astrophys. **360**, 65 (2000).
13. R. R. J. Antonucci and J. S. Ulvestad, Astrophys. J. **294**, 158 (1985).
14. X. Y. Hong, T. Venturi, T. S. Wan, *et al.*, Astron. Astrophys. **134**, 201 (1999).

15. M. P. Veron-Cetty and P. Veron, Astron. Astrophys. **412**, 399 (2003).
16. S. L. Ellison, L. Yan, I. M. Hook, *et al.*, Astron. Astrophys. **379**, 393 (2001).
17. M. Stickel, J. W. Fried, and H. Kühr, Astron. Astrophys., Suppl. Ser. **98**, 393 (1993).
18. C. P. O'Dea, S. A. Baum, and C. Stanghellini, Astrophys. J. **380**, 66 (1991).
19. L. M. Lister, A. P. Marscher, and W. K. Gear, Astrophys. J. **504**, 702 (1998).
20. M. Stickel, P. Padovani, C. M. Urry, *et al.*, Astrophys. J. **374**, 431 (1991).
21. D. C. Gabuzda, R. I. Kollgaard, D. H. Roberts, and J. F. C. Wardle, Astrophys. J. **410**, 39 (1993).
22. D. W. Murphy, I. W. A. Browne, and R. A. Perley, Mon. Not. R. Astron. Soc. **264**, 298 (1993).
23. N. A. Shaw, A. K. Tzioumis, and A. Pedlar, Mon. Not. R. Astron. Soc. **256**, 6 (1992).
24. T. M. Hechman, E. P. Smith, and S. A. Baum, Astrophys. J. **311**, 526 (1986).
25. P. N. Wilcinson, A. G. Polatidis, A. C. S. Readhead, *et al.*, Astrophys. J. **432**, L87 (1994).
26. I. Owsianik and J. E. Conway, Astron. Astrophys. **337**, 69 (1998).
27. A. L. Fey, A. W. Clegg, and E. B. Fomalont, Astrophys. J., Suppl. Ser. **105**, 299 (1996).
28. L. Xiang, C. Stanghellini, D. Dallacasa, and Z. Haiyan, Astron. Astrophys. **385**, 768 (2002).
29. M. L. Lister, K. I. Kellermann, R. C. Vermeulen, *et al.*, Astrophys. J. **584**, 135 (2003).
30. A. Witzel, C. J. Schalinski, K. J. Johnston, *et al.*, Astron. Astrophys. **206**, 245 (1988).
31. R. R. J. Antonucci, P. Hickson, E. W. Olszewski, and J. S. Miller, Astron. J. **92**, 1 (1986).
32. R. I. Kollgaard, D. H. Roberts, J. F. C. Wardle, and D. C. Gabuzda, Astron. J. **104**, 1687 (1992).
33. R. G. Strom and P. L. Bierman, Astron. Astrophys. **242**, 313 (1991).
34. P. Cassaro, C. Stanghellini, M. Bondi, *et al.*, Astron. Astrophys. **139**, 601 (1999).
35. H. Hirabayashi, H. Hirosava, H. Kobayashi, *et al.*, Publ. Astron. Soc. Jpn. **52**, 955 (2000).
36. D. C. Gabuzda, New Astron. Rev. **43**, 691 (1999).
37. D. C. Gabuzda and V. A. Chernetskii, Mon. Not. R. Astron. Soc. **339**, 669 (2003).
38. Yu. Yu. Kovalev, Y. A. Kovalev, N. A. Nizhelsky, and A. V. Bogdantsov, Publ. Astron. Soc. Austral. **19**, 83 (2002).

Translated by G. Rudnitskiĭ

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 909–920. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 998–1010.
Original Russian Text Copyright © 2004 by Sakhibov, Smirnov.

The Initial Mass Function and History of the Star-Formation Rate in Star-Forming Complexes in Galaxies

F. Kh. Sakhibov[1,2] and M. A. Smirnov[1]

[1] *Institute of Astronomy, ul. Pyatnitskaya 48, Moscow, 119017 Russia*

[2] *Institute of Astrophysics, Academy of Sciences of Tajikistan, ul. Bukhoro 22, Dushanbe, 734042 Tajikistan*

Received August 28, 2003; in final form, March 15, 2004

Abstract—The positions of star-forming complexes (SFCs) in color—luminosity, color—color, and chemical composition—luminosity diagrams are determined by the star-formation regime (history). Taking into account the fraction of Lyman continuum photons that are not absorbed by hydrogen, we find a strong correlation between the observed color indices and the total Lyman continuum flux from the stars in SFCs. The distribution of extragalactic SFCs in a plot of the slope of the initial mass function (IMF) versus the density of stars cannot be distinguished from this distribution for clusters in the Galaxy and the Large Magellanic Cloud, where the IMF slopes were derived directly from star counts. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

The stellar populations of star-forming complexes (SFCs) in galaxies contain information about the star-formation history and evolution of the chemical abundances in the SFCs, as well as the evolution of the galaxy as a whole. To understand the processes involved in galaxy formation, it is important to interpret the observed characteristics of SFCs in terms of physical parameters such as age, star-formation regime, and the initial mass function (IMF). Two main approaches to this problem have been developed. The first is population synthesis, in which the spectral evolution of a star-forming region is computed based on stellar-evolution theory and databases of stellar spectra for given IMFs, star-formation rates (SFRs), star-formation regimes, and chemical-composition evolutions (cf. [1–5] and references therein). The results depend on the adopted stellar-evolutionary tracks, IMF, and star-formation regime. Within the adopted assumptions, variations in the colors in star-forming regions are usually attributed to variations of age, chemical abundances, and internal reddening.

The second approach is stellar-population synthesis based on the observed parameters of stars and star clusters, or empirical population synthesis [6–9]. This method was most extensively developed in computations of synthetic spectra of galaxies using empirically constructed spectra of star clusters (cf. [7, 10–14] and references therein). The main problem with empirical population synthesis is that the solutions are not unique, due to the need to solve a strongly degenerate set of algebraic equations. The first version of the method [7] used 35 parameters combined in various ratios to calculate nine absorption-line equivalent widths in galactic spectra. Combinations of parameters giving 10% agreement between the nine computed equivalent widths and observations were considered acceptable solutions. The final set of parameters was calculated as the arithmetic mean of all the acceptable solutions.

There has been considerable progress in recent years in combating the lack of a unique solution in empirical population synthesis. Along with reducing the number of parameters from 35 to 12 [11] using the color indices in the stellar continuum [12], and extending the observational data for star clusters toward the far ultraviolet [13], a statistical procedure for finding the most probable solution has been proposed and formalized [14].

Our earlier papers [15, 16] presented an evolutionary synthesis method applied to young SFCs (giant HII zones) in external galaxies. The internal reddening and chemical abundances were determined from spectroscopic observations, making it possible to reduce the number of parameters (basic elements) to three: the slope and upper mass limit of the IMF and the cluster age. There were four computed parameters (integrated colors) to be compared with the observed values. Thus, the set of equations is overdetermined. However, even in this case, different combinations of IMFs, ages, and star-formation scenarios can sometimes correspond to the same combination of the four color indices; this is the so-called IMF—star-formation history degeneracy [17]. To remove this degeneracy, we considered the star-formation history in a complex to correspond with one of two extreme scenarios: an instantaneous burst (IB) or continuous, extended burst (EB) of star formation. This approach does not completely resolve the IMF—star-formation history ambiguity, since other star-formation regimes are possible. However, consideration of these two extreme cases can serve as a first rough approximation to removing the IMF—star-formation history degeneration.

We encountered another ambiguity that is related to the fraction of Lyman continuum photons that do not ionize gas, which can either be absorbed by dust or leave the SFC freely. This fraction can vary strongly from object to object, in the range (10–90)%. The SFC star-formation parameters in [15, 16] were obtained by fixing the fraction of Lyman continuum photons that did not participate in ionizing the HII region, and were accordingly either absorbed by dust or left the SFC freely, ionizing the diffuse interstellar gas, at 30%. It is natural to suppose that the fraction of Lyman continuum quanta that are not absorbed by neutral hydrogen cannot be the same for all SFCs, but instead varies over a wide range for various complexes and various galaxies. It was demonstrated in [18, 19] that up to 50% of the galaxies' combined Hα flux came from diffuse, ionized gas, and was not directly associated with HII regions. Based on a study of the diffuse, ionized gas in six spiral galaxies that demonstrated a spatial correlation between HII regions and the diffuse, ionized gas, it was suggested in [20] that the diffuse, ionized gas was ionized by the Lyman continuum photons that had freely left giant HII regions. The expected equivalent widths of the hydrogen Hβ line in young HII regions is 450–500 Å. However, such HII regions are known from observations to be very rare. This inconsistency between theory and observation also indicates that a large fraction of Lyman photons do not take part in ionization processes.

We used the empirical relation between age and size for star-forming regions, first found in [21], as an additional constraint to avoid these ambiguities. We were not able to remove the ambiguity in the star-formation regime for all SFCs. The remaining SFCs were considered to have no solution in the models considered. The use of the empirical age–size relation together with evolutionary-synthesis models to derive the star-formation parameters enables us to consider this approach to be "empirical population synthesis," an extended evolutionary population-synthesis method. The best objects for studies of stellar populations and their evolution using empirical evolutionary synthesis are young SFCs in spiral and irregular galaxies. At least three photometric observables corrected for reddening and a known chemical composition for the computation of the synthetic values are needed to determine the slope and upper mass limit of the IMF and the age of the SFC. We collected the required observational data for 180 SFCs in 22 spiral and irregular galaxies [22].

The proposed empirical evolutionary-synthesis method is an integrated, indirect method, with larger uncertainties compared to direct star counts, but enables us to determine both the slope and upper mass limit of the IMF, age, and SFR for a large number of objects in distant galaxies, for which direct star counts are not possible. This approach is helpful for studies of systematic differences in the properties of SFCs in different galaxies with different physical conditions. The derived relations of the star-formation characteristics in individual SFCs to local and global properties of the parent galaxy can provide useful input to the theory of star formation in galaxies.

Our previous papers [15, 16] concentrated on the method itself. The present study concerns intercomparisons and interpretations of the resulting IMF and SFR parameters and their application to investigations of star formation in a galaxy as a whole. The goals of the study are (i) to present a modified version of the method for deriving the star-formation parameters from the observed colors of an SFC, taking in into account the possible variations in the fraction of Lyman photons that are not absorbed by hydrogen in individual complexes; (ii) to analyze the method's sensitivity to the star-formation regime in a complex; (iii) to compare the results to IMFs derived directly using star counts in clusters; and (iv) to study the SFR in SFCs. To fulfill these goals, we propose the following plan for presenting our work. Section 2 describes the modified method for determining the IMF and SFR parameters and Section 3 the method's uncertainties. Section 4 discusses the distribution of star-formation regimes for the studied SFCs. The sensitivity of the observed SFC parameters to the star-formation regime is discussed in Section 5. The Lyman continuum luminosities of the SFCs are presented in Section 6. We compare the resulting IMF slopes to the results of direct star counts in Section 7, and apply our results to derive the SFR in the galaxies in Section 8. A discussion of the results is given in Section 9, and the main results and conclusions are summarized in Section 10.

This paper does not discuss the observational material used, since this information is presented and described in detail in [22, 15, 16]. Note that the numbers of objects in the color–luminosity, color–color, and age–size diagrams can vary since observations are not available for all the objects in our sample in all the bands used (U, B, V, and R), and linear dimensions are likewise unknown for some objects. Procedures to correct for peculiarities of the extinction in some SFCs in which absorption in emission lines is systematically in excess of the stars' continuum reddening are described in [23] and discussed in detail in [15]. We use the evolutionary-synthesis models for star clusters presented in detail by Myakutin and Piskunov [24] and discussed in our earlier papers [15, 16].

2. THE MODIFIED METHOD

Our method to derive the IMF and SFR parameters from the observed colors of a star cluster is described in detail in [15, 16]. A combination of colors ($U-B$, $B-V$, $V-R$, LCI) predicted by the evolutionary theory adopted and corresponding to the observed color combinations is identified by searching for the optimal values of the IMF parameters (slope, α, and upper mass limit, $M_{\max}$), age (t), and SFR. When computing the model colors, we fixed the cluster's chemical composition based on the observational data and corrected the observed colors for interstellar reddening. Two extreme cases of star-formation history were considered: an instantaneous burst (IB) and a continuous, extended burst (EB) of star formation. All other possible star-formation regimes are intermediate between these two extreme cases. When deriving the IMF parameters, age, and star-formation regime, we assume that only one of the considered regimes (IB or EB) operates within an individual SFC, namely, that corresponding to the observed spectral energy distribution, luminosity, and size. Problems associated with the lack of a unique star-formation regime in some of the complexes are considered in Section 4. When analyzing the IMF parameters, we treated the IB and EB complexes as separate groups. Those complexes with no acceptable solution or displaying ambiguity in their star-formation regime were not included in the subsequent analysis. We call the IMF and SFR parameters obtained in this way for 100 SFCs in 20 spiral and irregular galaxies the first-approximation parameters. The first-approximation star-formation parameters derived in [15, 16] are collected in Table 1 for the entire sample of 100 SFCs, without separation based on their star-formation regimes.

Note that SFCs with an IB star-formation regime are, on average, much younger than those with a continuous, EB regime. We demonstrated in [16] that the ages of complexes with an EB regime are correlated with their luminosities and linear sizes:

$$t \approx 5.37 \times 10^{-2} \times 10^{-(0.14\pm0.02)M_B}, \qquad (1)$$
$$t \approx 0.11 S^{1.02\pm0.30},$$

where t is the complex's age in millions of years, M_B is its absolute magnitude, and S is its linear size in parsecs. A correlation between the duration of the burst of star-formation and the size of the star-forming region was found earlier by Efremov and Elmegreen [21] for star clusters in the LMC.

We found for the IB SFCs that high-metallicity regions ($Z > 3Z_\odot$) show an inverse correlation between their ages and sizes. IB complexes with normal chemical composition show a direct, though weak, correlation between their ages and sizes. In this paper, we do not consider the high-metallicity SFCs, all of which show IB star formation, for the following reasons.

(1) There are no evolutionary models available for clusters with metallicities higher than three times the solar value, and the high-metallicity SFCs were compared to the existing highest-metallicity ($Z = 2.35Z_\odot$) model.

(2) About 30% of the high-metallicity SFCs in our sample belong to the peculiar interacting galaxy system NGC4038/39, in which a strong burst of star formation is observed [25]. The intense star formation in this system leads to an increased supernova rate, which may be partially responsible for the ionization of the HII regions. Thus, the observed line ratios are biased by the presence of an additional ionization source, and can lead to overestimation of the metal abundance in regions with normal chemical composition.

The first-approximation star-formation parameters were derived by fixing the fraction of Lyman-continuum photons that were partially absorbed by dust or left the SFC freely to be 50%. In this new study, we reject this assumption, and treat the fraction $1-f$ of Lyman continuum photons that did not participate in the ionization of the HII region as a free parameter that can vary from 0 to almost 100%. The modified procedure for determining the refined (second-approximation) IMF and SFR parameters can be divided into two stages.

In the first stage, we simultaneously derived the IMF and SFR parameters from the observed colors of a star cluster as described in [15, 16], using $1-f$ values from 0 to 90%, with a 10% increment. We treated $1-f$ as a free parameter at this stage. We thus obtained ten combinations of the IMF parameters (α, $M_{\max}$) and SFR parameters (t, regime) corresponding to the ten adopted $1-f$ values. Here, the star-formation regime can take on one of two possible values: instantaneous burst (IB) or extended-burst (EB).

In the second stage, the best set of parameters is selected from among the ten sets of IMF (α, $M_{\max}$) and SFR (t, regime) parameters based on additional constraints: the empirical age–absolute magnitude (t–M_B) and age–size (t–S) relations (1), which were independently confirmed in [21]. Using the refined

Table 1. First-approximation star-formation parameters for 100 SFCs in 20 galaxies

Parameter	Range	Mean	Standard deviation	Uncertainty
IMF slope, α	$-0.5\ldots-4.35$	-2.42	0.91	0.51
$M_{\max}$, $M_\odot$	30–120	74	20	33
$\log t$ (years)	5.9–8	6.84	0.55	0.29

estimates of the ages and star-formation regimes, we determined new empirical age–luminosity and age–size relations and then repeated the second stage of the procedure for the IMF and SFR parameters. The process of determining the star-formation parameters and the coefficients of the empirical relations (1) converged, and no further iterations were needed.

Below, we present the statistical characteristics of the distributions of the IMF parameters and ages separately for the IB and EB complexes. Complexes with ambiguous star-formation regimes (see Section 4) were not included in our subsequent analysis of the star-formation parameters. The new SFC age estimates occupy a narrower range than the first-approximation estimates (from $\log t = 5.9$ to $\log t = 7.5$). The IB complexes have ages from 0.8 to 6.0 million years, with a mean age of $1.8^{+1.4}_{-0.8}$ million years. The ages for the EB complexes range from 6.2 to 31.6 million years, with a mean age of $14.1^{+7.3}_{-5.8}$ million years. In both cases, the age of a complex is taken to be the entire period of the complex's existence, rather than the burst duration. The physical interpretation of the age gap between the IB and EB complexes is considered in Section 9. The IMF slope varies in the range -0.5 to -4.0, with the mean values $\alpha = -2.81 \pm 0.79$ for the IB complexes and $\alpha = -2.40 \pm 0.52$ for the EB complexes. Our estimates of the upper mass limit of the IMF range from 45 $M_\odot$ to 120 $M_\odot$ for the IB complexes and from 60 $M_\odot$ to 110 $M_\odot$ for the EB complexes, with a mean of 80 $M_\odot$ and standard deviation of $\sigma_{M_{\max}(obs)} = 20\ M_\odot$.

3. UNCERTAINTIES OF THE METHOD

We discussed the uncertainties of our method in [15, 16]. These include both the observational uncertainties and uncertainties in the empirical relations between the absorption of the light emitted by gas and by stars in SFCs [23], and between the chemical abundances and the observed relative intensities of the oxygen and nitrogen lines [26], as well as uncertainties in the calibrations used to convert the ratios of monochromatic continuum fluxes into broadband color indices in the standard $UBVR$ system [22]. Additional errors can result from disregarding the contribution of radiation by gas to the stellar continuum [26], and also in the estimates of the fraction of Lyman continuum photons that do not take part in ionizing the HII region. In [26], we used a semiempirical model to estimate the contribution of radiation by gas to various stellar-continuum bands for 96 SFCs (HII regions) in the galaxies NGC2403, NGC2903, NGC4038/39, and NGC5194. It turns out that this contribution does not exceed 7 Å even for the interacting galaxies NGC4038/39, in which a burst of star-formation is occurring. For the remaining galaxies without star-formation bursts, the contribution of radiation by gas to the stellar-continuum bands is less than 5 Å. The influence on the star-formation parameters derived from the observed color indices of uncertainties in the color indices ($\sigma_{obs} = 0.15^m$–0.25^m), metallicities ($\sigma_Z = 50\%$), and interstellar reddening ($A_v \leq 0.30^m$) is discussed in detail in [15, 16].

Since the IMF represents the statistical mass distribution of the stars, having a small number of stars in an SFC leads to uncertainties in the IMF parameters. In complexes with IMF slopes less steep than $\alpha = -3.32$, for the stellar luminosity function $L^*(m)$ [27], 80% of the luminosity L_B is due to stars with masses $m > 4\ M_\odot$. The uncertainties in the input photometric data are 15–20% [22]. Thus, the estimated slopes can be considered reliable only for the high-mass ($m\ >\ 4\ M_\odot$) part of the IMF. If a small number of massive stars produce 80% of the flux in a low-luminosity complex with a flat IMF, the upper-mass limit $M_{\max}$ can also be uncertain. $M_{\max}$ can be considered reliable only when the number of stars with masses exceeding the upper limit of the IMF is $N_p\ (m\ >\ M_{\max}) \geq 1$.

Otherwise, the IMF estimates can represent chance deviations from the mass distribution of the stars in the SFC. In [15], we used the condition that at least three stars have masses in the interval ($M_{\max}$, $M_{\max} + 30\ M_\odot$) for a given luminosity of the SFC as a criterion for the trustworthiness of the IMF parameters α and $M_{\max}$. Taking into account variations of the fraction $1 - f$ of Lyman-continuum photons that do not ionize the gas in the SFC reduced the initial uncertainties of the parameters, which become $\sigma_\alpha = 0.35 \pm 0.02$ for the IMF slope,

$\sigma_{M_{\max}} = 10 \pm 1\ M_\odot$ for the IMF upper-mass limit, and $\log t = 0.20 \pm 0.02$ for the SFC age.

4. STAR-FORMATION REGIMES

We noted in [15] that it was either not possible to find a good solution within the two star-formation regimes considered or the solution was ambiguous for 67 of the 180 SFCs studied. In ambiguous cases, when there were acceptable solutions for both regimes, we treated the SFCs as having no good solution. The reliability criterion adopted for the IMF parameters was not satisfied for 12 SFCs, and these objects were also excluded from our further analysis. The absence of acceptable solutions for some SFCs could be due to several reasons.

First, the colors of some of these objects are outside the range permitted by the theoretical models [15, Fig. 1]. For example, in the theory, $U - B$ cannot be bluer than -1.2 and $B - V$ cannot be bluer than -0.5. Similarly, according to the theory, $B - V$ values redder than $+0.2$ should correspond to redder $U - B$ values than those observed for these SFCs. The observed SFC colors could be outside the theoretically permitted range due to observational uncertainties, incorrect reddening corrections, or unidentified foreground stars overlapping the SFC image. These stars must be rather faint, but they can strongly bias the color indices.

Second, some objects can have solutions for both star-formation regimes. This means that the object's observed spectral energy distribution corresponds to the IB regime for one IMF and age, and to the EB regime for some other IMF and age. In other words, there is no unique solution. This is clearly visible in Fig. 1, which displays the distribution of star-formation regimes for the studied SFCs. We have chosen to represent the IB regime with the number 1 and the EB regime with the number 2. These two regimes are extremes among the possible star-formation regimes, and the SFCs with no solution for these two regimes probably have a more complex SFR history, and must be considered as intermediate cases between the IB and EB regimes. Accordingly, we have assigned all SFCs with no acceptable solution the number 1.5. There are not complete sets of input parameters for all of the 180 objects, and the absence of $U - B$ values for some of them leads to ambiguity and hence the impossibility of finding an acceptable solution. This color index is sensitive to the star-formation regime, and the availability of $U - B$ data makes it possible to refine the solutions obtained; i.e., to make them more definite. For this reason, the fraction of SFCs with no solutions is smaller for objects with $U - B$ measurements ($56/145 = 38.6\%$). This fraction for the SFCs without $U - B$ measurements is $23/35 = 65.7\%$, almost twice as high.

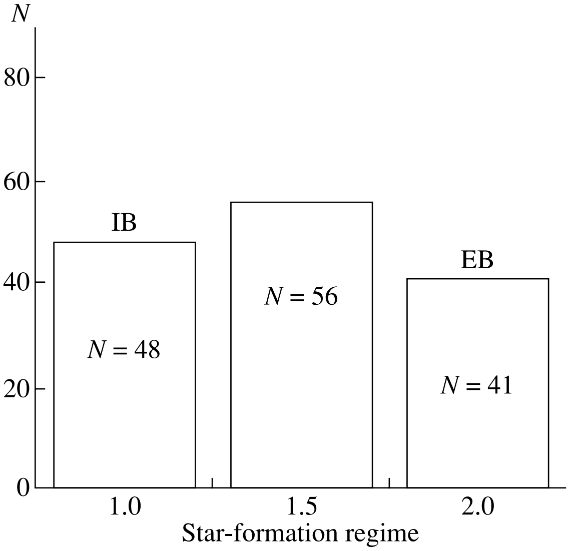

Fig. 1. Distribution of the star-formation regimes for SFCs with observed $U - B$ color indices. Objects with abnormally high metallicities are not considered here (see the text).

Third, objects with abnormally high metallicities display abnormally blue colors (both $U - B$ and $B - V$). This could be due to the use of incorrect reddening corrections. The reddening–absorption relation was derived for interstellar dust with a normal chemical composition. The absorption ratios at different wavelengths may not be the same in the presence of high heavy-element abundances. Our semiempirical model for the ratio of the extinctions of radiation by gas and by stars [23] may also not be fully applicable in the case of high metal abundances. In addition, many of these SFCs are heavily absorbed, further complicating the determination of the correct reddening corrections.

Fourth, there exist physical reasons for the lack of an acceptable solution for some objects. We have considered two extremely simple star-formation regimes, with all stars formed simultaneously t years ago (IB) or the star formation beginning t years ago and continuing until the present (EB). The real situation is probably different from this simple model. The stars in an SFC began to form t years ago and were formed during some interval Δt. Depending on $\Delta t/t$, we consider the star formation to have been instantaneous or to continue until the present. If $\Delta t/t \ll 1$, we have the IB regime, while, if $\Delta t/t \approx 1$, we have the EB regime. If the ratio has an intermediate value, our approach does not yield a unique solution.

All three possibilities are about equally probable; i.e., we expect to obtain no unique solution in approx-

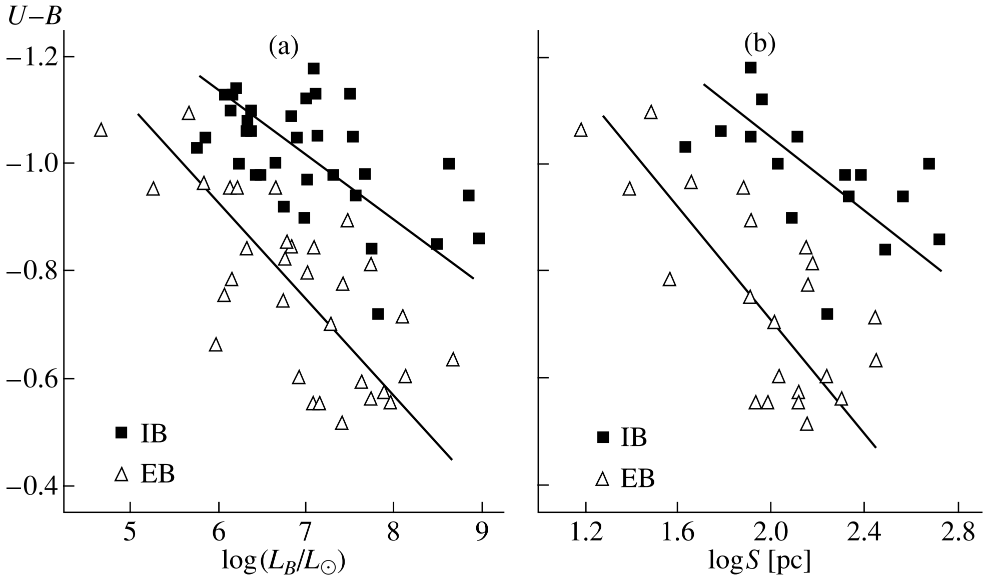

Fig. 2. (a) $U - B$ vs. L_B and (b) $U - B$ vs. S (pc) diagrams for IB SFCs (squares) and EB SFCs (triangles).

imately one-third of cases, or slightly more often, as is observed (Fig. 1).

5. SENSITIVITY OF THE OBSERVED SFC PARAMETERS TO THE STAR-FORMING REGIME

Let us consider the differences between the SFCs with the two different star-formation regimes and plot model-independent diagrams for their observed parameters. We begin with the distributions of the SFCs in color–luminosity and color–linear size diagrams (Fig. 2). We can see in Fig. 2 that the IB SFCs are systematically bluer than the EB SFCs at a given luminosity (size). The low accuracy of the observations and our choice of considering two distinct star-formation regimes lead to a low correlation for the relations shown in Fig. 2 ($r \approx 0.5$–0.7). The coefficients for all the empirical relations derived in this study are presented in Table 2.

The trend toward redder color indices for increasing luminosity (size) that is characteristic of both IB and EB SFCs is in agreement with the age–luminosity and age–linear size relations derived earlier [16, 21], as well as with the age–$U - B$ relation (see Fig. 9 below; the relationship between Fig. 2 and Fig. 9 will be discussed in Section 9). The actual difference between the SFCs with different star-formation regimes is readily visible in the $(U - B)$–$(B - V)$ two-color diagram for the averaged color indices (Fig. 3). The curve for the EB SFCs displays a constant displacement along the $U - B$ axis; i.e., these SFCs have redder $U - B$ values for a given $B - V$. The displacement is 0.1^m – 0.2^m, and is approximately equal to the dispersion of the averaged indices (and to the uncertainties in the observed color indices). This difference primarily indicates that these SFCs have a different distribution of stellar spectral types. The $U - B$ color index reflects the ratio of the number of O + early B stars to the number of late B + early A stars, while $B - V$ reflects the ratio of the numbers of B and A stars. Thus, the shift in the diagram provides clear evidence that the two types of SFCs have different ages, assuming that the stellar IMF is a power law without any breaks.

The third difference in the parameters of the SFCs with different star-formation regimes is presented in Fig. 4, which shows that the SFCs with high heavy-element abundances always display the IB regime, never the EB regime. Either star-formation regime can occur in complexes with low (solar or lower) metallicities. The direct relation between metal abundance and luminosity was first noted by us in [16].

6. LYMAN-CONTINUUM LUMINOSITIES

The number of Lyman-continuum photons emitted by a cluster's stars is usually estimated from the flux in the Balmer lines observed from the HII region. However, as is discussed above, a considerable fraction of the Lyman-continuum photons may not participate in the ionization of this gas, possibly resulting in underestimation of the number of Lyman-continuum photons emitted by the cluster stars. In turn, this leads to inconsistencies in the cluster's spectral energy distribution. One of the parameters determined in our modified method for deriving the IMF and SFR parameters from the observed

Table 2. Empirical relations and correlation coefficients for the SFCs

Relation	Regime	Correlation equation	r	Figure
$(U-B)$–$\log\left(\frac{L_B(obs)}{L_\odot}\right)$	IB	$y = -(1.86 \pm 0.09) + (0.12 \pm 0.03)x$	0.57	2a
$(U-B)$–$\log\left(\frac{L_B(obs)}{L_\odot}\right)$	EB	$y = -(2.00 \pm 0.12) + (0.18 \pm 0.03)x$	0.70	2a
$(U-B)$–$\log S$ [pc]	IB	$y = -(1.75 \pm 0.10) + (0.35 \pm 0.14)x$	0.55	2b
$(U-B)$–$\log S$ [pc]	EB	$y = -(1.82 \pm 0.14) + (0.54 \pm 0.12)x$	0.71	2b
$\log\left(\frac{N_{Lyc}}{L_B}\right)$–$(U-B)$	IB	$y = (6.71 \pm 0.06) - (4.15 \pm 0.10)x$	0.99	5
$\log\left(\frac{N_{Lyc}}{L_B}\right)$–$(U-B)$	IB	$y = (8.92 \pm 0.11) - (1.97 \pm 0.13)x$	0.94	5
$\log$ SFR [$M_\odot$ yr^{-1}]–$\log S$ [pc]	IB+EB	$y = -(10.15 \pm 0.80) + (3.82 \pm 0.44)x$	0.80	7
$\log t$ [years]–$\log S$ [pc]	IB	$y = (4.59 \pm 0.22) + (0.77 \pm 0.19)x$	0.58	9a
$\log t$ [years]–$\log S$ [pc]	EB	$y = (5.85 \pm 0.013 + (0.64 \pm 0.10)x$	0.78	9a
$\log t$ [years]–$(U-B)$	IB	$y = (8.29 \pm 0.09) + (2.06 \pm 0.18)x$	0.89	9b
$\log t$ [years]–$(U-B)$	EB	$y = (7.95 \pm 0.12) + (1.08 \pm 0.16)x$	0.77	9b
$\log S$ [pc]–$\log\left(\frac{L_B(obs)}{L_\odot}\right)$	IB+EB	$y = -(0.51 \pm 0.13) + (0.36 \pm 0.02)x$	0.92	10

integrated colors of an SFC is the fraction of Lyman-continuum photons that are not absorbed by hydrogen. Using this fraction and the flux in the Balmer lines, we can predict the Lyman-continuum flux from a particular SFC. The Lyman-continuum fluxes for the sample of SFCs predicted in this way are strongly correlated with the integrated $U-B$ colors (Fig. 5).

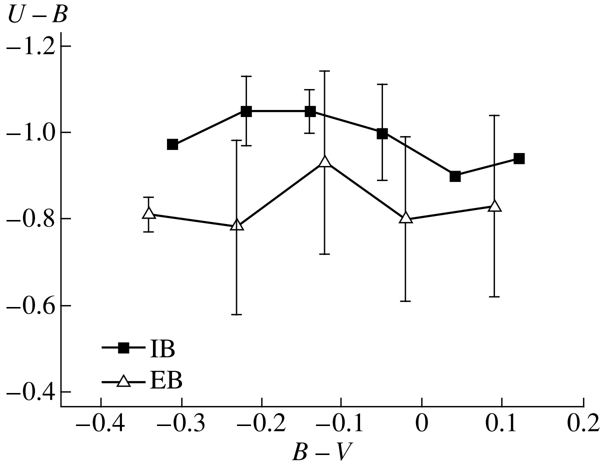

Fig. 3. Two-color diagram for the averaged colors of the SFCs.

7. COMPARISON OF DIRECT AND INDIRECT IMF-SLOPE ESTIMATES

The SFCs analyzed here include two objects in the Large Magellanic Cloud (LMC) whose IMF slopes have been estimated directly using star counts: 30 Doradus and Dem 152 [28–32]. Our indirect

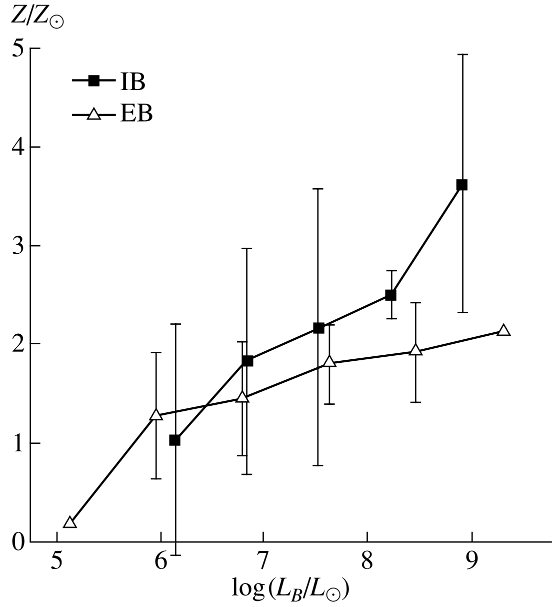

Fig. 4. Metallicity as a function of the SFC luminosity.

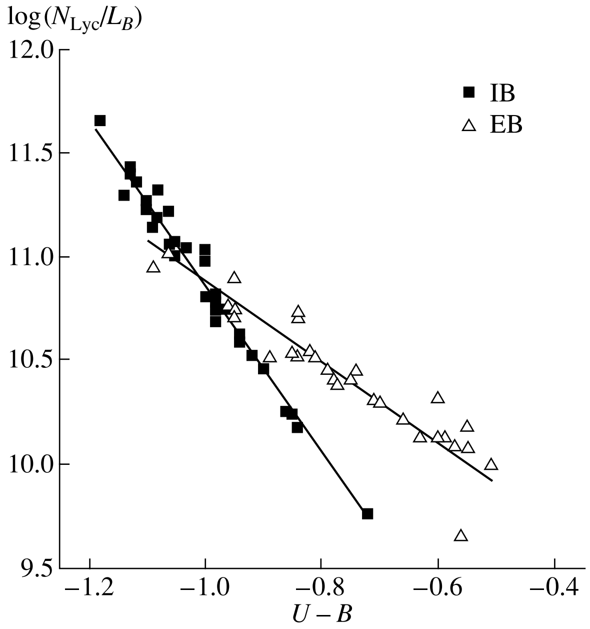

Fig. 5. Ratio of the predicted number of Lyman-continuum photons to the B luminosity versus $U-B$ for IB SFCs (crosses) and EB SFCs (triangles).

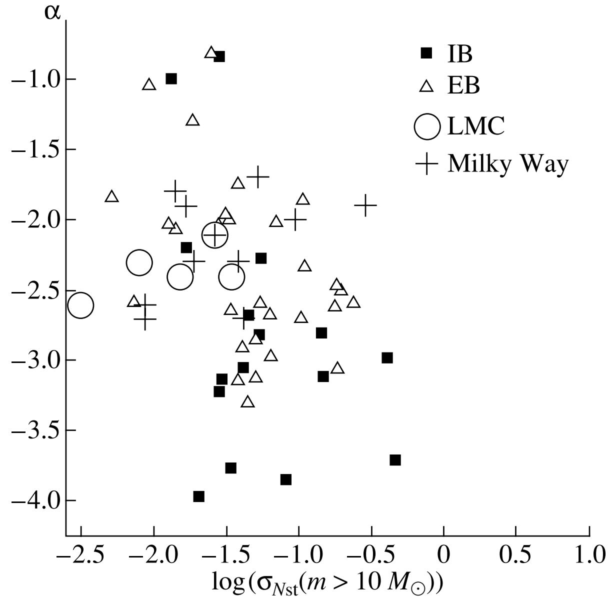

Fig. 6. IMF slope plotted against the surface number density of stars with masses exceeding 10 $M_\odot$ for IB SFCs (squares), EB SFCs (triangles), associations in the Galaxy (crosses), and associations in the LMC (circles) for which direct star counts are available.

IMF slopes, $\alpha = -2.00 \pm 0.35$ for 30 Dor and $\alpha = -2.70 \pm 0.35$ for Dem 152, are close to the IMF slopes derived directly: $\alpha = -2.2 \pm 0.3$ and $\alpha = -2.3 \pm 0.3$, respectively. To compare our IMF slope estimates and direct estimates derived from star counts, we plotted our objects on the plot of the IMF slope versus the surface density of stars with masses $m > 10\ M_\odot$ from [33].

The surface density of stars with masses exceeding 10 $M_\odot$, $\sigma_{N_{\rm st}}(m > 10\ M_\odot)$, can be determined from the IMF parameters, age, star-formation regime, integrated luminosity, and linear size of the SFC.

In the IB SFCs, all the stars were formed simultaneously t years ago, and the number of stars with masses in excess of 10 $M_\odot$ will be

$$N_{\rm st}(m > 10\ M_\odot) = A \int\limits_{10\ M_\odot}^{M_{\rm max}} m^\alpha dm, \qquad (2)$$

where the constant A is determined from the equation

$$L_B(\text{initial, at } t = 0) \qquad (3)$$

$$= A \int\limits_{M_{\rm min}}^{M_{\rm max}} L^*(m) m^\alpha dm.$$

On the left-hand side of (3), we have the SFC's B luminosity at the initial time, $t = 0$. $L^*(m)$ is the known B stellar luminosity function [27], m is a star's mass, α is the slope of the IMF, and $M_{\rm max}$ is the upper mass limit of the IMF. For the given stellar luminosity function, $L^*(m)$ [27], and an IMF slope less steep than $\alpha = -3.32$, the main contribution to the cluster's B luminosity function (80%) comes from stars with masses exceeding 4 $M_\odot$; we accordingly took the lower limit of the integral in (3) to be $M_{\rm min} = 4\ M_\odot$. The luminosity L_B at time $t = 0$ is determined using the evolutionary model for the cluster adopted here, together with the observed absolute magnitude, $M_B(obs)$, and the derived IMF parameters and age, t. Finally, the surface number density of stars with masses exceeding 10 $M_\odot$ is determined from the relation

$$\sigma_{N\rm st}(m > 10\ M_\odot) = \frac{N_{\rm st}(m > 10\ M_\odot)}{\text{SFC area (pc}^2)}. \qquad (4)$$

The total number of stars formed in EB SFCs can be determined from the SFR, $N_{\rm st}$ ($m > 10\ M_\odot$ per year), multiplied by the age of the SFC, t. The star-formation rate was taken to be the ratio of the number of stars formed at the initial time (2) to the lowest possible age of the EB SFC given by the evolutionary model. The surface number density of stars with masses exceeding 10 $M_\odot$ was derived as the total number of stars with masses exceeding 10 $M_\odot$ formed during the time t divided by the area of the SFC in square parsecs. Thus, the surface density of stars, $N_{\rm st}$ ($m > 10\ M_\odot$), is a function of the luminosity, size,

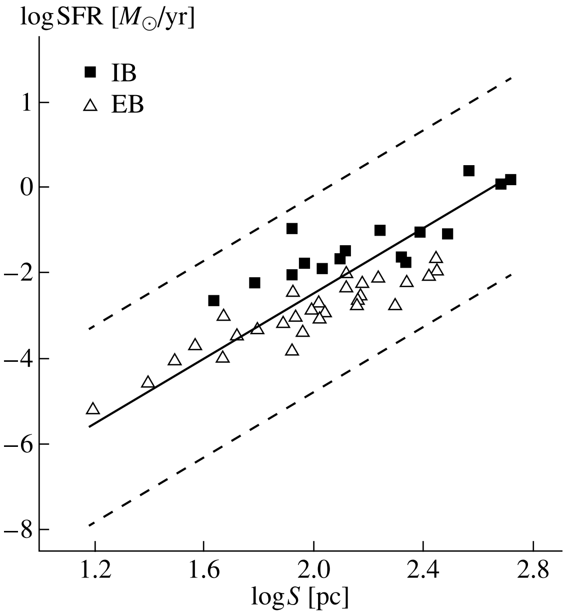

Fig. 7. SFR as a function of the linear size, S (pc), for IB SFCs (squares) and EB SFCs (triangles).

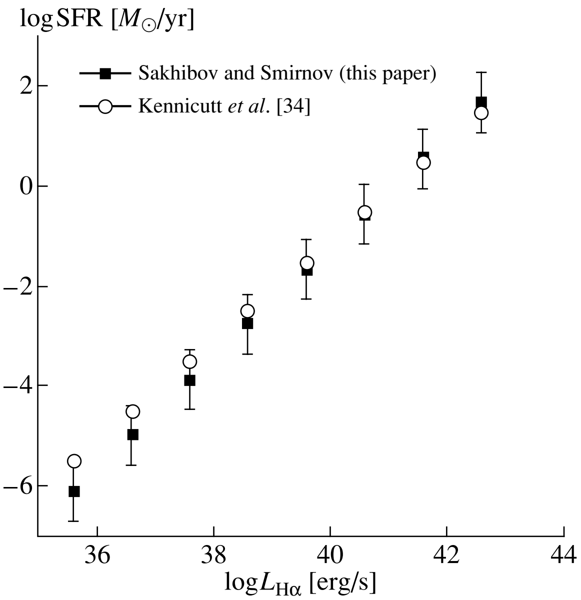

Fig. 8. Comparison of the SFR calibration from the SFR–S (size in pc) relation to the calibration of Kennicutt *et al.* [34] based on measured Hα fluxes.

IMF, SFR, age, and the star-formation regime of the complex.

Figure 6 shows the IMF slope plotted against the surface number density of stars with masses in excess of 10 $M_\odot$. We compare the direct IMF slope estimates derived from star counts for star clusters in the Milky Way (crosses) and LMC (circles) to our indirect estimates (squares and triangles). We can see from Fig. 6 that the distribution of objects with indirect IMF estimates is indistinguishable from that for the Milky Way and LMC star clusters for which there are direct estimates from star counts. Both groups of objects cover the same range, with the density varying over approximately a factor of 200. Figure 6 confirms the conclusion of [33] that there is no correlation between the IMF slope and the star density in SFCs.

8. SFC STAR-FORMING RATE

The SFR in the complexes was derived as the ratio of the total mass of stars formed at the initial time for the adopted IMF and the minimum lifetime of the SFC determined by the evolutionary model. We computed the total mass of initially-formed stars using the IMF, age, star-formation regime, and the luminosity of the SFC determined as is described above, similar to the technique used in the previous section to estimate the number of stars in the SFC. The SFR plotted against the size of the SFC is shown in Fig. 7. Though the IB SFCs (crosses) show systematically higher SFRs than the EB SFCs (triangles) for a given size, the distributions for both types of complexes are satisfactory described by a single relation (the solid line) with the correlation coefficient $r = 0.80$. The dotted lines delineate an area of three standard deviations from this correlation.

The SFR estimates for the SFCs are independent of their observed linear sizes. Thus, the correlation between the SFRs and the SFC sizes may relate the diameter distributions of HII regions found for many galaxies and the star formation rates in the disks of these galaxies. Figure 8 compares the calibration of the star-formation rates based on the SFR–size relation and the calibration of Kennicutt *et al.* [34], which is based on Hα fluxes.

9. DISCUSSION

The separation of the SFCs with different star-formation regimes in the $U - B$–luminosity or $U - B$–size diagrams appears natural in our approach, as we can see in the age–size and age–$U - B$ diagrams in Fig. 9. As noted above, the age–size relation for star-forming regions was first established by Efremov and Elmegreen [21] for LMC star clusters and then in [16] for SFCs in other galaxies.

The two different relations for the IB SFCs and EB SFCs in the diagrams in Fig. 9 reflect the existence of the two regimes. For a given age or luminosity, the IB complexes are considerably younger than the EB complexes. The observed shift in the two-color diagram (Fig. 3) provides additional evidence for the age difference between the IB and EB SFCs.

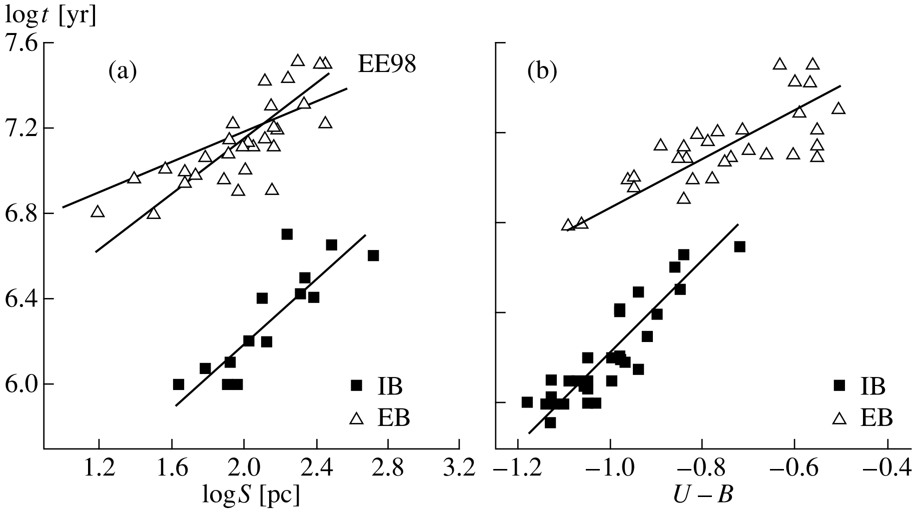

Fig. 9. SFC ages as a function of their (a) linear sizes S (pc) and (b) $U - B$ values for IB SFCs (squares) and EB SFCs (triangles). The dotted line labeled EE98 was derived independently by Efremov and Elmegreen [21] for star clusters in the LMC.

This age gap can be explained in a natural way. A region with a continuous star-formation regime and an age of less than six million years will contain most of the first-generation high-mass (>40 $M_\odot$) stars. The SFR is insufficient to distort the IMF of the forming stars within such a short time interval, and the color characteristics of the complex will not differ from those for an SFC with an instantaneous star-formation regime: our approach will assign such a complex to the IB SFC group. The typical gas velocities in an individual HII region are 15–30 km/s—the velocity of hydrogen atoms at a temperature of 10 000 K, as well as the stellar-wind velocity. For such velocities, the lifetime of an HII region with a typical size of about 100 pc is three to seven million years. For this region to exist for a longer time, a gas reservoir is needed, with a wave of star formation propagating across it; i.e., a continuous star-formation regime will be observed in this case. Thus, IB SFCs will be observed as HII regions until a time equal to the lifetime of an isolated region of ionized hydrogen, four to six million years. Older IB SFCs can no longer contain giant regions of ionized hydrogen.

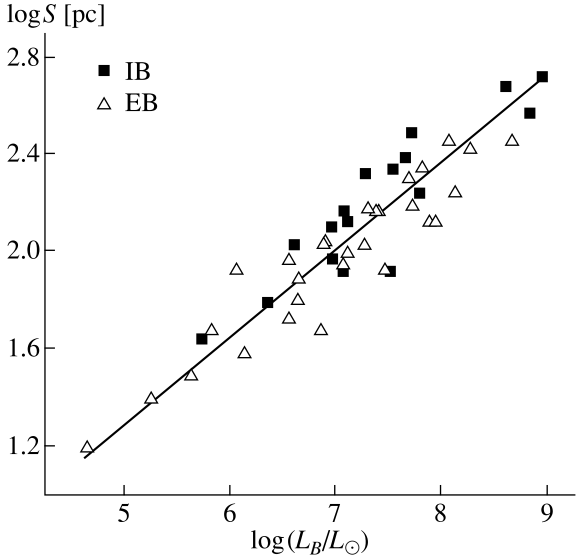

Fig. 10. SFC linear size as a function of SFC luminosity.

Our sample includes only those complexes that are giant HII regions, and thus our approach is not capable of finding IB SFCs older than six to eight million years. This is probably one of the physical distinctions between the regions with instantaneous and extended star-formation bursts. The presence of two different relations in Fig. 9 should distinguish the IB and EB complexes on this model-independent color–luminosity diagram. This conclusion is justified if there exists a direct relation between the linear size and luminosity that is the same for both IB and EB SFCs. Figure 10 demonstrates that the linear size–luminosity (S (pc)–L_B) relation for both SFC types is satisfactory described by the same line, with a correlation coefficient of $r = 0.92$.

An additional factor should be taken into account when discussing the different star-formation regimes in SFCs, having to do with the computed rather than the observed parameters. The IB complexes are younger than the EB complexes. In our models, this means that the burst duration, Δt, is different for SFCs with different star-formation regimes: about one to two million years or less for the IB SFCs (considerably less than the age of the complex) or several (or tens of) million years for the EB SFCs (comparable to the age). In other words, our study indicates that the difference between the regimes is real. The star-formation burst is either very short or more prolonged. The SFCs with no acceptable solutions are probably close to the EB group. Their burst of star formation is over, but the B stars providing the Lyman photons are still alive. The ages t of some of the EB SFCs exceed 20 million years, and only some of them have star-formation-burst durations that long. For most complexes that old, star formation was finished some 10 million years ago, though it was underway for approximately the same amount of time ($\Delta t \approx 10$ million years). Stars of different generations in the upper part of the mass spectrum will have enough time to form during this period. A generation lasts only several million years for OB stars.

In this scenario, which seems realistic, the star-formation history for some of the complexes should not give a unique solution corresponding to one of the two considered regimes. We can imagine that, if we translated the observed SFC colors into star-formation parameters using a discrete series of models between the instantaneous-burst and extended-burst star-formation models considered here, we would fill the area between the two relations in the age–size and age–color diagrams with a discrete series of similar relations. The color–luminosity diagram would demonstrate a gradual transition from the instantaneous-burst to the extended-burst regime. We may return to this problem once the needed series of theoretical models is available. The processes regulating the IMF and the SFR history in the complexes are so diverse and complicated that they cannot be described fully in a single unified model. We have considered two extreme star-formation scenarios in order to cover the whole space of models for more complex star-formation regimes. The IB and EB models we used should be considered tools to remove the ambiguity in the IMF and SFR regime.

10. CONCLUSIONS

Our main conclusions and results are the following.

For the first time, we have been able to distinguish SFCs with different star-formation histories in color–luminosity and two-color diagrams. The position of an SFC in the color–luminosity (size) diagram depends on the star-formation history (regime). Instantaneous-burst (IB) complexes are systematically bluer than extended-burst (EB) complexes, for a given luminosity (size). In the $(U-B)$–$(B-V)$ diagram, the EB complexes have redder $U-B$ color indices than the IB complexes, for the same $B-V$ color. The high-metallicity SFCs display only the instantaneous star-formation regime, never extended star formation. Complexes with low metalliticies (solar and lower) can have either star-formation regime.

The differences we have found in the distributions of SFCs with different star-formation regimes in three model-independent diagrams for the observed parameters of SFCs provide direct confirmation of the sensitivity of empirical evolutionary synthesis to the star-formation regime. The method's sensitivity to different star-formation regimes enabled us to find for the first time a formal way to resolve the IMF–SFR-history ambiguity when determining the star-formation parameters of the complexes. The fraction of Lyman-continuum photons that do not participate in the ionization of the HII region (and were accordingly either absorbed by dust or freely left the SFC, ionizing the diffuse interstellar gas) was varied over a wide range, from 10% to 90% in individual complexes, in agreement with the recent observations of emission from interstellar diffuse gas [18–20]. The strong correlation between the predicted number of Lyman-continuum photons and $U-B$ can be used to estimate the fraction, $1-f$, of Lyman-continuum photons that leave the HII region or are absorbed in the complex itself, contributing to the galaxies' infrared luminosity.

We found a single correlation between the star-formation rate and the linear size of the SFCs for both star-formation regimes, making it possible to study the distribution of the star-formation rate in galactic disks using a single observable—the linear size. This opens wide possibilities for comparisons between the radial distributions of the star-formation rate and of matter and luminosity in the disks, as well as for searches for connections between the distribution of the star-formation rate and the dynamic properties of the disks and the spiral structure.

REFERENCES

1. B. M. Tinsley, Astron. Astrophys. **20**, 383 (1972).
2. R. B. Larson and B. M. Tinsley, Astrophys. J. **219**, 46 (1978).
3. M. Fioc and R. Rocca-Volmerange, Astron. Astrophys. **326**, 1 (1997).
4. G. Bruzual and S. Charlot, Astrophys. J. **405**, 538 (1993).
5. C. Leitherer *et al.*, Astrophys. J., Suppl. Ser. **123**, 3 (1999).
6. H. Spinrad and B. J. Taylor, Astrophys. J., Suppl. Ser. **22**, 445 (1971).
7. E. Bica, Astron. Astrophys. **195**, 76 (1988).
8. D. Pelat, Mon. Not. R. Astron. Soc. **299**, 877 (1998).
9. C. Boisson, M. Joly, J. Moulaka, *et al.*, Astron. Astrophys. **355**, 99 (2000).
10. E. Bica and D. Alloin, Astron. Astrophys. **186**, 49 (1987).

11. A. A. Schmidt, M. V. F. Copetti, D. Alloin, and P. Jublonka, Mon. Not. R. Astron. Soc. **249**, 766 (1991).
12. C. Bonatto, E. Bica, M. G. Pastoriza, and D. Alloin, Astron. Astrophys. **355**, 99 (2000).
13. C. Bonatto, E. Bica, and D. Alloin, Astron. Astrophys., Suppl. **112**, 71 (1995).
14. R. Cid Fernandes, L. Sodre, H. Schmitt, and J. Leao, Mon. Not. R. Astron. Soc. **325**, 60 (2001).
15. F. Sakhibov and M. A. Smirnov, Astron. Astrophys. **354**, 802 (2000).
16. F. Kh. Sakhibov and M. A. Smirnov, Astron. Zh. **78**, 3 (2001) [Astron. Rep. **45**, 1 (2001)].
17. J. M. Scalo, Fundam. Cosmic. Phys. **11**, 1 (1986).
18. C. G. Hoopes, S. T. Gottesman, and B. E. Greenwalt, Astron. J. **112**, 1429 (1996).
19. M. S. Oey and R. Kennicut, ASP Conf. Ser. **131**, 322 (1998).
20. M. Rozas, A. Zurita, and J. E. Beckman, Astron. Astrophys. **354**, 823 (2000).
21. Yi. N. Efremov and B. Elmegreen, Mon. Not. R. Astron. Soc. **299**, 588 (1998).
22. F. Kh. Sakhibov and M. A. Smirnov, Astron. Zh. **76**, 419 (1999) [Astron. Rep. **43**, 361 (1999)].
23. F. Kh. Sakhibov and M. A. Smirnov, Astron. Zh. **72**, 318 (1995) [Astron. Rep. **39**, 281 (1995)].
24. É. A. Piskunov and V. I. Myakutin, Astron. Zh. **73**, 520 (1996) [Astron. Rep. **40**, 472 (1996)].
25. D. E. Rigopoulou, D. Lutz, R. Genzel, *et al.*, Rev. Mex. Astron. Astrofis. Ser. Conf. **6**, 87 (1997).
26. F. Kh. Sakhibov and M. A. Smirnov, Astron. Zh. **67**, 472 (1990) [Sov. Astron. **34**, 347 (1990)].
27. C. W. Allen, *Astrophysical Quantities, University of London* (Athlone Press, 1973).
28. A. Nota, M. Sirianni, C. Leitherer, *et al.*, STScI Newsletter **15** (4), 2 (1998).
29. D. A. Hunter, E. J. Shaya, J. A. Holtzman, *et al.*, Astrophys. J. **448**, 179 (1995).
30. D. A. Hunter, E. J. O'Nel, R. Lynds, *et al.*, Astrophys. J. **459**, L27 (1996).
31. B. Brandl, B. J. Sams, F. Bertoldi, *et al.*, Astrophys. J. **466**, 254 (1996).
32. M. S. Oey and P. Massey, Astrophys. J. **452**, 210 (1995).
33. P. Massey, K. E. Johnson, and K. Degioia-Eastwood, Astrophys. J. **454**, 151 (1995).
34. R. C. Kennicutt, Jr., P. Tamblyn, and C. W. Congdon, Astrophys. J. **435**, 22 (1994).

Translated by N. Samus'

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 921–933. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 1011–1023.
Original Russian Text Copyright © 2004 by Popov, Ustyugov, Chechetkin.

Development of the Geometric Structure of the Thermonuclear-Deflagration Front in Type Ia Supernovae

M. V. Popov, S. D. Ustyugov, and V. M. Chechetkin

Keldysh Institute of Applied Mathematics, Russian Academy of Sciences,
Miusskaya pl. 4, Moscow, 125047 Russia

Received March 4, 2004; in final form, March 15, 2004

Abstract—Three-dimensional hydrodynamical simulations of the development of a large-scale instability accompanying deflagration in the degenerate cores of rotating white dwarfs—progenitors of type-Ia supernovae—are presented. The numerical algorithm used is described in detail. An explicit, conservative, Godunov-type TVD difference scheme was employed for the computations. Large-scale convective processes are important as the deflagration front propagates. The supernova explosion is strongly nonspherically symmetric; a large-scale front structure emerges and propagates most rapidly along the rotational axis. The arrival of fresh thermonuclear fuel to the central region of the core can result in flares and the destruction of the core. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

It is usual to classify supernovae according to their optical spectra. Supernovae are classified as type I (SNI) if hydrogen lines are absent from their spectra and type II (SNII) if hydrogen lines are present. In turn, SNI's are subdivided into types Ia, Ib, and Ic. SNIa's are distinguished by the presence of a Si absorption line at about 6150 Å in the period following the explosion and by strong Fe emission lines at later times. Conversely, Si lines are absent from the initial spectra of SNIb's and SNIc's. Relatively strong He lines (especially near 5876 Å) are typical of SNIb's, while they are not observed (or are very faint) in SNIc's. Type II, Ib, and Ic supernovae are believed to be the products of the explosions of massive single stars (SNII) or binary systems (SNIb and SNIc). The progenitors of SNIa's are white dwarfs with masses close to the Chandrasekhar limit, $M_{ch} \sim 1.44\, M_\odot$ [1], consisting of a mixture of carbon and oxygen nuclei and a highly degenerate electron–positron gas. We study type Ia supernovae here.

The numerous unresolved issues in the theory of stellar evolution raise some questions in connection with the choice of an initial model for the presupernova star. Various SNIa models have been proposed. One type of model involves explosions in binary systems where the secondary in the system is either a similar degenerate white dwarf or an evolved red giant. Models for explosions of white dwarfs with masses $M \sim M_{ch}$ or $M < M_{ch}$ are also considered in the literature. Various evolutionary scenarios have been suggested for binary systems, including the absorption of the white dwarf by the secondary or the accretion of hydrogen and helium onto the white dwarf due to the transfer of matter from the secondary [2]. It is necessary to introduce binary systems as SNIa progenitors, since, otherwise, the white dwarf, whose initial mass is $\sim 0.6\, M_\odot$, cannot acquire the mass $\sim M_{ch}$ needed for the explosion. As the mass approaches the Chandrasekhar limit, the temperature in the central region of the star grows, providing the conditions required for the ignition of thermonuclear reactions and the propagation of a deflagration front. Numerical simulations of this process are rather difficult, since the deflagration front travels in an extended medium, with important roles played by gravity and, in the case of a rotating star, centrifugal forces. These factors are prerequisites for the development of various large-scale disturbances. Since some parameters of the process are indeterminate, it is possible to consider various scenarios for the explosion [2, 3].

The usual scheme for the classification of supernovae is based on the spectra of their envelopes, while the formation of the dense central cores in such stars is determined by the evolution of the progenitors. This is a complex process controlled by many factors—the character of the rotation, the presence of heavy elements, magnetic fields, convective processes, etc. A fully consistent evolutionary theory has not yet been developed, and the formation of a dense carbon–oxygen core surrounded by a massive envelope seems quite possible. In this case, the supernova could be observed as an SNIb, SNIc, or SNII [4].

We describe here our three-dimensional numerical simulations of the convective instability for the thermonuclear-deflagration front in the degenerate matter of a white dwarf with mass $M \sim M_{ch}$. We neglect the influence of the secondary in the system; i.e., we do not include its gravitational field or mass exchange between the stars. Our results can also be used to describe mechanisms for other types of supernova explosions provided their progenitors have dense carbon–oxygen cores.

2. FORMULATION OF THE PROBLEM

After a star with a mass of less than $8M_\odot$ leaves the main sequence in the course of its evolution and becomes a red giant, a degenerate carbon–oxygen core begins to form. This core is a developing white dwarf. Energy is released in the star by shell energy sources, with the innermost shell being located at the core–envelope interface. The mass of the core increases due to the inflow of deflagration products from this interface. The temperature and density of the core gradually increase if nuclear reactions do not occur. Ultimately, a temperature of $\sim 3 \times 10^8$ K is reached at the center of the star, and deflagration of the CO mixture sets in. The development of a thermal carbon–oxygen flash begins, which eventually results in the thermonuclear explosion of the star. During the deflagration, an “iron” core forms, which consists of the products of thermonuclear burning of carbon and oxygen; energy is released, increasing the entropy of the core material [5, 6]. Thermal and mechanical equilibrium is disrupted in the system, and pulsations begin to develop, giving rise to large-scale hydrodynamic instabilities, as we show below. It is important that, as shown by Imshennik *et al.* [7], a deflagration regime is characteristic of thermonuclear burning of a degenerate carbon–oxygen core.

We consider the rotating core of a presupernova in which the development of thermal instability in a deflagration regime has begun. The inner portion of the core consists of iron-peak elements (we denote this region the “iron core”) and an outer CO layer. The outer helium layer does not appreciably affect the processes in the central regions, and we did not include it in our computations. According to the modern theory of stellar evolution [8], the mass of the CO layer together with the iron core is about 1.5 $M_\odot$, close to the Chandrasekhar limit. We will assume that the mass fractions of carbon and oxygen are equal. The ratio of the rotational energy T to the gravitational energy W is $T/|W| = 0.01$, which corresponds to an angular rotational rate of $\Omega_0 = 2.0732$ s^{-1} [9]. The radius of the star is $R_0 = 1.5 \times 10^8$ cm, and the density at the center is $\rho_0 = 2 \times 10^9$ g/cm^3. We will use a tabulated equation of state for fully ionized matter $p = p(\rho, S)$ that describes the electron–positron component in terms of Fermi–Dirac statistics using various asymptotics and the ion component in an ideal-gas approximation [10].

3. EQUILIBRIUM CONFIGURATION

We construct an equilibrium configuration for a rotating CO sphere with constant entropy $S = S_0 =$ const. It follows from the thermodynamic relationship $TdS = dH - Vdp$ that, when $S =$ const, the enthalpy $H = \int dp/\rho$ depends solely on the density: $H = H(\rho)$. In this case, the equilibrium equation has the form [11]

$$H + \Phi_g - \int_0^R \Omega^2 \tilde{\omega} d\tilde{\omega} = C, \qquad (1)$$

where Φ_g is the gravitational potential, Ω is the angular rotational velocity, $\tilde{\omega}$ is the distance from the rotational axis, and $C =$ const. In a rigid-rotation approximation, $\Omega = \Omega_0 =$ const, Eq. (1) can be written

$$H + \Phi_g - \Omega_0^2 \frac{\tilde{\omega}^2}{2} = C. \qquad (2)$$

The constant C can be found from (2) when $\tilde{\omega} = 0$:

$$C = H(\tilde{\omega} = 0) + \Phi_g(\tilde{\omega} = 0). \qquad (3)$$

We calculate the gravitational potential using the formula [11]

$$\Phi_g = -G \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' = -4\pi G \int_0^\infty dr' \qquad (4)$$

$$\times \int_0^1 d\mu' \sum_{n=0}^\infty f_{2n}(r, r') P_{2n}(\mu) P_{2n}(\mu') \rho(\mu', r'),$$

where

$$f_{2n}(r, r') = \begin{cases} (r')^{2n+2}/r^{2n+1}, & r' < r, \\ r^{2n}/(r')^{2n-1}, & r' > r, \end{cases}$$

$\mu = \cos\theta$ is the cosine of the angle to the rotational axis, and $P_{2n}(\mu)$ are the Legendre polynomials. We will use the following algorithm to calculate the distributions of the density $\rho = \rho(r, \theta)$ and gravitational potential $\Phi_g = \Phi_g(r, \theta)$ of the rotating sphere.

(1) Using the known spherically symmetric density distribution $\rho = \rho(r)$ as a zeroth approximation, we calculate Φ_g from (4).

(2) Since $H = H(\rho)$ is a known function when $S =$ const, we find C from (3).

(3) We compute $H = H(r, \theta)$ using (2).

(4) Based on the tabulated function $H = H(\rho)$, we obtain $\rho(r, \theta)$ from $H(r, \theta)$.

(5) We calculate $\Phi_g(r, \theta)$ using (4).

We then return to step (2) and repeat the cycle until the condition $|C^n - C^{n-1}|/|C^n| < 1 \times 10^{-12}$ is satisfied; here, C^n is the nth iteration for the constant C.

The spherically symmetric density distribution $\rho = \rho(r)$ can be obtained for a nonrotating CO sphere by solving the system of equations

$$\begin{cases} \dfrac{dp}{dr} = -\rho \dfrac{Gm}{r^2}, \\ \dfrac{dm}{dr} = 4\pi\rho r^2, \end{cases} \tag{5}$$

where p is the pressure and m is the mass of the sphere of radius r. For a barotropic equation of state $p = P(\rho)$, (5) assumes the form

$$\begin{cases} \dfrac{dp}{dr} = \dfrac{1}{dP/d\rho}\left(-\rho \dfrac{Gm}{r^2}\right), \\ \dfrac{dm}{dr} = 4\pi\rho r^2 \end{cases}$$

and can be solved using a Runge–Kutta method with the boundary conditions $\rho = \rho_0$ and $m = 0$ at the center.

Thus, we have constructed a stable equilibrium configuration for a rotating CO sphere with constant entropy, $S = S_0$.

4. DEFLAGRATION

We will consider here and below a carbon–oxygen core with a burned central region consisting of iron-peak elements. The boundary of the iron core is a deflagration front, whose width we will neglect. We will also neglect the propagation speed of the deflagration front, since it is much smaller than the sound speed.

The heat release Q due to the reaction

$$2\,{}^{12}_{6}\mathrm{C} + 2\,{}^{16}_{8}\mathrm{O} \longrightarrow {}^{56}_{26}\mathrm{Fe} + Q$$

is determined as $Q = (2m_\mathrm{C} + 2m_\mathrm{O} - m_\mathrm{Fe})c^2$. The mass $m(A, Z)$ of the element ${}^{A}_{Z}\mathrm{X}$ can be calculated using the formula $m(A,Z)c^2 = (A - Z)m_n c^2 + Zm_p c^2 - Q_b(A,Z)$, where Q_b is the bond energy:

$$Q_b({}^{12}_{6}\mathrm{C}) = 0.92165 \times 10^8 \text{ eV},$$
$$Q_b({}^{16}_{8}\mathrm{O}) = 1.27624 \times 10^8 \text{ eV},$$
$$Q_b({}^{56}_{26}\mathrm{Fe}) = 4.92280 \times 10^8 \text{ eV}.$$

The burned region of the core has an increased entropy. This disrupts the equilibrium and gives rise to a large-scale convective instability, which we will study by solving the appropriate three-dimensional hydrodynamic equations. It can be shown that the lowest perturbation modes should develop most rapidly [12]. The core and envelope are described by different equations of state: $P = P_\mathrm{Fe}(\rho, S)$ and $P = P_\mathrm{CO}(\rho, S)$. We will use the function $S = S(r, \theta)$ as the initial data for our calculations, which undergoes a jump at the interface. In a first approximation, we can assume that the functions describing the density, $\rho = \rho(r, \theta)$, and gravitational potential, $\Phi_g = \Phi_g(r, \theta)$, are continuous at $t = 0$ and coincide with those obtained for a CO sphere.

5. HYDRODYNAMICAL EQUATIONS

If the stellar material can be considered a compressible inviscid fluid, the hydrodynamical equations in Eulerian variables are:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \operatorname{div}\mathbf{m} = 0, \\ \dfrac{\partial m_i}{\partial t} + \dfrac{\partial \Pi_{ik}}{\partial x_k} = \rho g_i, \\ \dfrac{\partial(\rho S)}{\partial t} + \operatorname{div}(\rho S \mathbf{v}) = 0, \end{cases} \tag{6}$$

where $\Pi_{ik} = P\delta_{ik} + \rho v_i v_k$, $\mathbf{m} = \rho\mathbf{v}$ is the momentum, and $\mathbf{g}$ is the free-fall acceleration. We chose a spherical coordinate system in which $(x_1, x_2, x_3) = (r, \theta, \phi)$ and the Lame coefficients are $(h_1, h_2, h_3) = (1, r, r\sin\theta)$. Formulas for the divergences of a vector and of a second-rank tensor in a curvilinear coordinate system are given in the Appendix. It is convenient to represent the system (6) in a divergence form. To this end, we introduce a symbolic density vector $\mathbf{w}$, which appears in the second derivative with respect to time, and the density-flux vectors $\mathbf{F(w)}$, $\mathbf{G(w)}$, and $\mathbf{H(w)}$, which appear in the spatial partial derivatives:

$$\mathbf{w}_t + \frac{1}{r^2}\frac{\partial}{\partial r}(r^2\mathbf{F}) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\mathbf{G}) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}\mathbf{H} = \mathbf{S}, \tag{7}$$

where

$$\mathbf{w} = (\rho, m_r, m_\theta, m_\phi, \rho S)^T,$$
$$\mathbf{F} = (m_r, p + \rho v_r^2, \rho v_\theta v_r, \rho v_\phi v_r, m_r S)^T,$$
$$\mathbf{G} = (m_\theta, \rho v_r v_\theta, p + \rho v_\theta^2, \rho v_\theta v_\phi, m_\theta S)^T,$$
$$\mathbf{H} = (m_\phi, \rho v_r v_\phi, \rho v_\theta v_\phi, p + \rho v_\phi^2, m_\phi S)^T,$$
$$\mathbf{S} = (0, \rho g_r + \frac{1}{r}[2p + \rho(v_\theta^2 + v_\phi^2)],$$
$$\rho g_\theta + \frac{1}{r}(p + \rho v_\phi^2)\cot\theta - \frac{1}{r}\rho v_r v_\theta,$$
$$\rho g_\phi - \frac{\rho v_\phi}{r}(v_r + v_\theta\cot\theta), 0)^T.$$

The system (7) is hyperbolic, so that the Jacobians $\mathcal{A} = \partial\mathbf{F}/\partial\mathbf{w}$, $\mathcal{B} = \partial\mathbf{G}/\partial\mathbf{w}$, and $\mathcal{C} = \partial\mathbf{H}/\partial\mathbf{w}$ have a full set of left and right eigenvectors corresponding to real eigenvalues. The matrices $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ have the form

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ (p_\rho)_s - \dfrac{S}{\rho}(p_s)_\rho - v_r^2 & 2v_r & 0 & 0 & \dfrac{1}{\rho}(p_s)_\rho \\ -v_r v_\theta & v_\theta & v_r & 0 & 0 \\ -v_r v_\phi & v_\phi & 0 & v_r & 0 \\ -v_r S & S & 0 & 0 & v_r \end{pmatrix},$$

$$\mathcal{B} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ -v_\theta v_r & v_\theta & v_r & 0 & 0 \\ (p_\rho)_s - \dfrac{S}{\rho}(p_s)_\rho - v_\theta^2 & 0 & 2v_\theta & 0 & \dfrac{1}{\rho}(p_s)_\rho \\ -v_\theta v_\phi & 0 & v_\phi & v_\theta & 0 \\ -v_\theta S & 0 & S & 0 & v_\theta \end{pmatrix},$$

$$\mathcal{C} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ -v_\phi v_r & v_\phi & 0 & v_r & 0 \\ -v_\phi v_\theta & 0 & v_\phi & v_\theta & 0 \\ (p_\rho)_s - \dfrac{S}{\rho}(p_s)_\rho - v_\phi^2 & 0 & 0 & 2v_\phi & \dfrac{1}{\rho}(p_s)_\rho \\ -v_\phi S & 0 & 0 & S & v_\phi \end{pmatrix}.$$

The following eigenvalues form the solution of the characteristic equation $\det|\lambda E - \mathcal{A}| = 0$: $\lambda_1 = v_r + c, \lambda_2 = v_r - c, \lambda_3 = \lambda_4 = \lambda_5 = v_r$, where $c = \sqrt{(p_\rho)_s}$ is the sound speed. The solution of the equation $\mathcal{A}\mathbf{e}_i = \lambda_i \mathbf{e}_i$ yields the corresponding right eigenvectors $\mathbf{e}_i$:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ v_r + c \\ v_\theta \\ v_\phi \\ S \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 1 \\ v_r - c \\ v_\theta \\ v_\phi \\ S \end{pmatrix},$$

$$\mathbf{e}_3 = \begin{pmatrix} 1 \\ v_r \\ 0 \\ 0 \\ -\rho\dfrac{(p_\rho)_s}{(p_s)_\rho} + S \end{pmatrix}, \mathbf{e}_4 = \begin{pmatrix} 0 \\ 0 \\ v_\theta \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ v_\phi \\ 0 \end{pmatrix}.$$

Similarly, we have for the matrix $\mathcal{B}$ $\lambda_1 = v_\theta + c$, $\lambda_2 = v_\theta - c$, $\lambda_3 = \lambda_4 = \lambda_5 = v_\theta$, and

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ v_r \\ v_\theta + c \\ v_\phi \\ S \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 1 \\ v_r \\ v_\theta - c \\ v_\phi \\ S \end{pmatrix},$$

$$\mathbf{e}_3 = \begin{pmatrix} 1 \\ 0 \\ v_\theta \\ 0 \\ -\rho\dfrac{(p_\rho)_s}{(p_s)_\rho} + S \end{pmatrix}, \mathbf{e}_4 = \begin{pmatrix} 0 \\ v_r \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ v_\phi \\ 0 \end{pmatrix}.$$

For the matrix $\mathcal{C}$, we find $\lambda_1 = v_\phi + c$, $\lambda_2 = v_\phi - c$, $\lambda_3 = \lambda_4 = \lambda_5 = v_\phi$, and

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ v_r \\ v_\theta \\ v_\phi + c \\ S \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 1 \\ v_r \\ v_\theta \\ v_\phi - c \\ S \end{pmatrix},$$

$$\mathbf{e}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ v_\phi \\ -\rho\dfrac{(p_\rho)_s}{(p_s)_\rho} + S \end{pmatrix}, \quad \mathbf{e}_4 = \begin{pmatrix} 0 \\ v_r \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{e}_5 = \begin{pmatrix} 0 \\ 0 \\ v_\theta \\ 0 \\ 0 \end{pmatrix}.$$

6. NUMERICAL SIMULATIONS

To find a numerical solution to system (7), we must break up the computational domain into cells. The solution will be a piecewise-smooth function. We reference the density vector $\mathbf{w}$ to the cell centers and the density-flux vectors $\mathbf{F}, \mathbf{G}$ and $\mathbf{H}$ to their boundaries. The method of Roe [13] was used to construct the finite-difference scheme.

Let us consider two adjacent cells in which the state of the substance is characterized by the quantities $\mathbf{w}_L$ and $\mathbf{w}_R$, which are close to some mean state $\mathbf{w}$. We expand $\Delta\mathbf{w} = \mathbf{w}_R - \mathbf{w}_L$ in the eigenvectors of each of the matrices $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$; i.e., we find the coefficients α_1, α_2, α_3, α_4, and α_5 satisfying the relationship

$$\Delta\mathbf{w} = \sum_{i=1}^{5} \alpha_i \mathbf{e}_i. \tag{8}$$

We obtain the following algebraic system for the matrix $\mathcal{A}$:

$$\begin{cases} \Delta\rho = \alpha_1 + \alpha_2 + \alpha_3, \\ \Delta(\rho v_r) = \alpha_1\left(v_r + \sqrt{(p_\rho)_s}\right) \\ \ + \alpha_2\left(v_r - \sqrt{(p_\rho)_s}\right) + \alpha_3 v_r, \\ \Delta(\rho v_\theta) = (\alpha_1 + \alpha_2 + \alpha_4)v_\theta, \\ \Delta(\rho v_\phi) = (\alpha_1 + \alpha_2 + \alpha_5)v_\phi, \\ \Delta(\rho S) = \alpha_1 S + \alpha_2 S + \alpha_3\left(-\rho\dfrac{(p_\rho)_s}{(p_s)_\rho} + S\right). \end{cases} \tag{9}$$

Since $\mathbf{w}_L$ and $\mathbf{w}_R$ are close to some mean $\mathbf{w}$, we can use the expression $\Delta(\rho U) = U\Delta\rho + \rho\Delta U + O(\Delta^2)$, where $U = v_r, v_\theta, v_\phi$, or S. Then, to second order, the solution of (9) is

$$\begin{cases} \alpha_1 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho + \dfrac{1}{2}\rho\Delta v_r\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_2 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho - \dfrac{1}{2}\rho\Delta v_r\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_3 = -\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_4 = \dfrac{\rho}{v_\theta}\Delta v_\theta - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_5 = \dfrac{\rho}{v_\phi}\Delta v_\phi - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S. \end{cases} \tag{10}$$

Similarly, the coefficients α_1, α_2, α_3, α_4, and α_5 satisfying (8) for the matrix $\mathcal{B}$ have the form

$$\begin{cases} \alpha_1 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho + \dfrac{1}{2}\rho\Delta v_\theta\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_2 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho - \dfrac{1}{2}\rho\Delta v_\theta\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_3 = -\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_4 = \dfrac{\rho}{v_r}\Delta v_r - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_5 = \dfrac{\rho}{v_\phi}\Delta v_\phi - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S. \end{cases} \tag{11}$$

We have for the matrix $\mathcal{C}$

$$\begin{cases} \alpha_1 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho + \dfrac{1}{2}\rho\Delta v_\phi\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_2 = \dfrac{1}{2}\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S + \dfrac{1}{2}\Delta\rho - \dfrac{1}{2}\rho\Delta v_\phi\dfrac{1}{\sqrt{(p_\rho)_s}}, \\ \alpha_3 = -\dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_4 = \dfrac{\rho}{v_r}\Delta v_r - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S, \\ \alpha_5 = \dfrac{\rho}{v_\theta}\Delta v_\theta - \dfrac{(p_s)_\rho}{(p_\rho)_s}\Delta S. \end{cases} \tag{12}$$

We now consider two radially adjacent cells. We can easily check that the coefficients (10) also satisfy the relationship

$$\Delta\mathbf{F} = \sum_{i=1}^{5} \lambda_i \alpha_i \mathbf{e}_i, \tag{13}$$

for these cells, where $\Delta\mathbf{F}$ is the difference of the radial density fluxes for the states $\mathbf{w}_L$ and $\mathbf{w}_R$:

$$\Delta\mathbf{F} = (\Delta(\rho v_r), \Delta p + \Delta(\rho v_r^2), \Delta(\rho v_\theta v_r), \\ \Delta(\rho v_\phi v_r), \Delta(\rho v_r S))^T, \\ \Delta(\rho U_1 U_2) = U_1 U_2 \Delta\rho + \rho U_1 \Delta U_2 \\ + \rho U_2 \Delta U_1 + O(\Delta^2), \\ (U_{1,2} = v_r, v_\theta, v_\phi \text{ or } S).$$

Equations (8) and (13) are satisfied for $\mathbf{w}_L$ and $\mathbf{w}_R$ that are close to some mean state $\mathbf{w}$. Let us now assume that $\mathbf{w}_L$ and $\mathbf{w}_R$ are arbitrary; instead of (8) and (13), we require that

$$\begin{cases} \Delta\mathbf{w} = \sum_{i=1}^{5} \alpha'_i \mathbf{e}'_i, \\ \Delta\mathbf{F} = \sum_{i=1}^{5} \lambda'_i \alpha'_i \mathbf{e}'_i, \end{cases} \tag{14}$$

where

$$\lambda'_{1,2,3,4,5} = v'_r + c', v'_r - c', v'_r, v'_r, v'_r,$$

$$\mathbf{e}'_1 = \begin{pmatrix} 1 \\ v'_r + c' \\ v'_\theta \\ v'_\phi \\ S' \end{pmatrix}, \quad \mathbf{e}'_2 = \begin{pmatrix} 1 \\ v'_r - c' \\ v'_\theta \\ v'_\phi \\ S' \end{pmatrix},$$

$$\mathbf{e}'_3 = \begin{pmatrix} 1 \\ v'_r \\ 0 \\ 0 \\ -\rho' \dfrac{(p_\rho)'_s}{(p_s)'_\rho} + S' \end{pmatrix},$$

$$\mathbf{e}'_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}'_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

$$\alpha'_1 = \frac{1}{2}\frac{(p_s)'_\rho}{(p_\rho)'_s}\Delta S + \frac{1}{2}\Delta\rho + \frac{1}{2}\rho'\Delta v_r \frac{1}{\sqrt{(p_\rho)'_s}},$$

$$\alpha'_2 = \frac{1}{2}\frac{(p_s)'_\rho}{(p_\rho)'_s}\Delta S + \frac{1}{2}\Delta\rho - \frac{1}{2}\rho'\Delta v_r \frac{1}{\sqrt{(p_\rho)'_s}},$$

$$\alpha'_3 = -\frac{(p_s)'_\rho}{(p_\rho)'_s}\Delta S, \quad \alpha'_4 = \rho'\Delta v_\theta - \frac{(p_s)'_\rho}{(p_\rho)'_s} v'_\theta \Delta S,$$

$$\alpha'_5 = \rho'\Delta v_\phi - \frac{(p_s)'_\rho}{(p_\rho)'_s} v'_\phi \Delta S.$$

We have factored out v'_θ and v'_ϕ from $\mathbf{e}'_4$ and $\mathbf{e}'_5$ to avoid indeterminacy of α'_4 and α'_5 at $v'_\theta = 0$ and $v'_\phi = 0$. Upon writing (14) explicitly and solving the resultant algebraic system of equations in ρ', v'_r, v'_θ, v'_ϕ, and S', we can show that [14, 15]

$$\rho' = \sqrt{\rho_L \rho_R}, \quad U' = \frac{\sqrt{\rho_L} U_L + \sqrt{\rho_R} U_R}{\sqrt{\rho_L} + \sqrt{\rho_R}}, \qquad (15)$$

$$U = v_r, \quad v_\theta, v_\phi, S.$$

To derive (15), we used the relationship

$$\Delta p = (p_\rho)'_s \Delta\rho + (p_s)'_\rho \Delta S.$$

Similarly, using (11) and (12) instead of (10), we can show that Eqs. (15) are valid not only for cells adjacent in radius but also for cells adjacent in θ or ϕ.

In the calculations, the derivatives of the pressure $(p_\rho)'_s$ and $(p_s)'_\rho$ must be computed according to the formulas [16]

$$(p_\rho)'_s = (p_\rho)''_s \left(1 + \frac{(p_\rho)''_s \Delta\rho}{\left((p_s)''_\rho \Delta S\right)^2 + \left((p_\rho)''_s \Delta\rho\right)^2} \delta p\right),$$

$$(p_s)'_\rho = (p_s)''_\rho \left(1 + \frac{(p_s)''_\rho \Delta S}{\left((p_s)''_\rho \Delta S\right)^2 + \left((p_\rho)''_s \Delta\rho\right)^2} \delta p\right),$$

where

$$\delta p = \Delta p - \left((p_\rho)''_s \Delta\rho + (p_s)''_\rho \Delta S\right).$$

The derivatives $(p_\rho)''_s$ and $(p_s)''_\rho$ can be determined from the tabulated equation of state $p = P(\rho, S)$ in terms of the known quantities ρ' and S'.

In the Roe scheme, the density fluxes at the boundary between two adjacent cells (e.g., those with radial numbers i and $i+1$) are calculated using the formula

$$\mathbf{F}_{i+1/2} = \frac{\mathbf{F}_i + \mathbf{F}_{i+1}}{2} - \frac{1}{2}\sum_{i=1}^{5} |\lambda'_i| \alpha'_i \mathbf{e}'_i. \qquad (16)$$

To construct the finite-difference scheme, we must carry out some simple manipulation of the original system (7). We write the equations for m_θ and m_ϕ explicitly:

$$\frac{\partial m_\theta}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \rho v_r v_\theta) \qquad (17)$$

$$+ \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left[\sin\theta(p + \rho v_\theta^2)\right] + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}(\rho v_\theta v_\phi)$$

$$= \rho g_\theta + \frac{1}{r}\left(p + \rho v_\phi^2\right)\cot\theta - \frac{\rho v_r v_\theta}{r},$$

$$\frac{\partial m_\phi}{\partial t} + \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \rho v_r v_\phi) \qquad (18)$$

$$+ \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta \rho v_\theta v_\phi) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}(p + \rho v_\phi^2)$$

$$= \rho g_\phi - \frac{\rho v_\phi}{r}(v_r + v_\theta \cot\theta).$$

It is convenient to introduce the term $-1/r\rho v_r v_\theta$ in (17) into the differentiated quantity:

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \rho v_r v_\theta) + \frac{\rho v_r v_\theta}{r} = \frac{1}{r^3}\frac{\partial}{\partial r}\left(r^3 \rho v_r v_\theta\right).$$

Similarly, we have for (18):

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \rho v_r v_\phi) + \frac{\rho v_r v_\phi}{r} = \frac{1}{r^3}\frac{\partial}{\partial r}\left(r^3 \rho v_r v_\phi\right),$$

$$\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta \rho v_\theta v_\phi) + \frac{\rho v_\theta v_\phi \cot\theta}{r}$$

$$= \frac{1}{r\sin^2\theta}\frac{\partial}{\partial\theta}(\sin^2\theta \rho v_\theta v_\phi).$$

System (7) then assumes the form

$$\begin{cases} \dfrac{\partial\rho}{\partial t} + \dfrac{1}{r^2}\dfrac{\partial}{\partial r}(r^2 m_r) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\theta}(\sin\theta m_\theta) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\phi}m_\phi = 0,\\ \dfrac{\partial m_r}{\partial t} + \dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left[r^2(p+\rho v_r^2)\right] + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\theta}(\sin\theta\rho v_r v_\theta) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\phi}(\rho v_r v_\phi) = \rho g_r + \dfrac{2p}{r} + \dfrac{\rho}{r}\left(v_\theta^2+v_\phi^2\right),\\ \dfrac{\partial m_\theta}{\partial t} + \dfrac{1}{r^3}\dfrac{\partial}{\partial r}\left(r^3\rho v_r v_\theta\right) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\theta}\left[\sin\theta(p+\rho v_\theta^2)\right] + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\phi}(\rho v_\theta v_\phi) = \rho g_\theta + \dfrac{1}{r}\left(p+\rho v_\phi^2\right)\cot\theta,\\ \dfrac{\partial m_\phi}{\partial t} + \dfrac{1}{r^3}\dfrac{\partial}{\partial r}\left(r^3\rho v_r v_\phi\right) + \dfrac{1}{r\sin^2\theta}\dfrac{\partial}{\partial\theta}(\sin^2\theta\rho v_\theta v_\phi) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\phi}(p+\rho v_\phi^2) = \rho g_\phi,\\ \dfrac{\partial}{\partial t}(\rho S) + \dfrac{1}{r^2}\dfrac{\partial}{\partial r}(r^2 m_r S) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\theta}(\sin\theta m_\theta S) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial\phi}(m_\phi S) = 0. \end{cases} \tag{19}$$

We will replace the derivatives by finite-difference expressions of the form

$$\frac{\partial\mathbf{w}}{\partial t} = \frac{\mathbf{w}^{n+1}-\mathbf{w}^n}{\tau},$$

$$\frac{1}{r^2}\frac{\partial}{\partial r}(r^2\mathbf{F}) = 3\frac{r_{i+1/2}^2\mathbf{F}_{i+1/2} - r_{i-1/2}^2\mathbf{F}_{i-1/2}}{r_{i+1/2}^3 - r_{i-1/2}^3},$$

$$\frac{1}{r^3}\frac{\partial}{\partial r}\left(r^3\mathbf{F}\right) = 4\frac{r_{i+1/2}^3\mathbf{F}_{i+1/2} - r_{i-1/2}^3\mathbf{F}_{i-1/2}}{r_{i+1/2}^4 - r_{i-1/2}^4},$$

$$\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\mathbf{G}) = \frac{\sin\theta_{j+1/2}\mathbf{G}_{j+1/2} - \sin\theta_{j-1/2}\mathbf{G}_{j-1/2}}{r_i\left(\cos\theta_{j-1/2} - \cos\theta_{j+1/2}\right)},$$

$$\frac{1}{r\sin^2\theta}\frac{\partial}{\partial\theta}(\sin^2\theta\mathbf{G}) = \frac{\sin^2\theta_{j+1/2}\mathbf{G}_{j+1/2} - \sin^2\theta_{j-1/2}\mathbf{G}_{j-1/2}}{r_i\sin\theta_j\left(\cos\theta_{j-1/2} - \cos\theta_{j+1/2}\right)},$$

$$\frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}\mathbf{H} = \frac{\mathbf{H}_{k+1/2} - \mathbf{H}_{k-1/2}}{r_i\sin\theta_j(\phi_{k+1/2} - \phi_{k-1/2})},$$

where $\mathbf{w}^n$ is the density vector at the nth time step, τ is the time step, integer indices refer to cell centers, and half-integer indices refer to cell boundaries. Note that all quantities are actually numbered by three indices on a three-dimensional grid; however, to avoid makng the formulas unwieldy, we will omit repeated indices. The indices i, j, and k vary with r, θ, and ϕ, respectively. The term $2p/r$ on the right-hand side of the second equation of (19) cancels out the corresponding term on the left-hand side. To preserve this property in the finite-difference approximation, the following relationship must be satisfied when $p=$ const:

$$3\frac{r_{i+1/2}^2 - r_{i-1/2}^2}{r_{i+1/2}^3 - r_{i-1/2}^3}p = \frac{2p}{r}.$$

We then find that

$$r_s = \frac{2}{3}\frac{r_{i-1/2}^2 + r_{i-1/2}r_{i+1/2} + r_{i+1/2}^2}{r_{i-1/2} + r_{i+1/2}}$$

and use $2p/r_s$ instead of $2p/r$ in the finite-difference equation. Similarly, for the term $p\cot\theta/r$ on the right-hand side of the third equation of (19), we should set $r = r_i$ and

$$\cot\theta = \frac{\sin\theta_{j+1/2} - \sin\theta_{j-1/2}}{\cos\theta_{j-1/2} - \cos\theta_{j+1/2}}.$$

The free-fall acceleration is determined by the formula $\mathbf{g} = -\nabla\Phi_g$. We neglect perturbations of the gravitational potential; i.e., we specify Φ_g in tabulated form as the initial condition and assume it to be time-independent. The components of $\mathbf{g}$ can be calculated using the formulas

$$g_r = -\frac{\Phi_{g\,i+1/2} - \Phi_{g\,i-1/2}}{r_{i+1/2} - r_{i-1/2}},$$

$$g_\theta = -\frac{\Phi_{g\,j+1/2} - \Phi_{g\,j-1/2}}{r_i(\theta_{j+1/2} - \theta_{j-1/2})}.$$

In view of the axial symmetry, $\Phi_g = \Phi_g(r,\theta)$ and $g_\phi = 0$. All other quantities on the right-hand sides of (19) are specified at the cell centers. The time step τ is determined by the Courant condition and can be calculated at each, nth, step according to the formula

$$\tau = C_{Cour}\min_{i,j,k}\left\{\frac{r_{i+1/2} - r_{i-1/2}}{|v_r| + c}, \frac{r_i(\theta_{j+1/2} - \theta_{j-1/2})}{|v_\theta| + c}, \frac{r_i\sin\theta_j(\phi_{k+1/2} - \phi_{k-1/2})}{|v_\phi| + c}\right\},$$

where $C_{Cour} =$ const is the Courant number, $0 < C_{Cour} < 1$, and c is the sound speed. The minimum is taken over the entire computational domain, but τ is determined by cells with small r due to the properties of the spherical grid.

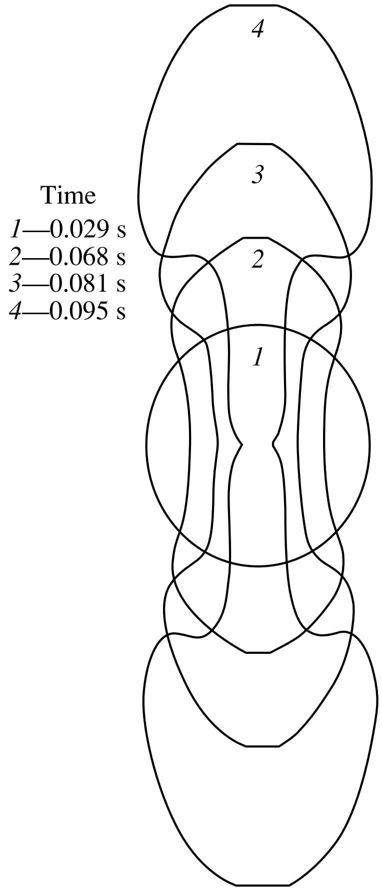

Fig. 1. Changes in the shape of the iron core during the development of large-scale convection.

7. RESULTS

In our computations, we made all quantities dimensionless using the following characteristic parameters:

(1) radius $R_0 = 1 \times 10^8$ cm,

(2) density $\rho_0 = 2 \times 10^9$ g/cm^3,

(3) pressure $p_0 = 1.2132 \times 10^{27}$ dyn/cm^2,

(4) entropy $S_0 = k_b/\mu = 8.3141 \times 10^7$ cm^2/s^2 K, where k_b is Boltzmann's constant and μ is the atomic mass unit,

(5) free-fall acceleration $g_{r0} = p_0/R_0\rho_0 = 6.0663\times 10^9$ cm/s^2,

(6) velocity $v_0 = \sqrt{p_0/R_0} = 3.4832 \times 10^9$ cm/s, and

(7) time $t_0 = R_0/v_0 = 0.0287$ s.

The dimensionless radius of the computational domain is 1.5. However, the initial equilibrium configuration was calculated for a radius of 1.8. In this way, we take into account the gravitation of layers of matter located outside the computational domain. The radius of the iron core formed by the time of the onset of the convective instability was assumed to be 0.25. The background entropy (the entropy of the rotating CO sphere) is $S_0 = 0.3564$, which corresponds to a temperature at the center of the star of $T_0 = 1 \times 10^8$ K and a density of $\rho_0 = 2 \times 10^9$ g/cm^3. The angular rotational speed is $\Omega_0 = 2.0732$ s^{-1}. At $t = 0$, we pass to the equation of state for iron in the central region, $r \leq 0.25$, and increase the entropy by a factor of ten. The values $r_{core} = 0.25$ and $S_{core} = 10S_0$ are free parameters of the problem. They are determined by the deflagration process, which controls the speed of propagation of the front, and by the dynamics of the core pulsations. We do not consider the relationship between these processes here. In a future study, using a more consistent formulation of the problem, and taking into account the deflagration rate, these parameters will be automatically determined during the computations.

We carried out our computations on an $N_\phi \times N_\theta \times N_r : 40 \times 80 \times 40$ grid with a Courant number of $C_{Cour} = 0.8$ using historical boundary conditions. In the course of the evolution, iron is mixed with unburned carbon and oxygen. Accordingly, we calculated the pressure using the formula $p = (1 - \alpha)p_{\rm CO} + \alpha p_{\rm Fe}$; i.e., we assumed a superposition of two equations of state. Since the entropy depends on the amount of matter that has burned (the iron content of the mixture), the coefficient $\alpha = \alpha(S)$ is a function of the entropy. We used the linear approximation $\alpha = (S - S_0)/(S_{core} - S_0)$.

Figure 1 shows entropy contours at the boundary of the iron core. Since the calculations implement a locally adiabatic approximation, i.e., they assume conservation of the entropy at any point of the Lagrangian coordinate space, these entropy levels characterize the transfer of material. As the computations progress, the entropy contours illustrate the departure of the shape of the core from the initial, nearly spherically symmetric configuration up until the time when the core breaks up and the regions of outflow acquire a jetlike structure.

Figure 2 illustrates the development of the explosion shown by the entropy contours and the momentum field in the meridional plane. Two blobs consisting of a mixture of the deflagration products with the initial carbon and oxygen begin to rise along the rotational axis. We can see that, by 0.05–0.07 s, the core loses its ellipsoidal shape, and its inner part begins to resemble two oppositely directed jets. Note the formation of two toroidal eddies encircling the jets (Fig. 3). This flow structure is due to the mushroom instability of a jet penetrating into a medium. Another interesting feature is the formation of streams from the unburned layers of the core to the central region. These streams have an equatorial configuration. The

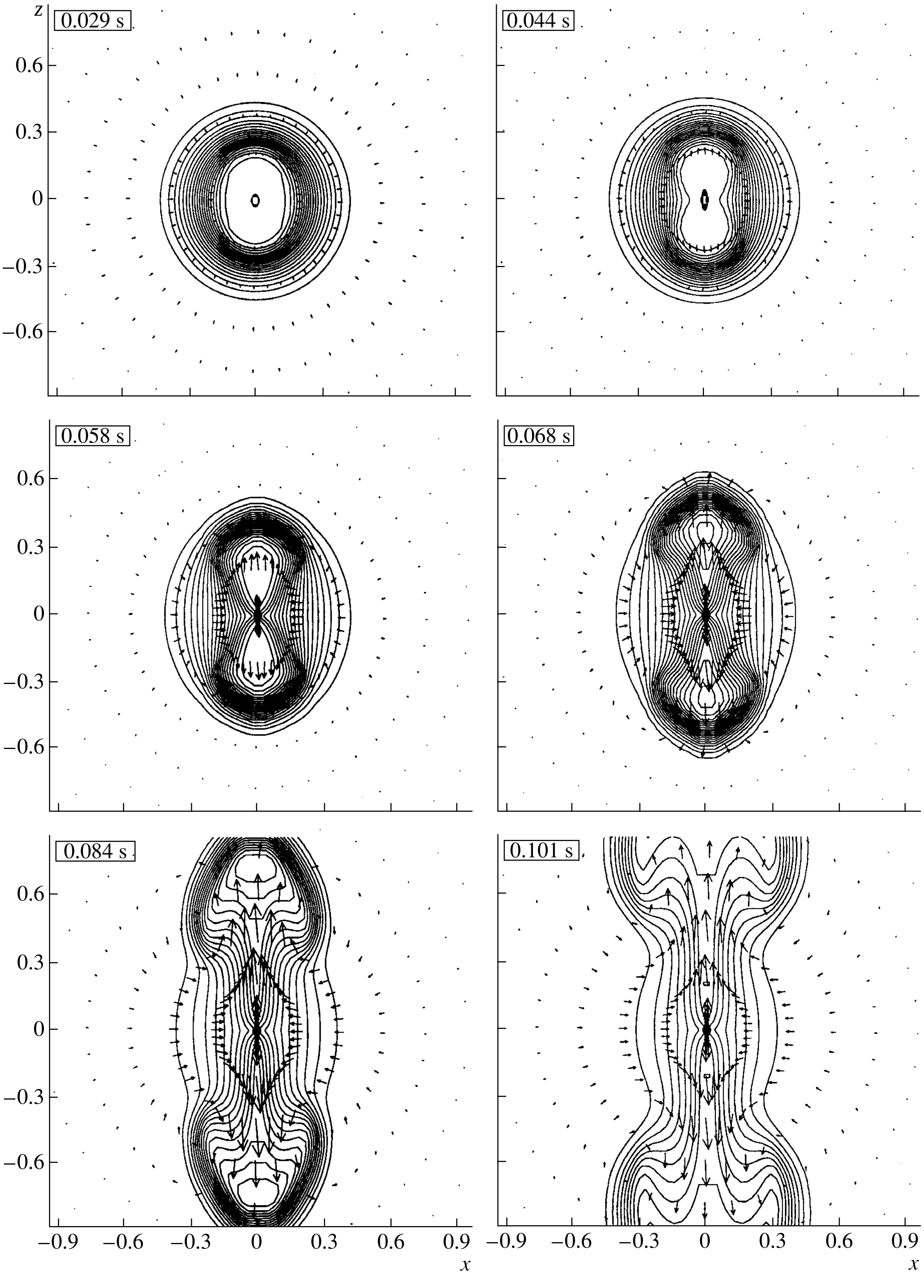

Fig. 2. Contours of entropy and the momentum field.

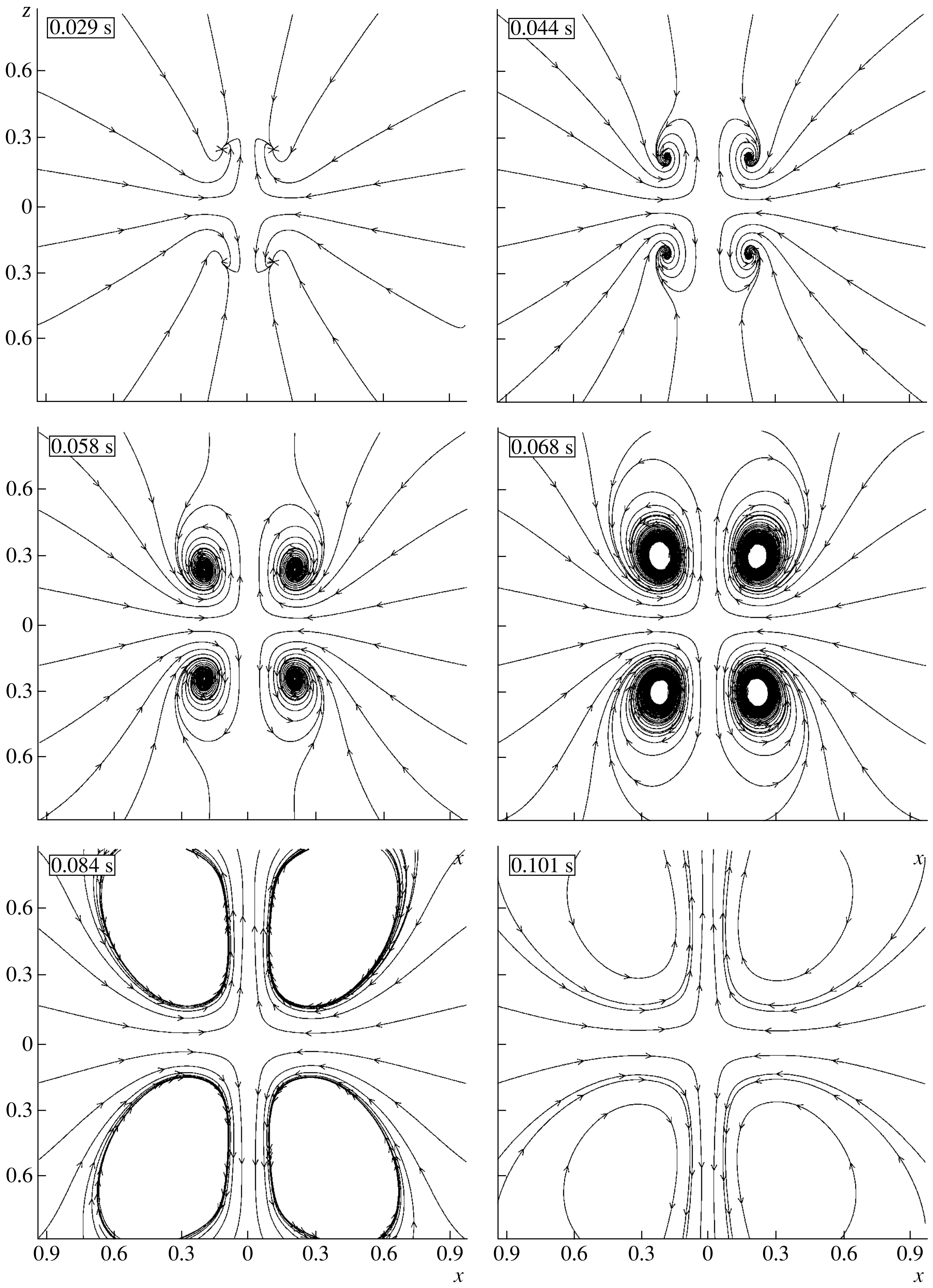

Fig. 3. Streamlines.

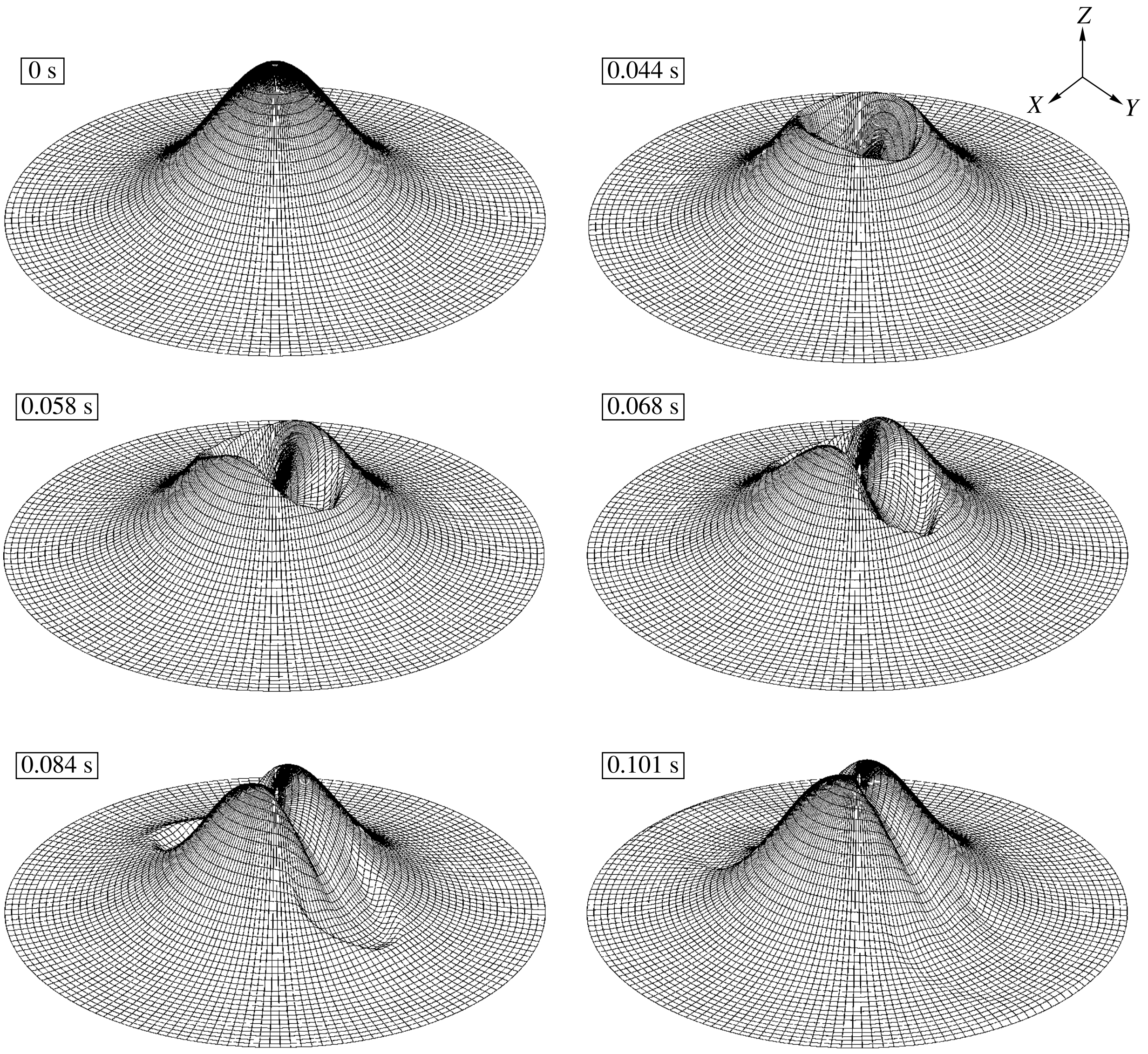

Fig. 4. Density distribution.

last two graphs in Fig. 2 ($t = 0.085$ and 0.105 s) exhibit only the jet pattern. During the explosion, virtually all the burned material emerges at the surface, with a carbon–oxygen mixture again appearing at the center.

The speed with which the blobs rise increases to a Mach number of about 1.5, which should result in the formation of a shock wave with an entropy jump behind the shock front. However, we cannot trace the formation of this shock, since we use hydrodynamical equations in a form that does not admit discontinuous solutions for the entropy (we approximate local adiabaticity). To resolve the shock-front structure in detail and study its influence on the explosion, the following system must be used in place of (6):

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0, \\ \dfrac{\partial \rho v_i}{\partial t} + \dfrac{\partial \Pi_{ik}}{\partial x_k} = \rho g_i, \\ \dfrac{\partial}{\partial t}\left(\dfrac{\rho v^2}{2} + \rho\varepsilon\right) + \operatorname{div}\left[\rho \mathbf{v}\left(\dfrac{v^2}{2} + \dfrac{p}{\rho} + \varepsilon\right)\right] = \rho \mathbf{v}\mathbf{g}, \end{cases} \tag{20}$$

where ε is the internal energy per unit mass of the substance. Switching the computations from (6) to (20) and back as necessary is also possible. This version of the procedure will be analyzed in a future study.

Figure 4 shows the evolution of the density distribution. The xy plane is the meridional plane, rotation occurs about the y axis, and density is plotted along the z axis. We can see that the density in the central region again grows during convection and the transport of material from the outer layers to the center. This is due to the fact that the unburned primary material has a lower temperature. Therefore, the conditions required for deflagration again arise in the central region, and a cyclic process forming a series of emerging large-scale structures is possible in the system.

During the formation of the iron core, its pressure increases by a factor of about 3.5. As a result, the primary disturbance is in the form of an acoustic wave, whose speed corresponds to a Mach number of no more than 0.5. The formation of this wave is predetermined by the initial configuration of the model. This disturbance does not substantially affect convective processes, since it leaves the computational domain long before convection begins. Furthermore, recall that the initial disturbance is spherically symmetric. The velocity field behind this disturbance becomes spherically symmetric, and the velocity is fairly low, so that it does not disrupt the core. The pressure jump rapidly becomes smoothed. This process could give rise to damped pulsations if there is no convective instability.

8. CONCLUSIONS

We have carried out numerical simulations of the development of a large-scale instability of the thermonuclear-deflagration front during the explosion of a type Ia supernova. Large-scale structures form in the rotating presupernova, which rise from the center to the outer layers of the star. This process is of paramount importance for understanding the explosion mechanism. The propagation of the deflagration front in type Ia supernovae is a strongly nonspherically symmetric process, and a large-scale front structure emerges, traveling most rapidly along the rotational axis. Fresh thermonuclear fuel continuously arrives at the center of the core, which should result in new flashes of burning. The fraction of burned CO fuel remains an open question. This quantity affects the peak height of the light curve of the supernova, and is commonly used to estimate cosmological distances. Another question concerns the production of chemical elements and the interpretation of light curves for an explosion lacking spherical symmetry. Khokhlov [17] carried out numerical simulations of the development of large-scale structures in a nonrotating supernova core. These resemble the structures that, according to [18], emerge during the development of large-scale convection.

Our results are indirectly supported by observations of SN 1987A, which is located in the Large Magellanic Cloud at a distance of 180 000 light years from the Earth. Although SN 1987A is a type II supernova, to all appearances, a similar explosion mechanism was realized there. Optical images of the SN 1987A explosion were obtained in 1995–2002 using the Hubble Space Telescope with violet, yellow, and red filters (see http://cfa-www.harvard.edu/cfa/oir/Research/sins.html). It is likely that there is no compact object at the center, so that the presupernova star exploded completely. During the destruction of this star, an elongated structure aligned with the rotational axis and normal to the plane of the ring is observed. This corresponds to the development of thermal instability in the degenerate CO core of the presupernova during its complete destruction. A Chandra X-ray image of SN 1987A obtained in January 2000 can be found at http://chandra.harvard.edu/photo/cycle1/sn1987a/). The results of our three-dimensional hydrodynamic calculations agree with this model for the explosion.

ACKNOWLEDGMENTS

This work was supported by INTAS (grant no. 01-491), the Russian Foundation for Basic Research (project no. 03-02-16548), and the Program “Leading Scientific Schools of Russia” (grant NSh-1029.2003.2).

Appendix

1. The divergence of a vector div**a** has the form

$$\operatorname{div}\mathbf{a} = \frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial x_1}(h_2h_3a_1) + \frac{\partial}{\partial x_2}(h_1h_3a_2) + \frac{\partial}{\partial x_3}(h_1h_2a_3)\right].$$

2. The divergence of a second-rank tensor div**T** has the form

$$(\operatorname{div}\mathbf{T})_{(1)} = \frac{1}{h_2h_3}\left[\frac{\partial}{\partial x_1}\left(\frac{h_2h_3}{h_1}T_{11}\right) + \frac{\partial}{\partial x_2}(h_3T_{12}) + \frac{\partial}{\partial x_3}(h_2T_{13})\right] - \frac{T_{11}}{h_1^2}\frac{\partial h_1}{\partial x_1} - \frac{T_{22}}{h_1h_2}\frac{\partial h_2}{\partial x_1} - \frac{T_{33}}{h_1h_3}\frac{\partial h_3}{\partial x_1} + \frac{(T_{12}+T_{21})}{h_1h_2}\frac{\partial h_1}{\partial x_2} + \frac{(T_{13}+T_{31})}{h_1h_3}\frac{\partial h_1}{\partial x_3},$$

$$(\operatorname{div}\mathbf{T})_{(2)} = \frac{1}{h_1h_3}\left[\frac{\partial}{\partial x_1}(h_3T_{21}) + \frac{\partial}{\partial x_2}\left(\frac{h_1h_3}{h_2}T_{22}\right) + \frac{\partial}{\partial x_3}(h_1T_{23})\right] - \frac{T_{11}}{h_1h_2}\frac{\partial h_1}{\partial x_2} - \frac{T_{22}}{h_2^2}\frac{\partial h_2}{\partial x_2} - \frac{T_{33}}{h_2h_3}\frac{\partial h_3}{\partial x_2}$$

$$+\frac{(T_{12}+T_{21})}{h_1h_2}\frac{\partial h_2}{\partial x_1}+\frac{(T_{23}+T_{32})}{h_2h_3}\frac{\partial h_2}{\partial x_3},$$

$$(\operatorname{div}\mathbf{T})_{(3)}=\frac{1}{h_1h_2}\left[\frac{\partial}{\partial x_1}(h_2T_{31})+\frac{\partial}{\partial x_2}(h_1T_{32})\right.$$

$$\left.+\frac{\partial}{\partial x_3}\left(\frac{h_1h_2}{h_3}T_{33}\right)\right]-\frac{T_{11}}{h_1h_3}\frac{\partial h_1}{\partial x_3}-\frac{T_{22}}{h_2h_3}\frac{\partial h_2}{\partial x_3}$$

$$-\frac{T_{33}}{h_3^2}\frac{\partial h_3}{\partial x_3}+\frac{(T_{13}+T_{31})}{h_1h_3}\frac{\partial h_3}{\partial x_1}+\frac{(T_{23}+T_{32})}{h_2h_3}\frac{\partial h_3}{\partial x_2}.$$

REFERENCES

1. K. Nomoto, K. Iwamoto, and N. Kishimoto, Science **276**, 1378 (1997).
2. W. Hillebrandt and J. C. Niemeyer, Annu. Rev. Astron. Astrophys. **38**, 191 (2000).
3. J. C. Niemeyer and S. E. Woosley, Astrophys. J. **475**, 740 (1997).
4. V. M. Chechetkin, S. S. Gershtein, V. S. Imshennik, *et al.*, Astrophys. Space Sci. **67**, 61 (1980).
5. L. N. Ivanova, V. S. Imshennik, and V. M. Chechetkin, Astron. Zh. **54**, 661 (1977) [Sov. Astron. **21**, 374 (1977)].
6. L. N. Ivanova, V. S. Imshennik, and V. M. Chechetkin, Astron. Zh. **54**, 1009 (1977) [Sov. Astron. **21**, 571 (1977)].
7. V. S. Imshennik, N. L. Kal'yanova, A. V. Koldoba, *et al.*, Pis'ma Astron. Zh. **25**, 250 (1999) [Astron. Lett. **25**, 206 (1999)].
8. V. S. Imshennik and D. K. Nadezhdin, Itogi Nauki Tekh. **21** (1982).
9. J.-L. Tassoul, *Theory of Rotating Stars* (Princeton Univ. Press, Princeton, 1979; Mir, Moscow, 1982).
10. D. K. Nadezhdin, Nauchn. Inform. Astron. Soveta Akad. Nauk SSSR **32**, 33 (1974).
11. I. Hachisu, Astrophys. J., Suppl. Ser. **61**, 479 (1986).
12. S. Chandrasekhar and R. N. Lebovitz, Astrophys. J. **138**, 185 (1963).
13. P. L. Roe and J. Pike, *Computing Methods in Applied Science and Engineering VI*, Ed. by R. Glowinski and J.-L. Lions (North-Holland, Amsterdam, 1984), p. 499.
14. P. Glaister, J. Comput. Phys. **74**, 382 (1988).
15. P. Glaister, J. Comput. Phys. **77**, 361 (1988).
16. Jian-Shun Shuen and Meng-Sing Liou, J. Comput. Phys. **90**, 371 (1990).
17. A. M. Khokhlov, *Three-Dimensional Modelling of the Deflagration Stage of a Type Ia Supernova Explosion*; astro-ph/0008463 v1 (2000).
18. V. M. Chechetkin, S. D. Ustyugov, A. A. Gorbunov, and V. I. Polezhaev, Pis'ma Astron. Zh. **23**, 34 (1997) [Astron. Lett. **23**, 30 (1997)].

Translated by A. Getling

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 934–940. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 1024–1031.
Original Russian Text Copyright © 2004 by Kasimova.

Influence of Binary Stars on the Chemical Evolution of Damped Lyα Systems

E. R. Kasimova

Rostov State University, Rostov-on-Don, Russia

Received January 18, 2004; in final form, May 27, 2004

Abstract—The impact of variations in the fraction of binary stars producing type Ia supernovae, β, on the chemical evolution of spiral galaxies is analyzed numerically. Even modest variations in β appreciably affect the evolution of the relative abundances of iron-group and alpha-process elements. If a substantial number of the damped Lα systems manifest in the spectra of quasars are due to spiral galaxies, the large scatter of the abundances of various elements displayed by these systems can be accounted for by variations in β.

1. INTRODUCTION

It is well known that a substantial fraction of stars are members of binary or multiple systems [1]. Estimates for the solar neighborhood show that nearly half of all G and M dwarfs are in binaries [2, 3]. According to available observations, depending on the type of star and the stellar population to which it belongs, the fraction of stars in binary systems varies from 30 to 60% [1–6].

In spite of the fact that nearly half of all stars are in binaries, as a rule, models for galactic evolution have taken into account only single stars in calculations of type Ia supernovae. Since it is thought that type Ia supernovae result from the thermonuclear explosion of accreting CO white dwarfs in intermediate-mass close binaries, of the entire range of possible binary masses, systems with total masses from 3 to 16 $M_{\odot}$ are explicitly included in the modeling. The fraction of binary systems giving rise to type Ia supernovae is a parameter in models for the chemical composition of galaxies, whose value is determined from observations of the rate of type Ia and type II supernovae.

The parameter β specifying the fraction of intermediate-mass binary stars whose evolution ends in type Ia supernovae in the standard scenario was first defined by Matteucci and Greggio [7], who derived the value $\beta = 0.1$ based on the observed supernova rate in the solar neighborhood. This value has been adopted in nearly all subsequent models for the chemical evolution of galaxies, although it is not clear that it is universal. We showed earlier [8] that, in the case of spiral galaxies, it is feasible to allow β to vary and determine its value based on agreement between the theoretical and observed [9] supernova rates. Depending on the type of spiral galaxy considered, good agreement with the observations of [9] was obtained for $\beta = 0.05$–0.1. Analysis for a more complex, multizone model for the chemical evolution of the Galaxy leads to similar values, $\beta = 0.05$–0.09 [10], but the specific value derived depends on the choice of the initial mass function.

Thus, the model for spiral galaxies studied in [8] yielded β values in the range 0.05–0.1 that were consistent with the observed supernova rates. We investigate here the applicability of this result to the chemical evolution of the absorbing systems producing saturated lines in quasar spectra, and whether the inferred variations of β can, to some extent, explain the observed dispersion in the elemental abundances in these systems. Section 2 gives a general description of the model and its parameters, Section 3 contains our results, and Section 4 presents a discussion of the problem studied, with our conclusions summarized in Section 5.

2. DESCRIPTION OF THE MODEL

Our model for the chemical and photometric evolution of spiral galaxies, including a self-consistent calculation of the dust component of the interstellar medium, was presented in [8]. The model is based on the standard chemical-evolution equations in the one-dimensional approximation of [7]:

$$dG(X,t)/dt = -\Psi(t)Z(X,t) \qquad (1)$$

$$+ \int_{m_{low}}^{M_b^{\min}} \Psi(t-\tau_m)Q_m(X,t-\tau_m)\varphi(m)dm$$

$$+ \beta \int_{M_b^{\min}}^{M_b^{\max}} \varphi(M_b)\left[\int_{\mu_{\min}}^{0.5} f(\mu)\Psi(t-\tau_{m_2})\right.$$

1063-7729/04/4811-0934$26.00 © 2004 MAIK "Nauka/Interperiodica"

$$\times Q_m(X,t-\tau_{m_2})d\mu\Bigg]dM_b + (1-\beta)$$

$$\times \int\limits_{M_b^{\min}}^{M_b^{\max}} \Psi(t-\tau_m)Q_m(X,t-\tau_m)\varphi(m)dm$$

$$+ \int\limits_{M_b^{\max}}^{m^{up}} \Psi(t-\tau_m)Q_m(X,t-\tau_m)\varphi(m)dm$$

$$+ Z_A(X)A(t) - Z_W(X)W(t),$$

where $\varphi(m)$ is the initial mass function (IMF) for stars with masses in the range (m_{low}, m_{up}), normalized to unity, $\Psi(t)$ is the star-formation rate in the galaxy, $G(X,t) = M(X,t)/M_t(t)$ is the total relative mass concentration of chemical element X in the gaseous and solid phases, $Z(X,t) = G(X,t)/G(t)$ the total mass concentration of element X in the gaseous and solid phases, $\tau(m)$ the lifetime of stars with mass m on the main sequence, $Z_A(X)$ is the mass concentration of element X in the accreting gas, and $Z_W(X)$ is the mass concentration of element X in the galactic wind. $Q_m(X,t-\tau_m) = (m - m_i)/m$ is the mass fraction of element X ejected by a star of mass m. It is known that supernovae are the main sources of heavy elements in the interstellar medium. The mass fraction $Q_m(X,t)$ of element X ejected by a star of mass m for type Ia supernovae determined for model W7 in [11] is in good agreement with the observed spectra of typical type Ia supernovae. The data presented in [12] are commonly accepted for the nucleosynthesis occurring during type II supernovae. In our calculations, we used the results of [11] and model B for the nucleosynthesis in type II supernovae from [12].

The system of equations (1) was solved together with equations describing the spectrophotometric evolution of galaxies and the evolution of dust in their interstellar media. More detailed descriptions of all notation used, the observational constraints on the model, and the star-formation scenario and IMF adopted are given in [8]. Like the generally accepted scenario [7], our model is concerned only with those binary systems that lead to the formation of type Ia supernovae. The contribution of these supernovae to the chemical evolution of galaxies is described by the third term on the right-hand side of (1). Let us consider this term in more detail [13]:

$$R_{\mathrm{Ia}}(t) = \beta \int\limits_{M_b^{\min}}^{M_b^{\max}} \varphi(M_b)/M_b \qquad (2)$$

$$\times \left[\int\limits_{\mu_{\min}}^{0.5} f(\mu)\Psi(t-\tau_{m_2})d\mu\right] dM_b.$$

Here, $M_b = M_1 + M_2$ is the total mass of the binary, M_1 is the mass of the initially more massive star, $\varphi(M_b)$ is the IMF for binary stars, which has the same form as the IMF for single stars, $\mu = M_2/(M_1 + M_2)$ is the relative mass of the secondary in the binary, and $f(\mu)$ is the distribution function of this quantity. According to [14], $f(\mu)$ is defined in the interval $0 < \mu \leq 0.5$ and has the form $f(\mu) = 2^{1+\gamma}(1+\gamma)\mu^\gamma$, where $\gamma = 2$. Since the Chandrasekhar limit is $M_{Ch} = 1.44\ M_\odot$ and stars more massive than 8 $M_\odot$ produce type II supernovae, only binary stars with total masses from $M_b^{\min} = 3\ M_\odot$ to $M_b^{\max} = 16\ M_\odot$ are considered as type Ia supernova precursors. All other stars in the galaxy are taken to evolve as single stars. In the system (1), the type II supernova rate is determined as

$$R_{\mathrm{II}}(t) = (1-\beta) \int\limits_{8M_\odot}^{M_b^{\max}} \Psi(t-\tau_m)\varphi(m)/mdm \qquad (3)$$

$$+ \int\limits_{M_b^{\max}}^{m^{up}} \Psi(t-\tau_m)\varphi(m)/mdm.$$

This means that all stars more massive than 8 $M_\odot$, both single and in binaries with total masses exceeding 8 $M_\odot$, produce type II supernovae at the end of their evolution.

Varying the parameter β in the interval 0.05–0.1 indicated by the observational constraints on models for spiral galaxies primarily influences the abundances of iron-group elements (Fe, Ni, Cr, Mn). Such variations could be one origin of the observed spread in the elemental abundances of galaxies. In the following section, we will consider the results of varying β in the models for spiral galaxies, and compare these with the observed elemental abundances for damped Lyα systems.

3. RESULTS OF THE COMPUTATIONS

Damped Lyα systems (DLA systems) are objects with neutral-hydrogen column densities $N(\mathrm{HI}) > 2 \times 10^{20}$ cm^{-2} that produce absorption lines in the observed spectra of quasars. As a rule, emission lines due to these objects are not observed. A few that have been identified at small redshifts confirm that DLA systems are associated with galaxies of various morphological types (spirals, dwarfs, low-surface-brightness galaxies) united by the presence in them of appreciable quantities of gas. We showed in [8] that

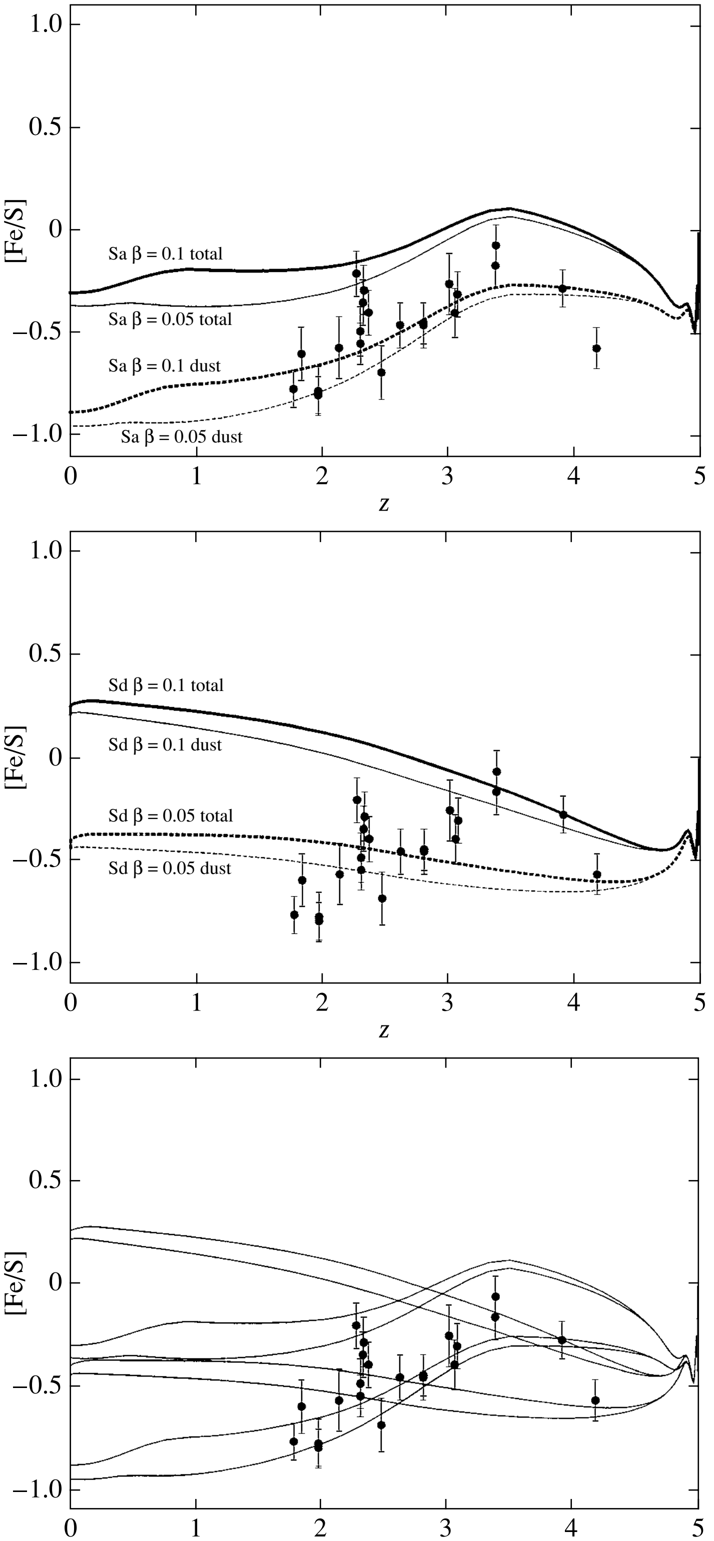

Fig. 1. Evolution of the relative content [Fe/S] with redshift z for models of (a) Sa and (b) Sd galaxies. The curves show the evolution of [Fe/S] for $\beta = 0.1$ for the total Fe and S abundances ("total," solid), and for the Fe and S abundances in the gaseous phase taking into account the condensation of some Fe and S atoms into the solid phase, forming dust ("dust," dotted). The thin curves show the same dependences for $\beta = 0.05$. The bottom panel shows the sums of all the curves presented in the upper and middle panels.

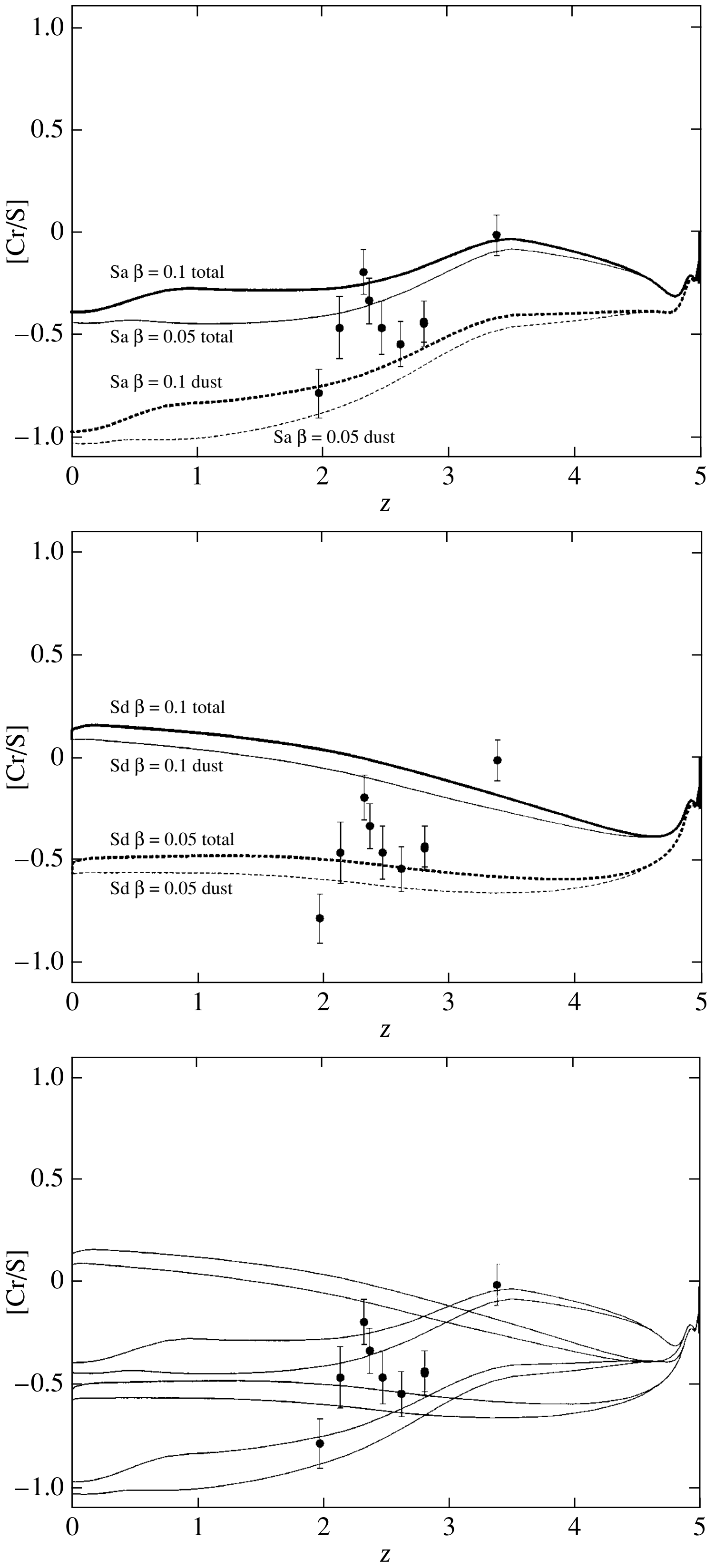

Fig. 2. Same as Fig. 1 for the ratio [Cr/S].

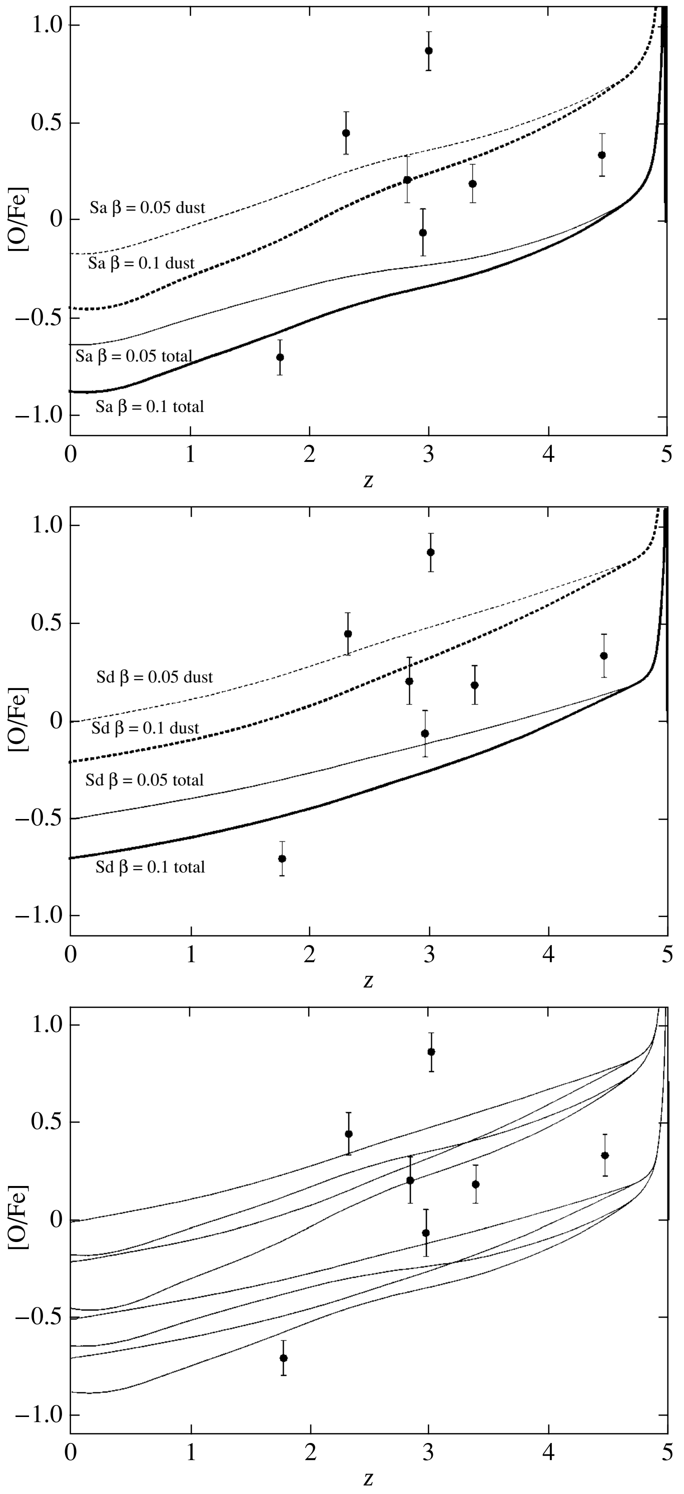

Fig. 3. Same as Fig. 1 for the ratio [O/Fe].

the chemical compositions of Lyα systems can be reproduced in model computations of the evolution of Sa to Sd spiral galaxies in the early stages of their evolution. Based on these results, we consider here the consequences of varying β in the interval 0.05–0.1 for DLA systems that are associated with spiral galaxies.

Figure 1 presents the evolution of the abundance ratio [Fe/S] as a function of redshift z. The points correspond to the observed abundances of these elements in DLA systems, while the various curves show the results of our numerical modeling. In particular, Fig. 1a shows, for the adopted evolutionary scenario for Sa galaxies, the evolution of [Fe/S] (solid bold curve) and the evolution of this abundance ratio for these elements in the gaseous phase, taking into account the condensation of Fe and S atoms into the solid phase (dust; dotted bold curve). Both scenarios were computed for $\beta = 0.1$. The thin curves show the corresponding results for $\beta = 0.05$. Figure 1b shows analogous curves for the evolution of [Fe/S] for the models of Sd galaxies, while Fig. 1c sums the information presented in the previous two panels for the models of Sa and Sd galaxies for $\beta = 0.1$ and $\beta = 0.05$, with and without a dust component.

Figures 2 and 3 show the evolution of [Cr/S] and [O/Fe] as functions of redshift z. The evolution of the relative abundances [Cr/S] and [O/Fe] for Sa galaxies is presented in Figs. 2a and 3a, and for Sd galaxies in Figs. 2b and 3b. As in Fig. 1c, Figs. 2c and 3c show summed computational results. The results for all intermediate models for galaxies between Sa and Sd, which are not shown here, can also satisfy the observations.

It is obvious that such regularities can be shown by the relative abundances of any iron-group element and any alpha-process element. They are also preserved when the evolution of [Fe/S], [Cr/S], or [O/Fe], for example, is considered as a function of the absolute elemental abundances. In this case, it is likewise possible to obtain a good agreement with the observations for DLA systems.

Since β is determined by the fraction of binary systems whose evolution leads to type Ia supernovae, which, in turn, determines the abundances of iron-group elements (Fe, Ni, Cr, Mn) in the corresponding galaxies, varying β within admissible limits is equivalent to varying the abundances of iron-group elements. This is why in Fig. 1 the relative abundance of Fe to S, which is a typical alpha-process element (together with O, Si, Ca), is sensitive to variations in β; naturally, Figs. 2 and 3 show this same behavior.

4. DISCUSSION

In accordance with the scenario of [7, 11], we specified the parameter β, determined by the fraction of binaries evolving into type Ia supernovae, to be constant. We then considered the interval of admissible values that were consistent with the observed supernova rates in spiral galaxies, and considered the consequences to the chemical evolution of spiral galaxies produced by variations within this interval.

Note that attempts to modify the classical scenario of [7, 11] and consider different models for the production of type Ia supernovae and their influence on the chemical evolution of galaxies were undertaken recently in [15]. These results showed that, as before, among various scenarios considered, that of [7, 11] can describe most adequately the chemical evolution of galaxies of early morphological types and the observed supernova rate.

It is known that the rate of formation of binaries is proportional to the metallicity of the interstellar medium [16, 17]. In general, we expect that β is a function of both the metallicity of the interstellar medium and of other properties determining the evolution of galaxies:

$$\beta = \beta[Z(t, \Psi(t), \phi(m), Q_m(m, X, t)]. \qquad (4)$$

We can see in (2) and (3) that this last assertion is also true for the rates of type Ia and type II supernovae, R_{Ia} and R_{II}. Thus, variations in β are physically fully justified even by the simple example considered in Section 3: we can see that even small variations in this parameter lead to appreciable changes in the relative abundances of iron-group and alpha-process elements.

It was already shown in the earliest studies that explicitly took into account the evolution of massive binaries [18, 19] that it was important to include the influence of binary systems on the photometric evolution of galaxies in computations of the star-formation rate. Due to the mass exchange between the components in massive, close binary systems, a population of O and WR stars is continuously generated even five million years after a burst of star formation; i.e., on time scales which are impossible to compute in evolutionary models that include only single stars [20]. Due to the presence of these O and WR stars in binaries, a star-forming region will appear younger than its actual age, and so evolutionary models including only single stars artificially underestimate the ages of active star-forming regions [20]. This conclusion was confirmed in other studies [21, 22] and further developed in [23, 24], where it was also confirmed that the situation is similar for the spectral characteristics of star-forming regions in the ultraviolet.

Another important consequence of taking into account binary systems in models of galactic evolution is the possibility of computing the rates of not only type Ia and type II, but also type Ib/c supernovae. Type Ib and Ic supernovae are produced either during the evolution of massive single stars that have lost their outer hydrogen layer due to their strong stellar winds or in massive close binary systems. Most modern models for the chemical evolution of galaxies consider only type II and type Ia supernovae. An exception is the two recent studies [16, 17], which were the first to consider in detail the evolution of binary stars of all masses and consequently producing all types of supernovae; it was shown that the relative rate of supernovae in the solar neighborhood is appreciably influenced by the evolution of close binary systems. It was also demonstrated in [16, 17] that the evolution of the abundances of several elements was sensitive to the fraction of the stars in binary systems. However, these results are not unambiguous. They differ quantitatively from standard models by no more than a factor of two to three. Since theoretical and observational uncertainties in models for galactic evolution are of the same order, it is currently difficult to interpret these results. Attempts to correctly describe and take into account all binary systems, on the one hand, provide more realistic models for stellar systems, but, on the other hand, appreciably complicate the models and increase the already comparatively large number of free parameters.

5. CONCLUSION

We have studied here the sensitivity of models for the chemical evolution of spiral galaxies to the paramter β, determined by the fraction of binary systems producing type Ia supernovae during their evolution. If damped Lyα systems are associated with spiral galaxies, the observed spread in elemental abundances for such systems could be due to differences in the morphological types of the galaxies, the selective condensation of heavy elements into dust, and variations in the parameter β. We have shown that varying β within admissible limits can produce a scatter in the abundances of heavy elements in DLA systems that is comparable to the scatter produced by selective condensation from the gaseous into the solid phase.

ACKNOWLEDGMENTS

The author thanks Yu.A. Shchekinov for useful comments and assistance with this study. This work was supported by the Russian Foundation for Basic Research (project no. 00-02-17689).

REFERENCES

1. N. Reid, Astron. J. **102**, 1428 (1991).
2. A. Duquennoy and M. Mayor, Astron. Astrophys. **248**, 485 (1991).
3. D. A. Fischer and G. W. Marcy, Astrophys. J. **396**, 178 (1992).
4. H. Zinnecker, astro-ph/0301078.
5. G. Duchene, T. Simon, J. Eislöffel, and J. Bouvier, Astron. Astrophys. **379**, 147 (2001).
6. B. D. Mason, D. R. Gies, W. I. Hartkopf, *et al.*, Astron. J. **115**, 821 (1998).
7. F. Matteucci and L. Greggio, Astron. Astrophys. **154**, 279 (1986).
8. E. R. Kasimova and Yu. A. Shchekinov, Astron. Zh. **81**, 387 (2004) [Astron. Rep. **48**, 353 (2004)].
9. E. Capellaro, R. Evans, and M. Turatto, Astron. Astrophys. **351**, 459 (1999).
10. F. Matteucci, astro-ph/0203340 (2002); *XII Canary Islands Winter School of Astrophysics: Cosmochemistry: the Melting Pot of Elements* (Tenerife, Spain, 2001).
11. K. Iwamoto, F. Brachwitz, K. Nomoto, *et al.*, Astrophys. J., Suppl. Ser. **125**, 439 (1999).
12. S. E. Woosley and T. A. Weaver, Astrophys. J., Suppl. Ser. **101**, 181 (1995).
13. L. Greggio and A. Renzini, Astron. Astrophys. **118**, 217 (1983).
14. A. V. Tutukov and L. R. Yungelson, *Close Binary Stars*: *Observations and Interpretation*, Ed. by M. J. Plavec, D. M. Popper, and R. K. Ulrich (Reidel, Dordrecht, 1980), p. 15.
15. F. Matteucci and S. Recchi, Astrophys. J. **558**, 351 (2001).
16. E. de Donder and D. Vanbeveren, New Astron. **7**, 55 (2002).
17. E. de Donder and D. Vanbeveren, New Astron. **8**, 817 (2003).
18. D. Vanbeveren, J. van Bever, and E. de Donder, Astron. Astrophys. **317**, 487 (1997).
19. J. van Bever and D. Vanbeveren, Astron. Astrophys. **334**, 21 (1998).
20. C. Leitherer, D. Schaerer, J. D. Goldader, *et al.*, Astrophys. J., Suppl. Ser. **123**, 3 (1999).
21. M. Cervino, J. M. Mas-Hesse, and D. Kunth, Rev. Mex. Astron. Astrofis. **6**, 188 (1997).
22. J. M. Mas-Hesse and M. Cervino, *IAU Symp. No. 193: Wolf-Rayet Phenomena in Massive Stars and Starburst Galaxies*, Ed. by K. A. van der Hucht, G. Koenigsberger, and P. R. J. Eenens (ASP, San Francisco, 1999), p. 550; astro-ph/9901350.
23. J. Van Bever and D. Vanbeveren, Astron. Astrophys. **400**, 63 (2003).
24. H. Belkus, J. van Bever, D. Vanbeveren, and W. van Rensbergen, Astron. Astrophys. **400**, 429 (2003).

Translated by D. Gabuzda

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 941–948. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 1032–1039.
Original Russian Text Copyright © 2004 by Titov.

Construction of a Celestial Coordinate Reference Frame from VLBI Data

O. A. Titov

Sobolev Astronomical Scientific Research Institute,
St. Petersburg State University, St. Petersburg, Russia

Received December 2, 2002; in final form, March 15, 2004

Abstract—A large number (∼2 million) of VLBI observations have been reduced in order to refine the measured coordinates of the observed radio sources. The data reduction was carried out in the OCCAM package using the least squares colocation method. Corrections to the coordinates of 642 objects were derived. The accuracy of the catalog is no worse than 0.2 milliseconds of arc for stable sources. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

Very Long Baseline Interferometry (VLBI) measures the difference in the arrival times of a wavefront emitted by an extragalactic radio source at two different radio telescopes. Such radio sources form the most stable realization of an inertial reference frame that is achievable at the current time. Because of their large distances from the Earth, their transverse shifts do not exceed 1 microarcsecond/year (μas/yr) on the sky, which is within the accuracy of VLBI measurements. Usually, from 3 to 20 radio telescopes separated by up to several thousand kilometers participate in a single 24-hour geodetic VLBI session. This network of telescopes carries out observations of a sample of quasars during a 24-hour session (from 10–15 sources at the beginning of the 1980s to 100 in modern observations). Scans of duration 1–3 min are simultaneously recorded for each source onto magnetic tapes at several stations. Observations are carried out at two frequencies: S band (2.3 GHz) and X band (8.4 GHz). This is done to enable correction for the frequency-dependent velocity of propagation of the wave in the ionosphere; only the X-band data are used directly for the analysis. When the observing session is finished, all the tapes are transported to a correlation center, where the data are correlated and the time delays to be used in the analysis are calculated.

During a 24-hour session, each radio telescope observes radio sources located in all directions, in order to exclude geometric correlations. Each radio source is observed many (sometimes more than 100) times with various combinations of antennas. In subsequent sessions, other groups of radio sources are observed using other VLBI networks. Thus, the large number of observations on intersecting quasar samples that has been accumulated makes it possible to derive the angular distances (arcs) between the sources on the sky, which can then be used to make a transformation to the usual coordinates of right ascension and declination, once a coordinate origin is chosen.

At the basis of constructing modern inertial reference frames lies the theoretical concept of the International Celestial Reference System (ICRS). The practical realization of this system is the International Celestial Reference Frame (ICRF), which is a catalog of objects whose coordinates have been referenced to a particular epoch. In accordance with a resolution of the IAU General Assembly in Kyoto in 1997, beginning in 1998 the ICRF has been defined using the coordinates of quasars derived from VLBI observations. This frame is now specified using 212 objects observed from 1979–1995, whose coordinates have been referenced to epoch J2000.0. The quasars that have been included in this list have a long observational history. The mean positional error for these quasars is ∼0.25 mas [1]. In addition, the ICRF catalog includes 294 so-called "candidate" objects, for which there are not yet a sufficient number of observations, as well as 102 unstable sources that are included to fill otherwide empty fields. Thus, the complete ICRF catalog consists of 608 quasars uniformly distributed over the celestial sphere [1].

Many sources have variable structure, leading to apparent shifts in the center of gravity of the radio brightness during short time intervals. Although these shifts can be much larger than the measurement accuracy, the quasar-based coordinate system is roughly a factor of 100 more accurate than the most accurate optical catalog—the FK5 catalog [2].

The ICRF catalog was obtained in 1996 by reducing ∼1.6 million observations carried out in 1979–1995 using the CALC/SOLVE program package developed at Goddard Space Flight Center (GSFC) [1]. All the sources are checked for their stability during the solutions. If the coordinates of some source display an apparent linear trend or random fluctuations, the source coordinates are considered to be variable, and are estimated separately for each session.

Work to improve the ICRF is ongoing. The ICRF-Ext. 1 supplement to the ICRF catalog came out in 2000; this contains 59 new quasars whose coordinates have been derived from ∼600 thousand observations in 1995–1999 [3]. However, the need to reexamine the main list of 212 ICRF quasars has arisen in recent years.

The construction of a catalog of radio sources from VLBI observations differs considerably from the construction of a fundamental catalog based on optical observations. For example, the optical FK5 catalog is comprised of a combination of individual catalogs containing objects in various declination zones, with each individual catalog being acquired using some particular instrument [2]. The ICRF radio catalog is constructed as a result of reducing all available observations with all radio telescopes; i.e., it is based on a single global solution. The advantage of this approach is that there is no need to combine various individual catalogs with their individual errors. The disadvantage is the impossibility of analyzing possible systematic errors. The use of only one software package in the construction of the ICRF can also be considered a disadvantage, since the derived coordinates could be subject to the influence of systematic errors associated with the reduction method realized in the SOLVE program. It is therefore important to have catalogs obtained using independent programs applying different estimation methods. We present here the results of reducing VLBI observations using the OCCAM package and the least squares colocation method.

2. OBSERVATIONAL MATERIAL AND REDUCTION METHOD

Nearly two million VLBI observations obtained in 1983–2001 were processed to construct the catalog. The observations were divided into two samples: those carried out on the IRIS-A and NEOS-A networks in 1983–2001 and those carried out on all other networks in 1988–2001. The former sample is of special interest, since precisely these observations were used by the International Earth Rotation Service (IERS). Only VLBI stations located north of the equator took part in the IRIS-A and NEOS-A programs, which limited the list of observed quasars to those with declinations $\delta = -45°$. Quasars located within 45° of the South pole were observed by only a few antennas in the Southern hemisphere, so that the number of observations, and consequently the positional accuracy, for these quasars were lower. The total number of studied objects was 642, of which 330 radio sources fell into the first sample and 620 into the second sample.

The IERS 2000 standards were used for the reduction of the data in the OCCAM package [4]. In accordance with the recommendations of the IERS, the effects of free-core nutation (FCN) were not included in the MHB2000 nutation reduction model [5]. We used the mapping function of Niell [6] when reducing the wet component of the troposphere delay to the zenith value. All radio sources observed at zenith distances $z > 85°$ were automatically reweighted.

Each sample was processed in two regimes: with and without estimation of the gradients of the wet component of the tropospheric delay. This enabled us to estimate the influence of these gradients on systematic errors in the quasar coordinate estimates. The coordinates of the reference stations and the Earth-rotation parameters were refined for each 24-hour observing session, and no-net-translation (NNT) and no-net-rotation (NNR) conditions were imposed for all the observing stations in order to avoid degeneracy of the matrix of normal equations. An NNR condition is also applied to the coordinates of the radio sources when corrections to these coordinates are being estimated.

Note also that we included all the radio sources (not only the 212 from the main ICRF list) in our list of "global" sources; i.e., those whose coordinates were taken to be constant over the entire observing interval. This is the main difference of our solution from that obtained at GSFC.

The second important difference is that CALC/SOLVE uses a segmented least-squares method [7, 8], while the OCCAM package uses a least squares colocation method in a three-group parametric model [9]. All the estimated parameters are divided into three groups:

—"global" parameters that are constant over the entire observational interval (corrections to the quasar coordinates);

—"daily" parameters that are constant over 24 hours (corrections to the nutational angles, corrections to the station coordinates, daily estimates of the tropospheric parameters and differences in the phases and clock rates for a given series of observations);

—"stochastic" parameters, which are variable over 24 hours (intraday fluctuations of the tropospheric delay at the zenith and of the clock phase differences).

To find the estimates, we formed the three-group parametric model

$$\mathrm{A}x + By + Cz + w = h, \qquad (1)$$

Table 1. Statistics of the solutions

	Solution 1	Solution 2	General solution
Number of series	1066	1400	2466
Number of observations	741.751	1.206.319	1.948.070
Time interval	1983–2001	1988–2001	1983–2001
Number of radio sources	330	620	642
Number of stations	21	51	51

where x, y, z are vectors for the global, daily, and stochastic parameters, respectively. A, B, C are matrices of partial derivatives for the indicated parameters, w is a vector of the observational errors, and h is a vector of the O–C values for the observations. Matrices of the *a priori* covariations for the stochastic parameters Q_Z and the observational errors Q_W are also introduced.

In accordance with the principles of the least squares colocation method, in order to estimate some unknown parameters, we must minimize the functional

$$S = w^T Q_W^{-1} w + z^T Q_Z^{-1} z. \qquad (2)$$

After extensive manipulation, the expression for the estimate of the global parameters has the form

$$\hat{x} = (A^T R A)^{-1} A^T R h \qquad (3)$$

or, allowing for the independence of the local and stochastic parameters in different observational series,

$$\hat{x} = \left(\sum_i (A^T R A)_i \right)^{-1} \sum_i (A^T R h)_i, \qquad (4)$$

where the sum is taken over all series included in the analysis. The matrix R in (2) and (3) is calculated using the formula

$$R = Q^{-1} - Q^{-1} B (B^T Q^{-1} B)^{-1} B^T Q^{-1}, \qquad (5)$$

and the matrix Q, in turn, is a combination of the matrices

$$Q = C Q_Z C^T + Q_W. \qquad (6)$$

Further, we can estimate the vectors y and z:

$$\hat{y} = (B^T Q^{-1} B)^{-1} B^T Q^{-1} (h - A\hat{x}), \qquad (7)$$

$$\hat{z} = Q_Z C^T (C Q_Z C^T + Q_W)^{-1} (h - A\hat{x} - B\hat{y}). \qquad (8)$$

The least squares error in the estimates of the global parameters (4) is

$$\sigma_{\hat{X}} = (\chi^2 \mathrm{diag}(A^T R A)^{-1})^{1/2} \qquad (9)$$

$$= \left[\chi^2 \mathrm{diag} \left(\sum_i (A^T R A)_i \right)^{-1} \right]^{1/2},$$

where χ^2 is a normalized chi-squared calculated using the formula

$$\chi^2 = \frac{(h - A\hat{x} - B\hat{y})^T Q^{-1} (h - A\hat{x} - B\hat{y})}{N - N_y - n_x}, \qquad (10)$$

where N, n_x, and N_y are, respectively, the total number of observations in all series, the number of global parameters, and the number of daily parameters, likewise summed over all series, $N_y = \sum_i n_y$.

The *a priori* autocovariance functions for the differences in the clock rates and the wet component of the tropospheric delay at the zenith were taken from [10]. The possibility of using the least squares colocation method for the reduction of VLBI data was demonstrated earlier in [9].

3. DISCUSSION OF THE RESULTS

Our reduction of the data using formulas (4)–(10) yielded the coordinates of 642 radio sources. Table 1 summarizes information about our two solutions obtained for the various samples, as well as about the overall solution. A full catalog of the radio sources is published on the web site of the Astronomical Scientific Research Institute of St. Petersburg State University (http://astro.pu.ru/astro/win/researches/ivs.html).

Figures 1 and 2 illustrate the influence of the tropospheric gradients on the corrections $\Delta\alpha\cos\delta$ and $\Delta\delta$ for the IRIS-A/NEOS-A observations and for the remaining data, respectively, while Fig. 3 displays the same information for all the observations together. These plots indicate that including the tropospheric gradients has virtually no effect on $\Delta\alpha\cos\delta$, and influences only $\Delta\delta$. The maximum effect (about 0.3 mas) is observed near the equator. It is believed that this is due to the thickening of the atmosphere from the polar regions toward the equator. This means that the radio waves from sources observed near the southern horizon from the Northern hemisphere must

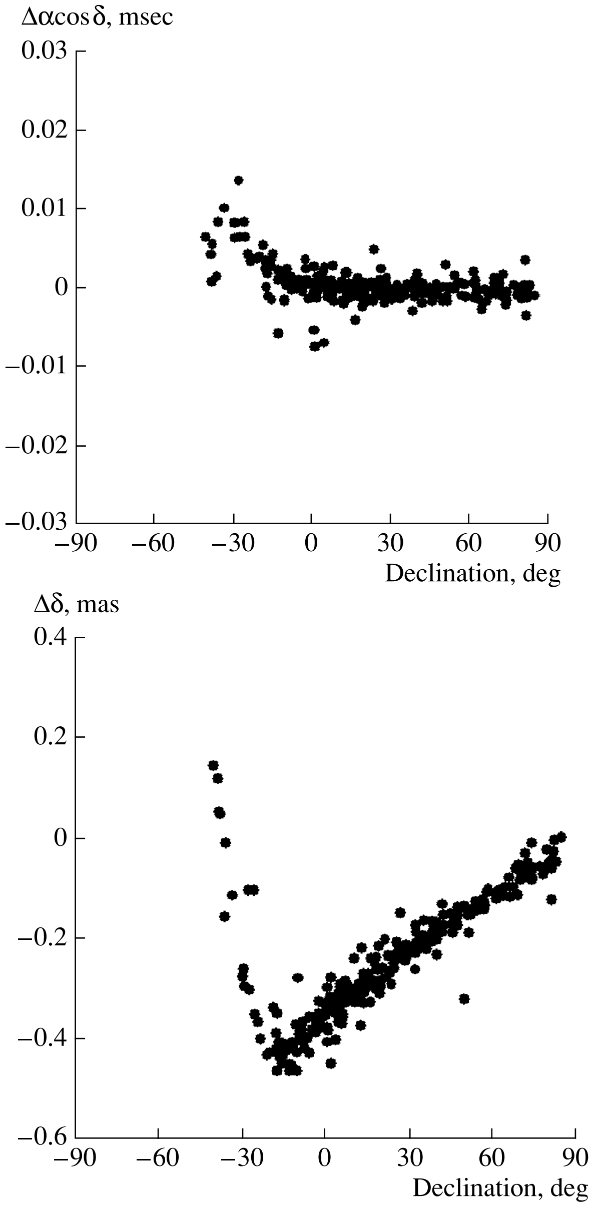

Fig. 1. Difference between the results with and without including the gradients of the tropospheric delay for the correction $\Delta\alpha\cos\delta$ (upper) and $\Delta\delta$ (lower) based on the observations on the IRIS-A/NEOS-A networks. Only radio sources with more than 20 observations are shown.

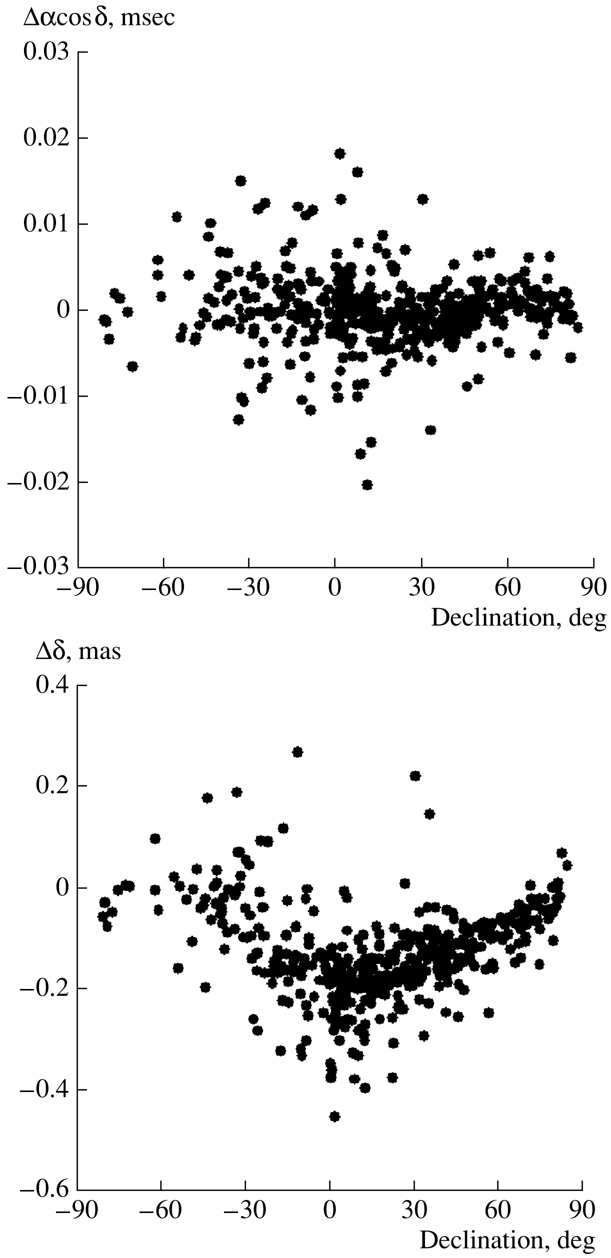

Fig. 2. Same as Fig. 1 but excluding the IRIS-A/NEOS-A data.

pass through a thicker layer of atmosphere than do radio waves for sources observed near the northern horizon with these same stations [11].

We can distinguish two groups of objects in Fig. 1: at declinations from −15° to 90° and from −45° to −15°. For objects in the first group, differences in the declination estimates are due to whether or not the tropospheric gradients are taken into account. For the second group, the determining role is probably played by imperfection of the mapping function at large zenith distances, since, as a rule, these objects were observed near the horizon. A similar, although weaker, effect is observed for the estimated right ascensions. We therefore conclude that analyses of the observations for the IRIS-A/NEOS-A programs (i.e., carried out from the Northern hemisphere) can yield accurate coordinates only for radio sources north of declination $\delta = -15°$.

The systematic shift between the two solutions is not as large in Fig. 2, but the random scatter of points about the mean curve is larger than in Fig. 1. This is due to the fact that the first solution is more

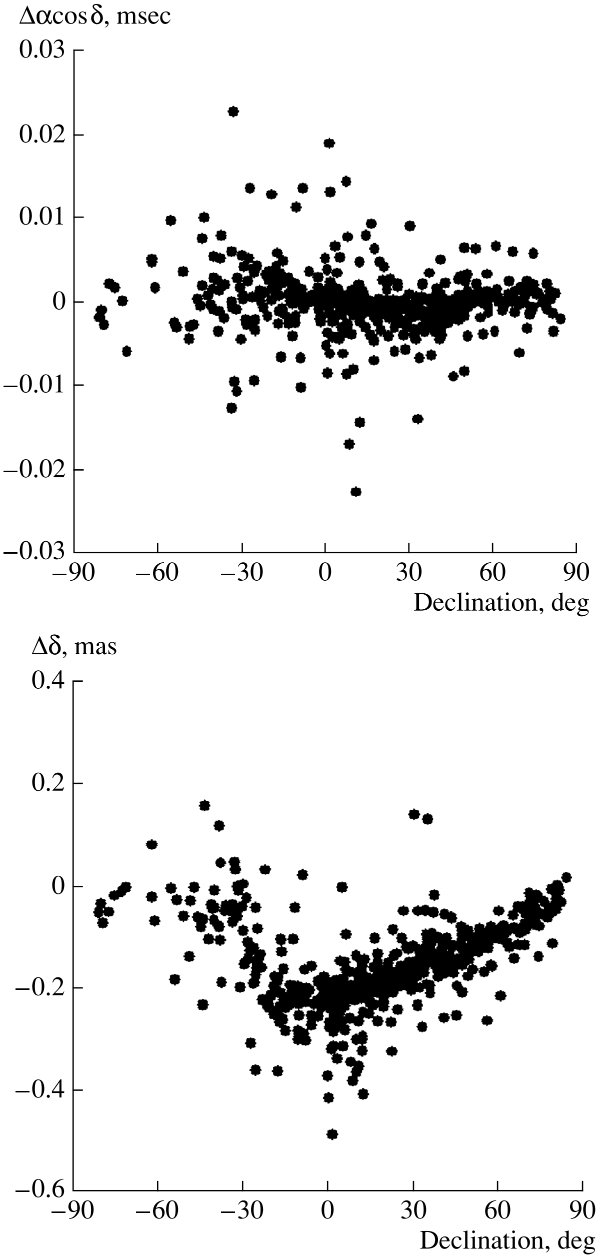

Fig. 3. Difference between the results with and without including the gradients of the tropospheric delay for the corrections $\Delta\alpha\cos\delta$ (upper) and $\Delta\delta$ (lower) based on all the observations. Only radio sources with more than 20 observations are shown.

uniform, being based on observations carried out on (as a rule, four to six) northern stations that were well distributed in latitude and longitude, which made it possible to observe the radio sources at a range of zenith angles. The second solution was obtained using various networks, including some for which the distribution of the stations is not as uniform. Therefore, the systematic effects in Fig. 1 are associated with the absence of VLBI stations in the Southern hemisphere, and in Fig. 2 with insufficient accuracy of the observations themselves.

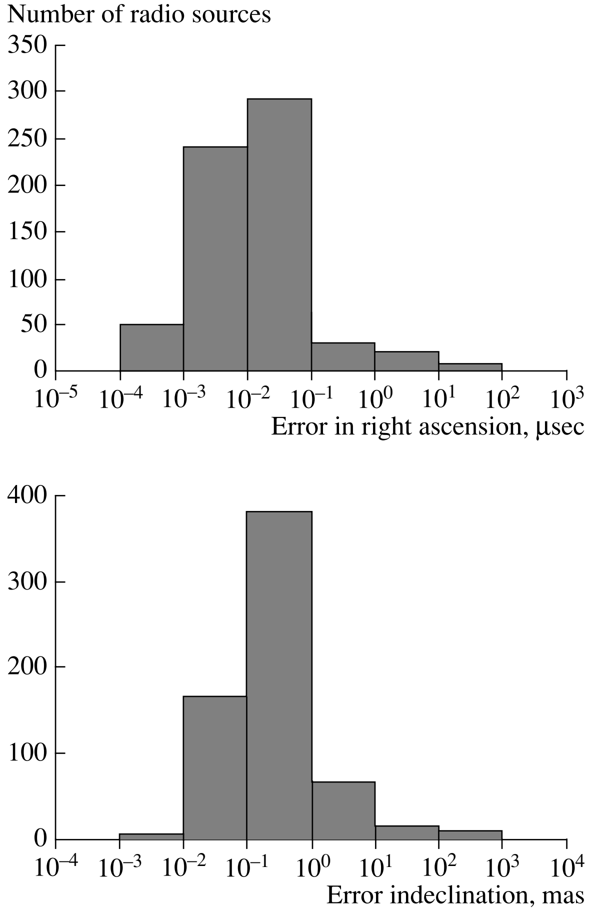

Fig. 4. Distribution of the formal errors in the coordinates $\Delta\alpha\cos\delta$ (upper) and $\Delta\delta$ (lower).

Figure 3 displays the same shifts obtained using all the observations together. Overall, Fig. 3 repeats the behavior shown in Fig. 2, although a small asymmetry is visible in the plot for $\Delta\delta$, due to the addition of observations from the Northern hemisphere.

Figure 4 depicts the distributions of the errors in each coordinate. These histograms show that the median position error is approximately 0.3 mas. To estimate the external accuracy of the catalog, we compared our errors with those for two other catalogs, obtained by the US Naval Observatory (USNO) [12] and the Leipzig Cartography Institute (BKG). Both of these catalogs were obtained using the CALC/SOLVE package. It has recently come to light that many sources in the main ICRF list have variable structure. Therefore, the calculations were carried out for the list of 199 quasars compiled by Feissel-Vernier [13] based on an analysis of time series of coordinate measurements for 1979–2002 at the USNO [12]. The quasars in this list passed several tests to verify the stability of their positions

Table 2. RMS deviation for the coordinate differences for three catalogs in mas (in cells above the empty diagonal) for $\Delta\alpha\cos\delta$ (upper values) and $\Delta\delta$ (lower values), together with the number of quasars included in the statistical analysis (in cells below the empty diagonal). Quasars in the list of Feissel-Vernier [13] are used

	BKG	USNO	SPSU
BKG		0.212	0.195
		0.127	0.225
USNO	181		0.209
			0.201
SPSU	195	180	

[13]. The formal positional errors for these quasars are ∼0.1 mas.

In the comparisons, we included only quasars with at least 20 observations for each of the catalogs. The results are presented in Table 2, which shows that the agreement between all three catalogs for the sample of stable quasars from [12] is 0.1–0.2 mas. The estimated accuracies for the BKG–USNO $\Delta\delta$ differences are nearly a factor of two better than for the SPSU–USNO and SPSU–BKG differences. This may be associated with the different methods used to correct for the tropospheric delay at the zenith in the OCCAM and CALC/SOLVE packages, as well as with the fact that, in contrast to CALC/SOLVE, the OCCAM package estimates all the quasar coordinates as global parameters.

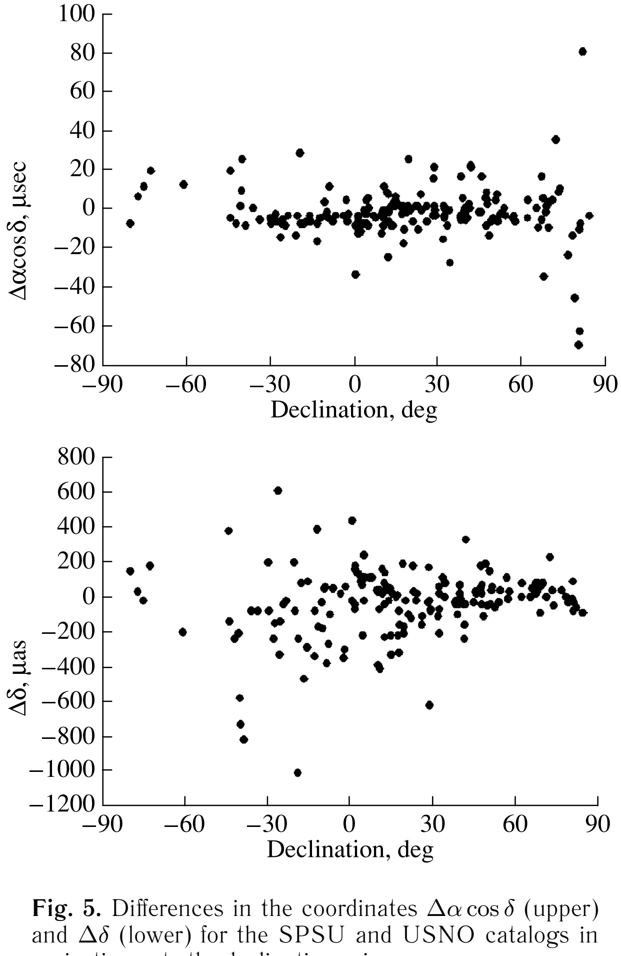

Fig. 5. Differences in the coordinates $\Delta\alpha\cos\delta$ (upper) and $\Delta\delta$ (lower) for the SPSU and USNO catalogs in projection onto the declination axis.

Figure 5 presents as an illustration the difference in the SPSU–USNO coordinate corrections $\Delta\alpha\cos\delta$ and $\Delta\delta$ projected onto the declination axis. It is interesting that, while a sharp degradation in the accuracy of $\Delta\alpha\cos\delta$ is observed in near-polar regions, the $\Delta\delta$ differences show a smooth increase in the scatter with distance further into the Southern hemisophere, presumably due to the smaller number of observations in this part of the sky. Similar effects are observed for the other combinations of catalogs.

Thus, the errors in the estimated coordinates for "stable" quasars are at the level of 0.2 mas, which is a factor of two higher than the formal errors for this group of objects (0.1 mas), confirming the conclusions of other studies [14, 15]. This is due to the influence of various systematic factors, which, for various reasons, have not been taken into account when deriving corrections based on VLBI observations.

One such factor is instability in the apparent positions of the quasars. In order to estimate the level of this instability, we carried out additional studies aimed at estimating the coordinates as daily parameters. As an example, Fig. 6 shows variations in the estimated corrections $\Delta\alpha\cos\delta$ for the quasar 0923+392 (4C 39.25) from 796 daily sessions, derived from the IRIS-A/NEOS-A observations for 1986–2001. We can see a trend on which are superposed systematic fluctuations. A spectral analysis enabled us to identify oscillations with a period of ∼2070 days and an amplitude of 72 ± 17 μas. A least-squares estimate of the linear trend (66.2 ± 2.6 μas/yr) is in agreement with the result of [13] for data obtained in 1986–1997 (59.8 ± 2.2 μas/yr). The presence of this trend in the derived corrections $\Delta\alpha\cos\delta$ for this quasar had been reported earlier in [16, 17]. Astrophysical studies of 0923+392 showed the presence of rapid apparent motions associated with the emergence of a superluminal component in 1980 [17]. A quadratic model was also used in [14], which yielded the quadratic term -13.6 μas/yr^2, suggesting deceleration of the motion of the superluminal component. The extrapolation carried out in [14] showed that the motion should have ceased in the middle of 1997. New estimates incorporating the results of observations obtained in 1998–2001 and taking into account the 2070-day period yielded the

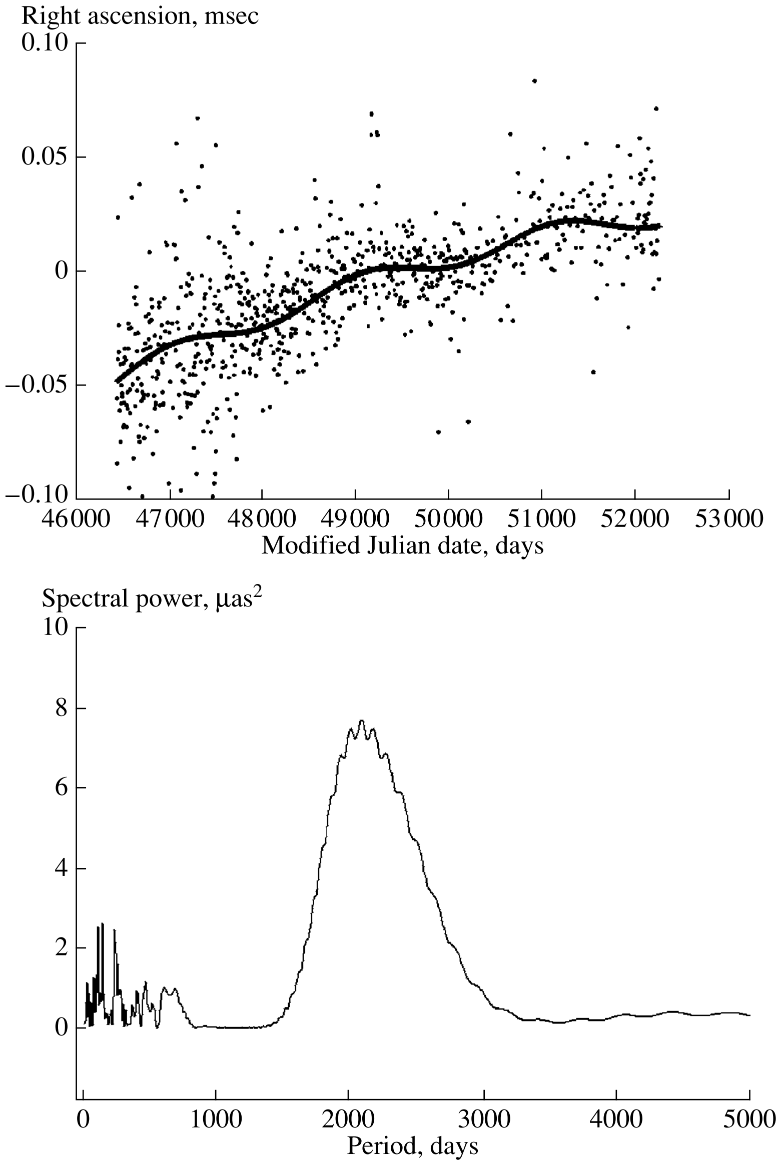

Fig. 6. Variations in the daily right-ascension corrections for the quasar 0923+392 (4C 39.25) in the interval 1986–2001, together with the spectral power for these variations. The solid curve in the upper panel shows an approximate quadratic model plus a signal with a period of 2070 days.

deceleration -2.4 ± 0.5 μas/yr^2. Thus, the deceleration of the superluminal component was not as rapid as originally believed, and, in fact, its motion should have ceased in the middle of 1999, as can be observed in Fig. 6. At present, it is difficult to suggest a likely origin for the periodic oscillations.

It appears that the specified scale for the linear shift (60–70 μas/yr) is fairly unique, but it is possible that linear trends due to instability of the radio structure will be a more common phenomenon at levels of <10 μas/yr. This complicates the creation of a high-accuracy celestial coordinate system. This problem can be solved only by comparing radio images, calculating the structural delays for each source individually, and using this information during the analysis of the VLBI observations. Detailed experiments of this kind have already been carried out for the quasar 3C273 [18]. However, the construction of maps for several hundreds of quasars would require appreciable computational resources.

4. CONCLUSIONS

Modern VLBI observations can be used to create a reference frame based on the coordinates of radio sources with accuracies of 0.2–0.4 mas. Further increase in the coordinate accuracy is limited by various factors. Most of the observed objects have variable structures, with the scale for the variations exceeding the formal accuracies obtained during the coordinate estimation. These effects are not predictable, making it difficult to model them.

Another source of errors is the troposphere, which introduces errors at a level of 0.1–0.2 mas, especially for observations at large zenith distances.

In addition, the effect of aberration due to the revolution of the solar system around the center of the Galaxy can lead to systematic shifts in the quasar coordinates at a level of 5 microseconds/yr [19]; this has led to errors of 0.1 mas over 20 years of observations.

Further improvement of the accuracy of the ICRF requires work in several directions: increasing the number of observations, increasing the number of radio sources, especially in the Southern hemisphere, the creation of more flexible observational programs, studies of the apparent motions of radio sources via mapping of the radio structure, and the construction of more accurate models for the motions of VLBI stations.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 01-02-17070) and the Ministry of Education of the Russian Federation (grant 11.36). The author thanks V.V. Vityazev (ASRI, SPSU) for a number of useful comments about this paper, and also C. Ma and D. MacMillan (GSFC), A. Fey (USNO), and M. Feissel-Vernier (Observatoire de Paris) for consultations about VLBI data reduction and useful discussions.

REFERENCES

1. C. Ma, E. Arias, T. Eubanks, *et al.*, Astron. J. **116**, 516 (1998).
2. W. Fricke, H. Schwan, and T. Lederle, Veroeffentl. Astron. Rechen-Inst. Heidelberg **28** (1988).
3. C. Ma, *Proceedings of the 15th Working Meeting on European VLBI for Geodesy and Astrometry*, Ed. by D. Behrend and A. Rius (Institut d'Estudis Espacials de Catalunya, Consejo Superior de Investigaciones Científicas, Barcelona, 2001), p. 187.
4. *IERS Conventions 2000*, Ed. by D. McCarthy (Paris Obs., Paris, 2000).
5. P. M. Mathews, T. A. Herring, and B. A. Buffett, J. Geophys. Res. **107**, 2068 (2002).
6. A. E. Niell, J. Geophys. Res. **101**, 3227 (1996).
7. C. Ma *et al.*, Astron. J. **92**, 1020 (1986).
8. C. Ma, J. Sauber, L. Bell, *et al.*, J. Geophys. Res. **95**, 21991 (1990).
9. O. Titov, in *IERS Technical Notes 28*, Ed. by B. Kolaczek, H. Schuh, and D. Gambis (Paris Obs., Paris, 2000), p. 11.
10. V. S. Gubanov and O. A. Titov, Commun. No. 60 IPA RAN (Inst. of Appl. Astron., Russ. Acad. Sci., 1994).
11. D. S. MacMillan and C. Ma, Geophys. Res. Lett. **24**, 453 (1997).
12. A. L. Fey, private communication (2002).
13. M. Feissel-Vernier, Astron. Astrophys. **403**, 105 (2003).
14. A. L. Fey, M. Eubanks, and K. A. Kingham, Astron. J. **114**, 2284 (1997).
15. A. L. Fey, IAU Coll. **180**, 20 (2000).
16. A. Alberdi, J. M. Marcaide, A. P. Marscher, *et al.*, Astrophys. J. **402**, 160 (1993).
17. A. Alberdi, J. L. Gomez, J. M. Marcaide, *et al.*, Astron. Astrophys. **361**, 529 (2000).
18. P. Charlot, in *Proceedings of the URSI/IAU Symposium*, Ed. by T. Sasao, S. Manabe, O. Kameya, and M. Inoue (Tokyo, Japan, 1993), p. 287.
19. O. Sovers and J. Fanselow, Rev. Mod. Phys. **70**, 1393 (1998).

Translated by D. Gabuzda

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 949–953. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 1040–1044.
Original Russian Text Copyright © 2004 by Gabov, Sokoloff.

Current Helicity and the Small-Scale Dynamo

A. S. Gabov[1] and D. D. Sokoloff[2]

[1]Research Computer Center, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia

[2]Moscow State University, Vorob'evy gory, Moscow, 119992 Russia

Received February 25, 2004; in final form, May 27, 2004

Abstract—The small-scale dynamo inherent to mirror-asymmetric turbulence can generate a magnetic field characterized by substantial mirror asymmetry of the associated electric currents. In general, the corresponding helicity should be taken into account in calculations of the helicity balance, which is now used as a basis for models describing the suppression of the large-scale dynamo. However, the mirror asymmetry of the fluctuating magnetic fields is concentrated on scales much shorter than the magnetic-loop diameter. Therefore, the unaccounted-for contribution to the helicity balance is, in fact, not important.

1. INTRODUCTION

The origin of the large-scale magnetic fields of celestial bodies via the dynamo mechanism is commonly attributed to the joint action of differential rotation and the helicity of turbulent (or convective) flows. As this mechanism operates, the originally weak magnetic field grows exponentially, although this growth will obviously be stabilized at a later time. In principle, various factors can be responsible for this stabilization. However, it seems likely for the solar and galactic dynamos that it results from the suppression of the weakest link in the chain of magnetic-field self-excitation: helicity.

Until recently, this idea was largely speculative, since all considerations of helicity and its role in magnetic-field generation were based exclusively on theoretical calculations (which can hardly be entirely realistic) and order-of-magnitude estimates. Recently, considerable progress has been achieved due to observations of one helicity component at the solar surface. These observational data can be interpreted in the framework of dynamo theory to some extent, and a nonlinear-stabilization scenario for the dynamo that reproduces at least the basic outline of the observed pattern can be derived (see [1] and references therein). For definiteness, we will consider here the solar dynamo, since it is precisely for the solar dynamo that an observational basis for helicity estimates was first found.

Far from all stages of this scenario have been developed theoretically, and many of the assumptions on which it is based await justification. We will substantiate one of these assumptions here.

The root of the matter is as follows. Helicity is a measure of the breaking of mirror symmetry in an MHD system. This includes contributions from the velocity field and magnetic field, which are called the hydrodynamic and the magnetic helicity, respectively. It is the magnetic helicity that can be obtained from observations, and the magnetic helicity is, in addition, a conserved quantity. In turn, the magnetic helicity is composed of parts produced by the mean and the fluctuating magnetic field. The mean magnetic field and its helicity grow due to the action of the solar dynamo; thus, to conserve helicity, we must assume that variations in the helicity of the small-scale magnetic field have the same magnitude but are opposite in sign. It is this helicity that appears in the coefficients of the mean-field-dynamo equations (the so-called α effect) and stabilizes the field-generation process. Relevant calculations are presented in [1, 2].

The gap that we are going to fill is as follows. We are interested in the behavior of the mean solar magnetic field and the fluctuating magnetic fields that originate together with the mean field and participate, for instance, in the solar-activity cycle. However, the generation of fluctuating magnetic fields is not necessarily associated with the mean field. Naturally, such fluctuating fields are not involved in the solar cycle. For the corresponding generation mechanism, known as the small-scale dynamo, the α effect does not play any role; starting from the pioneering work of Kazantsev [3], this has been studied for the case of nonhelical velocity fields. The magnetic field generated by this mechanism is likewise not helical, and this field should be taken into account when analyzing the suppression of the generation of the large-scale field.

However, solar convection is actually helical, and both theoretical and observational evidence for this is now available [1]. The α effect, which has the

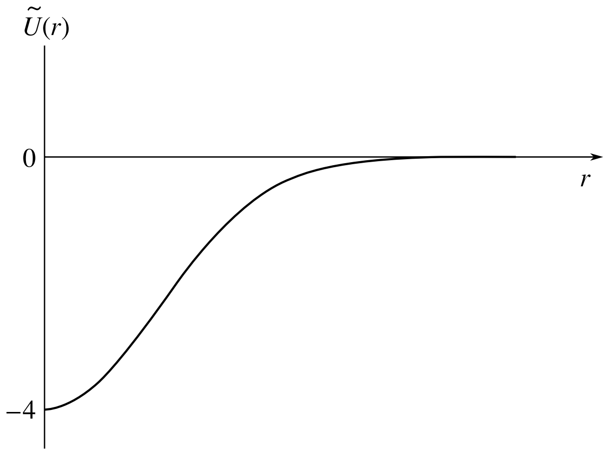

Fig. 1. Typical profile of the potential $\tilde{U}(r)$.

dimension of velocity, can reach 10% of the convective velocity (although it may be much smaller). Therefore, we cannot rule out the possibility that the helicity of the small-scale magnetic field, which is not coupled with the large-scale field, is not negligible. We will give some relevant calculations below. Indeed, it turns out that the small-scale dynamo associated with mirror-asymmetric convection does generate helical magnetic fields. However, this helicity is fortunately concentrated on small spatial scales on which the conservation of helicity is violated, which are very small compared to those on which the mean magnetic field is expected to be generated. Therefore, the already cumbersome helicity-balance equations used to calculate the stabilization of the mean-field dynamo need not be made even more unwieldy by including this phenomenon.

Note that the concepts of the mean magnetic field and small-scale magnetic fields were introduced theoretically, and some care should be taken when identifying them with concepts suggested by observations. For this reason, we particularly avoid denoting the mean magnetic field as a large-scale field. However, we assume that the scale of the mean field localized deep in the convection zone and penetrating to the surface, e.g., in the form of sunspots, is comparable with the solar radius. The spatial scale of the fluctuating magnetic fields is assumed to be comparable with the size of supergranules.

Two technical points should be noted before considering the results of our computations.

There are various (but expressible in terms of one another [4]) measures for the mirror asymmetry of the magnetic field. These can be found by calculating the number of linkages between magnetic field lines, electric field lines, or other lines related to the magnetic field [5]. The magnetic helicity is a conserved quantity (the coefficient of magnetic-field line linkage), but it is technically simpler to trace the dynamics of the current helicity; i.e., the coefficient of electric-current line linkage.

To describe convection, we use a short-correlation convective (or turbulent) velocity field as a model; i.e., we assume that its memory time τ is much shorter than the ratio of its correlation scale l_0 to the rms velocity v_0 of the convective motions: $\tau \ll l_0/v_0$). This assumption enormously simplifies the analysis, and it is used, to some degree, in virtually all analytical approaches to studying small-scale dynamos. The short-correlation approximation reproduces fairly well the properties of interstellar turbulence produced by supernova explosions, although there are obviously no reasons to expect its applicability to solar convection. However, as far as can be investigated (see, e.g., [6]), the use of more realistic models for convective (turbulent) fields yields similar results. Therefore, it seems sufficient to us to restrict our analysis to this simple model.

2. GOVERNING EQUATIONS

The correlation tensor of a homogeneous, isotropic, and mirror-symmetric random velocity field in an incompressible fluid is

$$\langle \tau v_i(\mathbf{x}) v_k(\mathbf{x}+\mathbf{r})\rangle \tag{1}$$
$$= 2\tau v_0^2/3[F(r)\delta_{ik} + rF'/2(\delta_{ik} - r_i r_k r^{-2})]$$
$$- v_0\chi(r) e_{ikl} r_l/3,$$

where v_0 is the rms velocity, τ is the correlation time of the velocity field, e_{ikl} is the antisymmetric tensor, $F(r)$ is the longitudinal-correlation function ($F(0) = 1$), $\chi(r) = \dfrac{\tau}{v_0}\langle \mathbf{v}(\mathbf{x})\nabla\mathbf{v}(\mathbf{x}+\mathbf{r})\rangle$, and $\chi(0)$ is the hydrodynamic helicity [7]. Here, $\langle\ldots\rangle$ denotes averaging, and a prime symbolizes differentiation with respect to r. In turn, the correlation tensor of the fluctuating magnetic field can similarly be expressed in terms of the longitudinal-correlation function W and the current helicity $\mu = \langle \mathbf{H}(\mathbf{x})\nabla\mathbf{H}(\mathbf{x}+\mathbf{r})\rangle$. Recall that the magnetic-field line linkage coefficient is measured by the magnetic helicity $\nu(r) = \langle \mathbf{A}(\mathbf{x})\mathbf{H}(\mathbf{x}+\mathbf{r})\rangle$, where $\mathbf{A}$ is the vector potential of the magnetic field $\mathbf{H}$.

The evolutionary equation for the correlation tensor of the magnetic field in a short-correlation flow was obtained by Kazantsev [3]; we use it in the form suggested in [8]. The Kazantsev equation for the correlation tensor of the magnetic field and the Steenbeck–Krause–Rädler equation for the mean magnetic field can be derived by averaging the induction equation over an ensemble of convective pulsations. However, the Kazantsev equation in its general form is very cumbersome, and we will write it only for the particular case of interest to us. To this end, it is useful to introduce an auxiliary quantity R instead of the longitudinal-correlation function W for the magnetic field and an auxiliary quantity κ instead of the current helicity μ (the definitions of R and κ will be given below). We will then utilize the fact that the magnetic Reynolds number Rm is large in the solar convection zone (up to 10^8), so that the inverse quantity $\varepsilon = \mathrm{Rm}^{-1}$ can be used as a small parameter.

This will reduce the Kazantsev equation to the form

$$(2m)^{-1}R'' + (E-U)R = -4(2m)^{1/2}r(\chi - \chi(0))\kappa, \tag{2}$$

$$(2m)^{-1}\kappa'' + (E-\tilde{U})\kappa = 2(2m)^{-1/2}rV(R,\chi), \tag{3}$$

where $(2m)^{-1} = 2\mathrm{R}^{-1} + F(0) - F(r)$, $U(r) = 1/mr^2 + 1/(2r)f' - 1/(8m^3)(m')2$, $\tilde{U}(r) = 2F'/r$, $f(r) = \langle v_i v_i\rangle$, $W = \sqrt{2m}R/r^2$ (this relationship defines the auxiliary function R), $\mu(r) = 2m\kappa/r$ (this relationship defines the auxiliary function κ), $\gamma = -E$ is the growth rate of the magnetic energy, $V(R,\chi) = (\chi - \chi(0))(R'' + MR' + 3M^2/4 - 2/r^2 + mF'')R + \chi'(2R' + MR) + \chi''R$, and $M = (\ln m(r))'$. To be specific, we set $\chi(r) = \chi(0)\exp(-r^2/r_0^2)$, $F(r) = \exp(-r^2/r_0^2)$, and choose r_0 as the unit length.

A typical form of the potential $\tilde{U}(r)$ is shown in Fig. 1. The boundary conditions can be obtained using the condition that the correlations vanish at infinity, a smoothness condition, and the normalization condition $\langle H^2\rangle = 1$:

$$R(0) = 0, \quad R(\infty) = 0, \quad \kappa(0) = 0, \quad \kappa(\infty) = 0. \tag{4}$$

To analyze the Kazantsev equation, it is useful to break the region of the variations of r into three parts [9]—$0 \le r \le a\sqrt{\varepsilon}$ (region I), $a\sqrt{\varepsilon} \le r \le b$ (region II), and $b \le r \le \infty$ (region III), where a and b are constants of order unity. The behavior of the solution in regions I and III ensures that the boundary conditions are satisfied, while the magnetic-energy growth rate is determined by region II, which is thus of most interest to us. In this region, the Kazantsev equation reduces to

$$\xi^2 R'' + (4-\gamma)R = 4\varepsilon^{1/2}\chi(0)\xi^2\kappa, \tag{5}$$

$$\xi^2\kappa'' + (4-\gamma)\kappa = -2\varepsilon^{1/2}\chi(0)\xi^2(\xi^2 R'' + 6\xi R' - 2R), \tag{6}$$

where $\xi = \varepsilon^{-1/4}r$ is a new variable. After this subdivision, we can use the small parameter $\varepsilon = \mathrm{Rm}^{-1}$.

3. NONHELICAL VELOCITY FIELD

Our aim is to study the magnetic field generated by a helical flow. To this end, however, we must consider the properties of the Kazantsev equation for the nonhelical case ($\chi(r) = 0$) in more detail than was done in [9, 10]. In this case, $R(r)$ and $\kappa(r)$ are not coupled, and the corresponding eigenfunctions can be found independently. Recall that the highest-order derivative in the first of these equations is to leading order

$$R(\xi) \sim \varepsilon^{5/4}\ln\varepsilon\,\xi^{1/2}\sin\left(\frac{4\pi}{\ln\varepsilon}\ln\xi\right), \quad \kappa = 0, \tag{7}$$

$$\gamma_R = \frac{15}{4} - \left(\frac{4\pi}{\ln\varepsilon}\right)^2 + O(\ln^{-3}\varepsilon).$$

We will need below the form of this solution in region I: $R(r) \sim \varepsilon^{1/2}r^2$. This solution is normalized to $\langle H^2\rangle = 1$.

The fundamental eigenfunction for κ is similar:

$$\kappa(\xi) \sim \varepsilon^{1/4}\xi^{1/2}\sin\left(\frac{4\pi}{\ln\varepsilon}\ln\xi\right), \quad R = 0, \tag{8}$$

$$\gamma_\kappa = \frac{15}{4} - \left(\frac{4\pi}{\ln\varepsilon}\right)^2 + O(\ln^{-3}\varepsilon).$$

In region I, it has the form $\kappa^{(I)}(r) \sim \varepsilon^{-1/8}r \times \ln^{-1}\varepsilon$. It is more convenient here to use the normalization condition $\int_0^\infty \kappa^2 d\xi = 1$. It is important that not all solutions to the Kazantsev equation are physically realizable; i.e., correspond to the correlation functions of any random fields. In particular, the solution (8) is not realizable, since the magnetic energy vanishes, while the current helicity does not.

4. HELICAL RANDOM FLOW

Let us return to our consideration of a helical flow. The equations for R and κ are now coupled by an operator that has the following form in region II (the first row corresponds to R and the second to κ):

$$\hat{V} = \begin{pmatrix} 0 & -4\varepsilon^{1/2}\chi(0)\xi^2 \\ \chi(0)\left(\xi^2\dfrac{\partial^2}{\partial\xi^2} + 6\xi\dfrac{\partial}{\partial\xi} - 2\right) & 0 \end{pmatrix}. \tag{9}$$

We do not aim to solve this problem fully, but instead will utilize the fact that the helicity is normally small, treat $\chi(0)$ as a small parameter, and apply a perturbation technique. We will only find out via calculations how the perturbation "admixes" the eigenfunctions (8) and (7). We emphasize that, although the function (8) is not physically realizable, this does not prohibit such "admixing," since the mixed state is realizable.

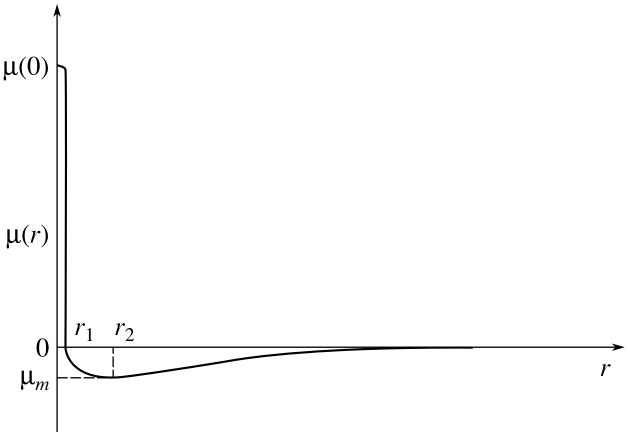

Fig. 2. Current-helicity profile $\mu(r)$; $r_1 = O(l_0\mathrm{Rm}^{-1/2})$, $r_2 = O(l_0\mathrm{Rm}^{-1/4})$.

We assume that the eigensolutions (7) and (8) are not degenerate. This is obviously a severe restriction. It means that we neglect $\chi(0)$ compared to quantities of order $(\ln\varepsilon)^{-3}$. For the solar convection zone, a straightforward application of this condition implies that $\chi(0)$ does not exceed 10^{-3}. This condition may not be satisfied in the solar convection zone, although the uncertainties of the estimates admit the possibility that it is. If it is not satisfied, we must use the perturbation theory for degenerate levels. According to our estimates, the admixture of the helicity to the eigenfunction (7) will not exceed the value given by the formulas for the nondegenerate case. The basic conclusions of our study are not sensitive to this difference [which depends on the logarithmic factor in (12), rather than the power-function factor in the final formula (13)]. We present here the calculation for nondegenerate levels as our basic calculation, since we cannot be sure of the applicability of the logarithmic asymptotics (7) and (8) to high orders, and, in general, two Schrödinger-type equations for different potentials will have different eigenvalues. Of course, this subtle point should be clarified numerically, but this goes beyond the scope of this study.

Standard calculations based on the perturbation formulas [11] reveal that the fundamental eigenfunction of the small-scale dynamo problem in the case of helical turbulence is

$$\mathbf{T} = \begin{pmatrix} R_0 \\ C(\ln\varepsilon)^3\chi(0)\kappa_0 \end{pmatrix}, \qquad (10)$$

where $C = O(1)$ is a constant, and the functions R_0 and κ_0 have unit norms, as is conventional in perturbation theory. After implementing the normalization $\langle H^2\rangle = 1$ natural to the dynamo problem and returning to the variable r, we obtain

$$\mu(r) = C\mathrm{Rm}^{-9/8}\ln^3(\mathrm{Rm}^{-1})\chi(0)\frac{2m}{r}\kappa_0(r). \qquad (11)$$

Figure 2 shows the spatial distribution of the current helicity. The function $\mu(r)$ reaches its maximum at $r = 0$, which can be found using the expression for κ in region I:

$$\mu(0) = C\ln^2(\mathrm{Rm}^{-1})\chi(0). \qquad (12)$$

The estimated value (12) is fairly large, so that the helicity must obviously be included in the calculation of the helicity balance. However, the distribution of current helicity (11) is concentrated on scales of order $r_0\mathrm{Rm}^{-1/2}$, while the dynamo operates on scales $r \sim r_0\mathrm{Rm}^{-1/4}$. For these two scales (see Fig. 2),

$$\mu \simeq C\mathrm{Rm}^{-5/8}\ln^3(\mathrm{Rm}^{-1})\chi(0), \qquad (13)$$

and the degree of mirror asymmetry of the small-scale magnetic field proves to be small, since $\mathrm{Rm} \gg 1$.

5. DISCUSSION

The small-scale dynamo mechanism generates a magnetic field in the form of twisted ropes of thickness $r_0\mathrm{Rm}^{-1/2}$. On scales of the rope thickness, the electric currents are helical. One implication of this helicity is present in the physical pattern of magnetic-field generation suggested by Zel'dovich, according to which a magnetic loop is stretched and twisted into a figure eight. This figure is, indeed, twisted, and the mirror asymmetry of the turbulence is able to choose the preferred sign of twisting. However, the real dynamo mechanism is associated with larger spatial scales, on which this asymmetry is negligible, providing hope that analyses that neglect the generation of helicity by the dynamo mechanism nonetheless give fairly reliable results. Therefore, calculations of the helicity balance can be restricted to including the production of helicity by the large-scale dynamo and the compensation of this process by the small-scale magnetic field coupled with the large-scale field. For example, the helicity balance in the solar convection zone was taken into account in a similar way in [1], and the result of this calculation was compared with the observed current helicity at the solar surface.

We have analyzed here the helicity of fluctuating magnetic fields in the context of the solar dynamo. However, our results are, naturally, also applicable to stellar dynamos, and probably galactic dynamos as well, although the role of the multiphase composition of the interstellar medium requires further study. The application of our findings to the geodynamo and planetary dynamos is most problematic, since the magnetic Reynolds number is not large in these cases.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 03-05-64074), a joint project of the RFBR and the National Natural Science Foundation of China (joint-project no. 02-02-39027), and INTAS (grant 03-51-5807).

REFERENCES

1. N. Kleeorin, K. Kuzanyan, D. Moss, *et al.*, Astron. Astrophys. **409**, 1097 (2003).
2. N. Kleeorin, D. Moss, I. Rogachevskii, and D. Sokoloff, Astron. Astrophys. **361**, L5 (2000).
3. A. P. Kazantsev, Zh. Éksp. Teor. Fiz. **53**, 1806 (1967).
4. N. Kleeorin, I. Rogachevskii, and A. Ruzmaikin, Astron. Astrophys. **297**, 159 (1995).
5. Ya. B. Zeldovich, A. A. Ruzmaikin, and D. D. Sokoloff, *Magnetic Fields in Astrophysics* (Gordon and Breach, New York, 1983).
6. V. G. Lamburt, D. D. Sokolov, and V. N. Tutubalin, Astron. Zh. **77**, 743 (2000) [Astron. Rep. **44**, 659 (2000)].
7. H. K. Moffat, *Magnetic Field Generation in Electrically Conducting Fluids* (Cambridge Univ. Press, Cambridge, 1978; Mir, Moscow, 1980).
8. S. A. Molchanov, A. A. Ruzmaikin, and D. D. Sokoloff, Magn. Gidrodin., No. 4, 67 (1983).
9. Ya. B. Zeldovich, A. A. Ruzmaikin, and D. D. Sokoloff, *The Almighty Chance* (World Sci., Singapore, 1991).
10. V. G. Novikov, A. A. Ruzmaikin, and D. D. Sokoloff, Zh. Éksp. Teor. Fiz. **85**, 909 (1983) [Sov. Phys. JETP **58**, 527 (1983)].
11. L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory* (Nauka, Moscow, 1974, 1989; Pergamon Press, Oxford, 1977).

Translated by A. Getling

Astronomy Reports, Vol. 48, No. 11, 2004, pp. 954–964. Translated from Astronomicheskiĭ Zhurnal, Vol. 81, No. 11, 2004, pp. 1045–1056.
Original Russian Text Copyright © 2004 by Kobanov, Makarchik.

Pulsating Evershed Flows and Propagating Waves in a Sunspot

N. I. Kobanov and D. V. Makarchik

Institute for Solar-Terrestrial Physics, P.O. Box 4026, Irkutsk, 664033 Russia

Received March 19, 2004; in final form, May 27, 2004

Abstract—A comparative analysis of oscillatory spectra based on 66 time series for 14 active regions observed in 2001 shows that, although the chromospheric and photospheric oscillations in the Evershed flow zone possess many common features, there is no firm evidence that the direct and inverse flows have the same physical origin. The interactions between the various oscillation modes and stationary flows results in a complex pattern of wave motions in a sunspot. We studied the Doppler-velocity variations in the sunspot NOAA 0051 during its motion over the disk. The spatial—temporal distribution of the line-of-sight velocity in the chromospheric umbra displays a chevron structure, clearly indicating the presence of propagating waves. These waves move from the center of the umbra to outer regions with a phase speed of 45—60 km/s, a period of 2.8 min, and a measured Doppler speed of 2 km/s. The amplitude of these oscillations decreases abruptly at the boundary between the umbra and penumbra, and the observed waves are not directly related to propagating penumbral waves. Furthermore, the observed pattern of the photospheric velocities shows periodic motions (with a period of 5 min) directed from the inner boundary of the penumbra and superpenumbra toward the line of maximum Evershed velocity. © *2004 MAIK "Nauka/Interperiodica"*.

1. INTRODUCTION

Many puzzles are associated with sunspots, whose solutions will enable considerable progress in understanding fundamental problems in solar physics [1]. A sunspot involves a large number of interactions between motions of material and the magnetic field. The umbra, where the magnetic field is vertical at the photospheric level, is characterized by a downflow of material and weak five-minute oscillations of the entire region [2, 3]. The umbral chromosphere is dominated by powerful three-minute oscillations, which have been interpreted as standing acoustic waves [2, 4]. An even more complex pattern of motions is observed in the sunspot penumbra, where the magnetic field is approximately horizontal.

The former type of motion is associated with oscillations and waves, while the latter is quasi-stationary, and corresponds to so-called Evershed radial flows. These flows are characterized by a fairly clear radial symmetry and a height inversion: they are directed outward from the geometric center of a sunspot in the photosphere and inward in the chromosphere. Is there any relation between these two types of motion? Do the direct and inverse flows represent two independent systems, or do they form a single system? These are only a few questions that arise in studies of the plasma motions in a sunspot penumbra. According to our current understanding, the approximately horizontal magnetic field is concentrated in narrow penumbral filaments. Such filaments appear as dark structures in white light and are oriented in the radial direction for regularly-shaped sunspots. The motions of the material inside the filaments are commonly explained by the siphon mechanism, due to the difference in the gas pressures at the ends of the filaments [5, 6]. It is reasonable to suppose that the preferred direction for the propagation of all waves in the penumbra should be horizontal, with propagation in the vertical direction being suppressed by the horizontal magnetic field. Nevertheless, as follows from observations, penumbral oscillations also propagate upwards, at least those at frequencies of 3—6 mHz [2].

This is quite natural, because the horizontal magnetic field does not cover the entire surface of the penumbra, and there are radial gaps between the dark filaments. When acoustic oscillations propagate in the penumbra atmosphere, they can excite oscillations at various frequencies. The possibility of such a transformation of oscillations in sunspots was first noted by Pikel'ner and Lifshits [7], and later by Zhugzhda and Dzhalilov [8]. In addition, there are traveling waves in the penumbra chromosphere, which propagate in the direction opposite to the quasi-stationary flow identified with the St. John effect. In general, the motions in a sunspot are quite complex, and comparing the characteristics of photospheric and chromospheric oscillations in order to identify the relationships between them is difficult. There is hope that studying the characteristics of oscillations observed simultaneously in a penumbra photosphere and chromosphere may provide new

1063-7729/04/4811-0954$26.00 © 2004 MAIK "Nauka/Interperiodica"

information about the possibly unified origin of Evershed and St. John effects.

We believe that this work should include two stages. First, it is necessary to identify oscillations associated in some way with radial flows. Next, close frequencies at two heights (photospheric and chromospheric) should be separated out for these oscillations and used for a comparative analysis. The first stage by itself is observationally quite difficult. Various researchers have obtained substantially different results for the spectral components of oscillations in the penumbra. Some [2] believe that there are no clear oscillations in the penumbra photosphere, while the period of oscillations at the chromospheric level varies from 4–5 min at the inner boundary of the penumbra to 8–10 min at the outer boundary. In contrast, other researchers [9] have detected oscillations with 5–7 min periods in the penumbra photosphere, whose amplitude is maximum at the outer boundary. Oscillations with longer periods are also observed in the penumbra photosphere, which may be related to the direct Evershed flow [10–12]. There have been far fewer attempts to identify chromospheric oscillations with the inverse Evershed flow [13, 14].

Another important area of study is the spatial–temporal characteristics of the oscillatory motions in a sunspot umbra as a probable source of traveling waves in the penumbra, and of oscillations observed recently in the transition zone and corona above sunspots [15–17]. Although oscillations in sunspot umbras have been widely studied during the last 30 years, the number of questions associated with such oscillations has not decreased (see, for example, the recent reviews on this subject [2, 18, 19]). It was commonly believed that the three-minute oscillations represent standing waves in the umbra chromosphere. However, it is often argued that these oscillations are identical to the "umbral flashes," and are responsible for traveling waves in the penumbra [4, 20]. In addition, we still have few high-quality experimental data that can be used to determine whether waves in an umbra chromosphere are standing or traveling, and to directly measure their phase velocity in the latter case.

2. METHODS AND INSTRUMENTS

The degree to which we can accomplish the tasks outlined above depends on the observational method used. As was noted earlier, the differential method [21] makes it possible to separate out waves propagating in some specific direction from the noiselike mixture of the various wave motions in the sunspot penumbra. In our case, we are interested in the radial direction, along the penumbra filaments. Therefore, precisely those wave motions whose direction coincides with the Evershed flows are predominantly detected even at the observing stage. However, the stationary component of the Evershed velocity, which is the same for both regions of the penumbra under consideration, is excluded from the signal when the differential method is used. This leads to some uncertainty in choosing regions of the penumbra with the most clearly expressed direct and inverse Evershed flows.

Nonmodulational methods for measuring line-of-sight velocities and magnetic fields using modern multichannel photodetectors developed at the Institute for Solar and Terrestrial Physics of the Siberian Division of Russian Academy of Sciences [22] combine the advantages of differential and ordinary methods. Observations were carried out using the horizontal solar telescope of the Sayany Observatory. The coelostat plane mirror with a diameter of 800 mm and the spherical principal mirror with a diameter of 900 mm and a focus of 20 m were made from a glass ceramic to ensure thermal stability. There is a hole in the center of the principal mirror, where an auxiliary spherical mirror with a diameter of 100 mm and a focal length of 19 m is fixed by quartz wedges inclined slightly to the optical axis of the principal mirror. A photoguide–coordinatograph with moving photodetectors for the tracking system is located at the focus of the auxiliary mirror. The photoguide also compensates for the drift of the image due to the rotation of the Sun, and is used for automatic scanning of a specified surface. The scanning parameters can be input both from a control desk or by computer. The set of optical–mechanical units of the telescope provide fixing or scanning of an image with an accuracy of 1″ or better.

The form, amount, and quality of the observational information obtained depend substantially on the characteristics of the multichannel CCD photodetector used. We used two such devices: a Toshiba CCD line (4096 pixels with height 200 μm and width 7 μm) and a Princeton Instruments RTE/CCD-256H array (256×1024 pixels of size 24×24 μm). One pixel corresponds to 0.24″ in the vertical direction. The array is equipped with a cooling system, which can automatically maintain a temperature down to −40°C with an accuracy of 0.05°C, substantially reducing the thermal noise in the receiver and increasing the measurement sensitivity. The array is operated using an ST-133 controller operated by either the WinSpec-32 or WinView-32 software package. This software controls the observations and also provides a number of tools for the preliminary processing of the data obtained.

The observations of velocity oscillations in the Evershed flow zone using the CCD line in the Doppler

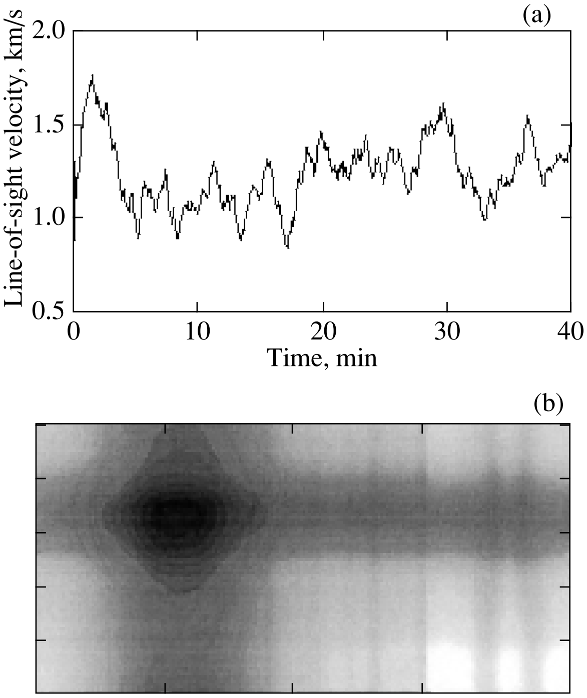

Fig. 1. Primary data: (a) the time series obtained using the CCD line and (b) a single frame obtained with the CCD array. The horizontal axis coincides with the direction of the spectrograph dispersion, and the vertical axis is directed along the entrance slit.

compensation regime have certain characteristic features that must be taken into account. In particular, the control software always sets the first measurement to zero. Thus, the time series must begin with measurements in a region of undisturbed photosphere, usually located at the same meridian as the penumbral region to be studied (so that the rotational velocity will not contribute to the signal). After two or three measurements, without interrupting the running of the software, we began to study the target object. This procedure enabled us to select regions with maximum Evershed velocities. Note that the positions of the maxima in the photosphere and chromosphere may not coincide. As usual, the inverse Evershed effect is observed further from the outer boundary of the penumbra than in the photosphere.

3. OBSERVATIONAL MATERIAL

We shall analyze here data obtained in 2001–2002. We focused our observations on large, isolated, regularly-shaped sunspots. When these were absent, we observed sunspots incorporated into active regions. Examples of the primary data obtained using the CCD line and array are shown in Fig. 1. The CCD-line observations are organized so that a computer calculates the line-of-sight velocity, magnetic-field intensity, and brightness during the measurements and writes them into a common file. An example of our reconstruction of the line-of-sight velocity using the standard software package Excel is presented in Fig. 1a. The CCD-array data cannot be processed using this same procedure directly during the observations due to their large volume. The data volume was slightly reduced by writing only selected parts of the spectrum containing interesting lines into the data file, rather than the entire image. The file for a time series contains a sequence of frames with selected fragments. One frame from such a sequence is presented in Fig. 1b. We can see here two components of the FeI 6569 Å line that result from the action of a deflector [22], which is amplified in the region of the magnetic field.

We used the simple criterion that one of the lines be formed in the deep photosphere and the other in the chromosphere to select the spectral lines to be observed. Since both lines must be measured simultaneously by the same photodetector, they must also be located close to each other in the solar spectrum. The most suitable pairs of spectral lines were Hβ and FeI 4859.8 Å, and Hα and FeI 6569.2 Å; their characteristics are listed in Table 1. A very useful property of the photospheric lines is their magnetic sensitivity, characterized by the Landé factor. Due to this sensitivity, we also obtained, in some of the observations, additional valuable information about the longitudinal component of the magnetic field. The strong Hβ and Hα lines possess very deep and broad profiles, so that we can easily identify separate parts associated with specific heights in the chromosphere. Note that the sensitivity of the measurements of the line-of-sight velocity or the magnetic-field intensity depends on the steepness of the profile, and is maximum in the steepest parts of the profile (i.e., in the middle of the wings, where the profile is almost linear). The width of the spectrograph entrance slit was chosen to find a reasonable compromise between the following two requirements.

On the one hand, the slit width must be large enough to ensure sufficient illumination of the spectrograph, due to a number of effects: the large depth of the Hβ and Hα lines, the decrease of the intensity in the sunspot umbra and penumbra, and limb darkening (in some of our observations, the sunspots are located near the limb, where the Evershed effect is more prominent). On the other hand, increasing the width of the entrance slit can worsen the spectral resolution, first and foremost, for narrow photospheric lines, due to broadening of the instrumental contour. In most cases, the width of the entrance slit was taken to be between 100 and 200 μm, which corresponds to $1''$–$2''$ on the image and is in satisfactory agreement

Table 1. Spectral lines used

Spectral line	λ, Å	Equivalent width, mÅ	Landé factor	Behavior in the sunspot (S/s/W/w)	Parts of the line wing used, Å	Height of formation, km
Hβ	4861.3	3680	1.1	w	±0.2	1000–1500
Hα	6562.8	4020	1.1	W	±0.2; 0.4; 0.7	1500–2000
FeI	4859.8	108		s	±0.1	300–350
FeI	6569.2	71	1.4	w	±0.15	250–300

with the actual attainable spatial resolution, which is limited by the influence of the Earth's atmosphere.

A short description of the observational data for 2001 is presented in Table 2. The number of observations of penumbral regions near the limb and near the center of the solar disk are denoted NL and NC, respectively. As we can see in the table, the sunspots were usually observed at longitudes of 20°–60°. This is quite natural, since the direct and inverse Evershed effects are most clearly visible in these positions. We were sometimes able to track the sunspots along their entire path across the observable part of the disk. For convenient localization of the required object in the spectrograph entrance slit, we used a Dove prism installed just in front of the slit. When working with the CCD array, a sunspot was usually rotated so that the slit was parallel to a line of solar latitude (east–west). As a result, we were able to observe Evershed flows and oscillatory processes simultaneously in two radially-opposite penumbral regions, with the penumbral filaments being located predominantly along the slit. For sunspots of moderate size, the aperture also covered the superpenumbra. The corresponding time series enabled us to study both the characteristics of Evershed flows, and possible relationships between the umbral oscillations and traveling waves in the penumbra.

4. RESULTS

4.1. Spectral Components of Variations in the Line-of-Sight Velocity in the Evershed-Flow Zone

The observational material for 2001 was obtained using the CCD line, while the 2002 observations were carried out using CCD array (1024×256). We can see in Fig. 1 that both the form of the primary data and their informational and quantitative characteristics for these two types of photodetectors are substantially different. In the former case, the primary material can be used directly for a rough visual analysis of the periods, phases, and amplitudes of the observed oscillations. In the latter case, a number of reduction procedures must be applied before beginning the analysis of the above characteristics. One advantage of the CCD-array observations is the possibility of considering the studied process simultaneously in 256 spatial elements, for example, along the east–west section of the sunspot. Such data are most suitable for the analysis of traveling waves.

In 2001, we obtained 66 time series for 14 sunspots. These can be separated into two groups: those in which low-frequency oscillations were detected visually (i.e., were dominant) at both heights and those in which such oscillations were observed at one height only (Fig. 2).

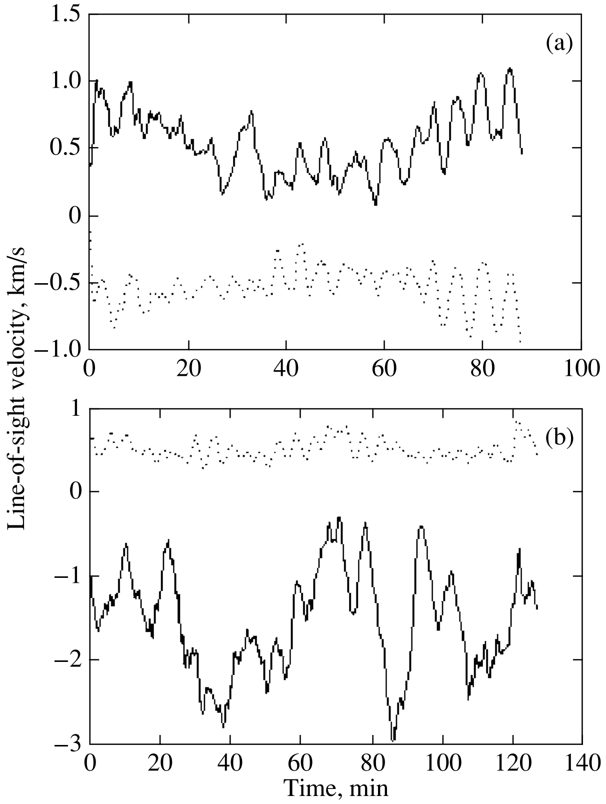

Fig. 2. Explicit manifestation of low-frequency variations in the line-of-sight velocity: (a) at both heights simultaneously and (b) at one height only.

Table 2. Characteristics of the observational data for 2001

Active region, NOAA	Date of observations	Coordinates of the sunspot observed	Number of time series (NC, NL)	Spectral lines
9433	20–24.04	15°N, 60°E–10°E	9C, 5L	Hβ, FeI 4859.8 Å
9435	23.04	20°S, 15°W	1C	Hβ, FeI 4859.8 Å
9436	24.04	10°S, 44°E	1C, 1L	Hβ, FeI 4859.8 Å
9484	03.06	6°S, 21°E	1C	Hα, FeI 6562.9 Å
9487	03–04.06	20°N, 70°E–58°E	2C, 2L	Hα, FeI 6569.2 Å
9488	03–04.06	18°S, 77°E–63°E	1C, 2L	Hα, FeI 6569.2 Å
9529	05, 11–12.07	7°N, 60°E–27°W	1C, 3L	Hα, FeI 46569.2 Å
9535	12.07	6°N, 61°E	1C	Hα, FeI 6569.2 Å
9575	17–20.08	12°N, 6°E–36°W	4C, 7L	Hα, FeI 6569.2 Å
9580	17–24.08	25°N, 46°E–40°W	8C, 2L	Hα, FeI 6569.2 Å
9616	17–19.09	12°S, 16°E–14°W	2C, 3L	Hα, FeI 6569.2 Å
9620	17–19.09	13°N, 67°E–42°E	2C, 4L	Hα, FeI 6569.2 Å
9621	18.09	16°N, 61°E	1C	Hα, FeI 6569.2 Å
9624	19.09	3°N, 57°E–54°E	2C, 1L	Hα, FeI 6569.2 Å

A more accurate quantitative analysis of the power spectra of the line-of-sight velocity carried out using a fast Fourier transform reveals a number of other hidden periods. The histogram in Fig. 3 represents the frequency of occurrence of specific periods in the complete set of 66 series. The periods presented in the histogram denote groups covering all the detected periods. For example, the 5-min group contains all periods greater than 4 min and less than 6 min, the 14-min group covers the interval from 12 to 17 min, and so on.

The low frequency of occurrence of the 3-min (2 to 4 min) oscillations is striking. These oscillations dominate in the chromosphere over sunspot umbras, as well as in the undisturbed chromosphere, beyond active regions. Using the CCD line, we observed primarily oscillations in the middle of the penumbra. Therefore, this result may reflect the fact that there was no significant scattered light in most of our images, so that such light did not lead to faulty periods associated with neighboring regions.

Note the large fraction of simultaneous observations of oscillations with the same periods in the photosphere and chromosphere. This is especially clear for the groups of 5- and 8-min oscillations. There are far fewer such coincidences in two other groups of periods in the middle of the histogram (14 and 11 min), although precisely these periods are commonly assumed to be associated with variations in the Evershed flows at the photospheric level [10, 23, 24]. The phase relations between the photospheric and chromospheric line-of-sight velocity in these groups of periods are random. This provides indirect evidence that the direct and inverse Evershed flows are probably not part of a unified system, with variations in the velocities of photospheric flows leading to corresponding variations in the velocities of chromospheric flows. The periods of 35 min or more observed in the measured Evershed velocity may be produced by torsional oscillations of the sunspots [25, 26].

4.2. Spatial–Temporal Characteristics of the Oscillations, Wave Motions

At the end of July and the beginning of August 2002, we were able to conduct a series of observations of the isolated sunspot NOAA 0051 during its passage across the disk. Using the CCD array, we obtained ten time series from July 29 to August 6, 2002. The average duration of each series was 45 min, with an interval between frames of 5 s. The Doppler velocity in the Hα line was determined as the difference of the intensities in the red and violet wings normalized to their sum at levels of ±0.2 Å, ±0.4 Å, and ±0.7 Å; and in the FeI 6569 Å line at a level of ±0.15 Å. The instrumental shifts of the spectrum were determined using an H_2O telluric line located near Hα. These shifts were subtracted from the calculated signals after recalculation to the equivalent Doppler velocity.

The resulting gray-scale images of the spatial and temporal distributions of the line-of-sight velocity in the Hα line show a clear periodic structure resembling a chevron. The location of this chevron on the time axis directly demonstrates the presence of traveling waves in the umbra chromosphere (the right-hand diagrams in Fig. 4 and the left-hand image in Fig. 5). First and foremost, this structure indicates the spatial symmetry of the process: the observed motions are directed from the center to outer regions of the umbra. The light areas in all the gray-scale diagrams correspond to velocities directed toward the observer, whereas dark areas denote motion away from the observer.

It is difficult to imagine any artificial mechanism that could create such a structure. In addition, the influence of an artefact should be the same in the photosphere, so that the two distributions should be similar to each other. In fact, they differ sharply (Fig. 4) in terms of both the periodicity and the location of the structures. The chevron pattern in some parts of the time series is so clear that the period T and phase velocity $V = L/T$ of the traveling wave can be derived directly from the pattern (left-hand image in Fig. 5). After averaging over several parts of the series with the most clearly expressed chevron structure, we obtained the approximate values $T = 3$ min and $V =$ 45–50 km/s. According to our estimates, the top of the chevron coincides with the center of the umbra. If the distance between the elements corresponding to points B and C in Fig. 5 is 2″, the temporal distributions of the line-of-sight velocity at Hα ±0.2 Å should resemble the right-hand plot of Fig. 5. The signal for the line-of-sight velocity at point C is characterized by an average time delay of about 25 s relative to the signal at point B. Since the distance between these points is 1500 km, the phase velocity turns out to be slightly less than 60 km/s, which almost coincides with the values derived directly from the chevron pattern.

Thus, the estimates of the phase velocity obtained using the two methods coincide. There is a maximum in the power spectrum at 2.8 min. The horizontal size of the region assumed to be the source of the wave motions and localized at the center of the umbra can be estimated as 1.5″–2″. Note that the amplitude of the line-of-sight velocity measured at Hα±0.2 Å exceeds the corresponding values at the levels ±0.4 Å and ±0.7 Å. We did not find a phase delay between the line-of-sight velocities measured at Hα ±0.2 Å and Hα ±0.7 Å for the same spatial element. This may indicate that the vertical extent of the source is comparable to the observed height scale (1000–1500 km); in any case, it should not be below this value. Therefore, the problem of more accurate height localization of the source remains unsolved. Our preliminary study shows that the 3-min oscillations are very weakly expressed in the brightness of the Hα line, and the chevron structure is absent. In the first stage, we did not detect any definite phase relation between the observed oscillations in the line-of-sight velocity and the intensity. Note that the chevron structure is also manifest to some degree in other time series. It is expressed most clearly in the observations of July 31, 2002. We have analyzed these data most completely, and will discuss them here.

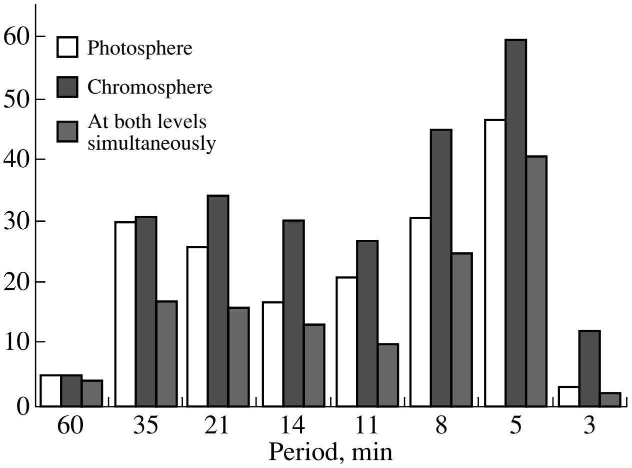

Fig. 3. Frequency of occurrence of the oscillatory periods.

The well-defined localization of the wave motions within the boundaries of the sunspot umbra, determined by the intensity of the continuous spectrum, is striking. The spatial size of the chevron pattern (11″) coincides with the size of the umbra (10″). Our primary observations on July 31 show that there are no appreciable features associated with the continuation of the wave motions to the penumbra region in the directions observed. They finish at or just behind the boundary of umbra, as we can see in both Fig. 4 and Fig. 6. The traveling penumbral waves are probably not seen in Fig. 4 (upper right-hand diagram) due to the small amplitude of the line-of-sight velocity (1 km/s). However, the direct and inverse Evershed flows are expressed very clearly in Fig. 4, although their projection onto the line of sight in this position of the sunspot yields a close velocity (1.1 km/s). Thus, it seems that traveling penumbral waves are very weak or completely absent in the observed directions. Nevertheless, we applied a special procedure to try to detect them.

The initial spatial–temporal distributions of the Doppler velocities (upper right-hand diagram in Fig. 4) involve both Evershed flows and oscillatory–wave motions. This representation may be convenient

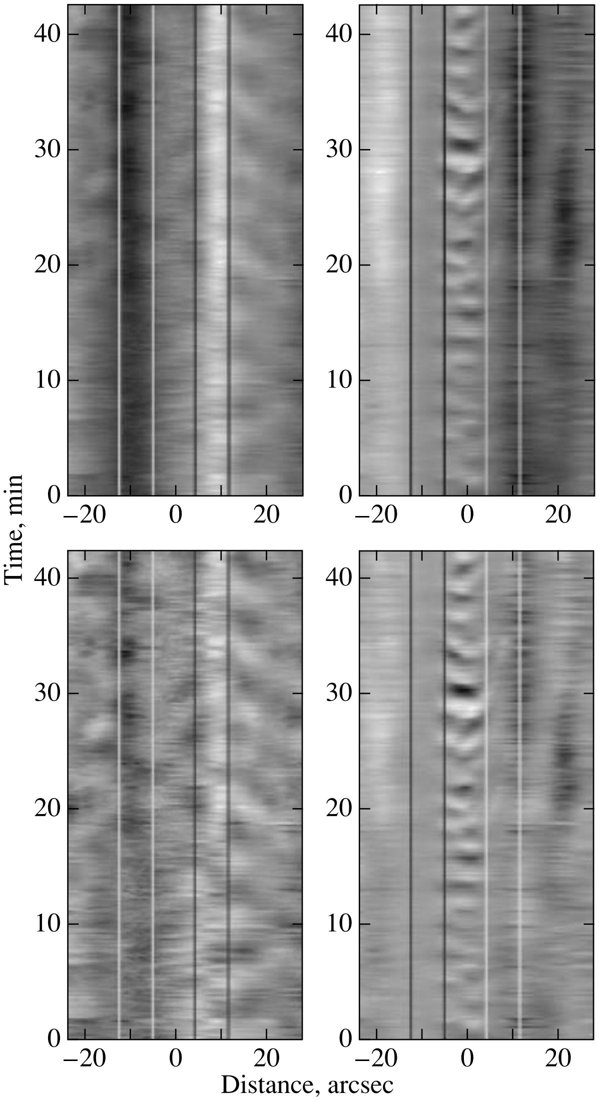

Fig. 4. Spatial–temporal distributions of the Doppler velocity in the photosphere (left-hand column) and chromosphere (right-hand column) of the sunspot NOAA 0051 (Hα ±0.2 Å). The Evershed velocities in the bottom row are decreased by a factor of four to show the weak wave motions. The vertical lines denote the inner and outer boundaries of the penumbra.

for investigating the relationship between these phenomena. In this case, the Evershed flows presented in the pictures can serve as additional calibration information. However, the speeds of the direct and inverse Evershed flows are substantially greater than the speeds associated with other types of motions. As a result, motions with small velocities will not be seen in the initial gray-scale diagrams due to the insufficient dynamic range of the gray scale. In order for the structure of weak wave motions to be presented more broadly, the constant component of the Evershed velocity should be subtracted (or substantially reduced).

The results of such a transformation are presented by the two bottom diagrams in Fig. 4, which show weak wave structures in the penumbra chromosphere at various time intervals. A more detailed analysis shows that these structures usually correspond to those parts of the diagram where the chromospheric chevron in the umbra becomes more gently sloping and flat. Such behavior may indicate an increase in the velocity of propagation of the wave in the umbra in the corresponding time intervals. We can suppose that such waves cross the boundary of the penumbra before they lose energy and disappear. In fact, precisely the high speed of propagation of the wave combined with insufficient frequency of exposure of the sunspot may be responsible for the fact that the umbral chevrons in the spatial–temporal diagrams published in [20, 27, 28] degenerated into flat structures. The typical intervals between the exposures in these works were 12 to 36 s, while this interval was 5 s in our study. Naturally, the authors of [20, 27, 28] could not quantitatively estimate the speed of propagation of the wave based on their diagrams directly in the sunspot umbra, since this would lead to extremely large values for the phase velocity. (A standing wave possesses an infinite phase velocity.) In this situation, these incorrect quantitative estimates would seem to provide evidence against the propagation of waves in the umbra.

It is interesting to compare the velocity oscillations at points located symmetrically about the top of the chevron ($\pm 3''$). The phases of the signals are close to each other, especially in the time interval 20–40 min (Fig. 7b). This suggests the synchronous propagation of chromospheric oscillations from the center of the umbra to the east and west. It is natural to suppose that the dominant direction of the wave motions in the umbra coincides with the magnetic field, i.e., it is approximately vertical. In this case, the measured horizontal velocity of propagation is a projection of the true velocity, which may be substantially greater. If we consider the Doppler velocities at both sides of the fir-tree-like structure observed at the photospheric level, the situation will be exactly the opposite: the signals will be in antiphase (Fig. 7a). Since two oppositely-directed wave motions take place here, there should be a considerable decrease in their amplitude due to interference. Precisely this effect is observed near the maximum of the Evershed velocity in Fig. 7a.

We do not know now if this is a typical situation that also occurs in other sunspots, or if it is characteristic of the particular sunspot studied here. This may be related to the double structure of the Evershed-velocity maximum in Fig. 4. Features with such double structures were noted by Bumba *et al.* [29].

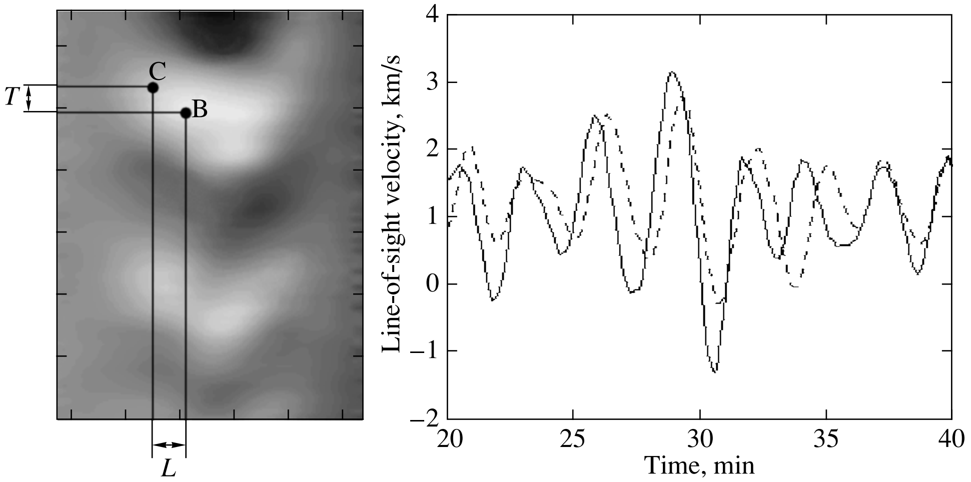

Fig. 5. Scheme for determining the velocity of propagation of the wave process (a) directly from the chevron pattern and (b) from the phase delay of the signals at points B (solid curve) and C (dashed curve).

We can see in Fig. 4 that the pattern of the motions at the photospheric level is quite complex, even for this isolated sunspot with a regular shape. Oscillatory motions are observed in the sunspot umbra, and it seems that the Evershed flow itself pulsates with the 5-min period in some regions; in any case, the curves representing the maxima Evershed velocity are wavy.

The observed line-of-sight velocity can be represented as the sum of two orthogonal components: the vertical (perpendicular to the solar surface) component V_r and the horizontal component V_h. The same approach was used in [30]. Due to the axial symmetry of the motions in a sunspot, the Evershed flow and its variations will be presented primarily by the horizontal component. Therefore, this provides a new means for identifying the oscillations associated with Evershed flows.

We can study the separate oscillatory modes in more detail using frequency filtration. We performed a direct wavelet transformation of the data on the space and time distribution of the analyzed velocity component. Next, we identified interesting frequencies and restored the initial distribution using the inverse wavelet transformation. The resulting patterns describe more clearly the behavior of individual oscillatory modes. For example, Fig. 8a shows that the 5-min photospheric oscillations experience a sharp phase jump near the middle of the penumbra. The 10–12-min oscillations in the photosphere (Fig. 8b) experience two such jumps; the first appears approximately in the same place as for the 5-min mode, 2″–3″ closer to the inner boundary, while the second jump occurs just behind the outer boundary of the penumbra. The zone of maximum Evershed flow plays an unclear, but obviously important, role in the propagation of oscillations in the sunspot photosphere. It is difficult to interpret the corresponding data, and more extensive observational data, supplemented by series of sensitive filtergrams with high temporal resolution, are needed.

5. DISCUSSION AND CONCLUSIONS

Could the well-known umbral flashes [31–34] and the traveling umbral waves observed in our study represent the same phenomenon? They have a number of common features: the periods of oscillations are close to 170 s, the proagation velocities nearly coincide (40 km/s for umbral flashes and 45–50 km/s

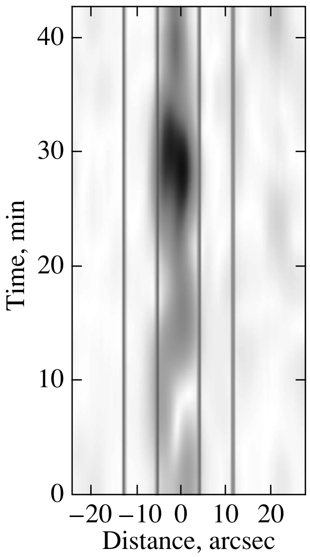

Fig. 6. Wavelet diagram illustrating the spatial–temporal distribution of the power in the 3-min mode in the chromosphere of the sunspot NOAA 0051 (July 31, 2001).

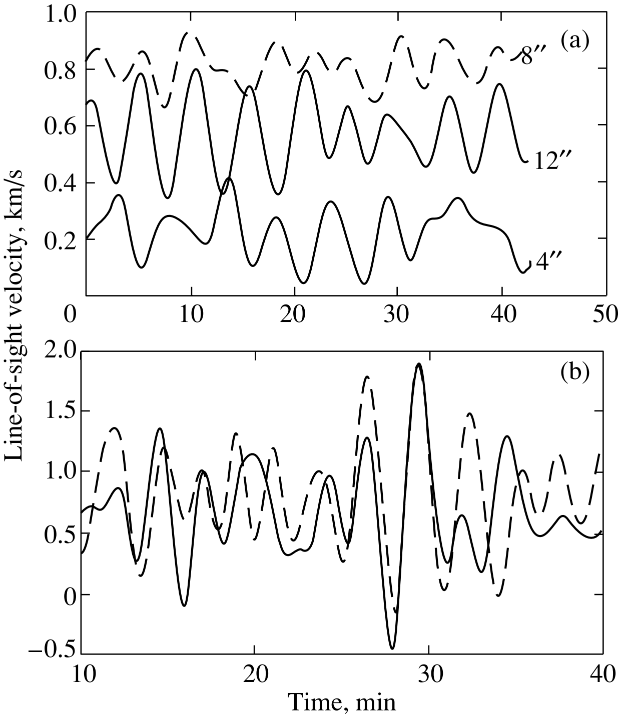

Fig. 7. Phase relations between the Doppler-velocity signals in (a) the photosphere and (b) the chromospheric chevron of a sunspot. The coordinates are measured in arcseconds from the center of the sunspot (negative to the left, positive to the right). For the photosphere, the coordinates are given near the corresponding curves. For the chromosphere, the solid curve corresponds to $-3''$ and the dashed curve to $+3''$.

in our case), the scales are no greater than 3″, and both phenomena are observed within the sunspot umbra. Differences between the two phenomena are as follows. First, no brightening in Hα is observed in our case. The intensity variations are very small and do not show an unambiguous phase relation relative to the velocity variations. Second, the periods of the umbral flashes are different for different positions in the sunspot umbra, and several umbral flashes with different periods can be observed simultaneously. In our case, the period remains constant over the entire observed umbra (which is not small).

Next, let us consider the possible relation of the observed phenomenon to the traveling penumbral waves [35, 36]. Giovanelli [35] found an absence of oscillations propagating in the radial direction inside the sunspot umbra ($r = 0.9$). According to his observations, the traveling waves are formed in a narrow zone of the umbra (between $r = 0.9$ and 1) and propagate outwards with speeds of 20 km/s in the sunspot penumbra. The period of these waves is about 300 s, and the amplitude of the line-of-sight velocity is 1 km/s.

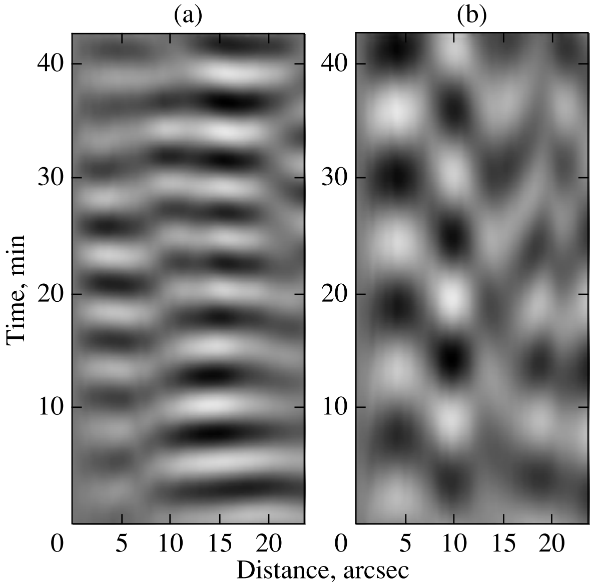

Fig. 8. Jumplike variations in the phases of separate modes of the photospheric oscillations: (a) 5-min period and (b) 10.5-min period. The distance is measured in arcseconds from the center of the sunspot to the west.

Our results are substantially different. We observed waves propagating in the radial direction directly from the center of the sunspot umbra. Moreover, these waves disappeared abruptly at or near the boundary between the umbra and penumbra. The propagation velocity in the radial direction was 45–60 km/s, the period ~170 s, and the amplitude of the line-of-sight velocity about 2 km/s.

The later works [20, 27] report that the traveling waves are formed in the umbra and next propagate through the penumbra with speeds of 20–30 km/s. Unfortunately, these authors do not discuss the disagreement between the two periods (170 s in the umbra and 280–300 s in the penumbra). They observed several oscillatory elements with various periods and amplitudes in the sunspot umbra, and also found some features oscillating with a period of 80 s. In addition, they did not estimate quantitatively the propagation velocity for the wave in the umbra, only noting that the phase velocity was very large. They suggest that the umbral oscillations and the traveling waves are produced by the same resonator.

Moore and Tang [32] believe that the sources of umbral oscillations and traveling penumbral waves are physically independent, and are localized in deeper subphotospheric layers. Returning to the bottom left-hand diagram in Fig. 4, we can see a sharp difference of the spatial–temporal distributions for the line-of-sight velocity in the photosphere measured by the FeI 6569 Å and Hα lines. There are no 3-min oscillations in the umbra at the photospheric level, but there are weak 5-min-oscillation features. On the other hand, the penumbra region is characterized by a clear periodic structure, whose center of symmetry is the line of maximum Evershed velocity. The period of this firtree-like structure is about 5 min or, according to a more accurate spectral analysis, 5.2 min. The wave motions are directed in opposite directions, from the inner part of the penumbra and the outer penumbra. The waves flow into the zone of maximum Evershed velocity, simulating pulsations of the Evershed flow with a 5-min period.

In fact, our results contradict the suggestion of [4, 28] that the 3-min oscillations in the umbra chromosphere are composed primarily of standing waves [37]. In the photosphere of the inner penumbra, there are waves directed inwards, toward the sunspot umbra, whereas the waves in the outer penumbra (0.7 of the distance between the umbra and penumbra boundaries) are directed outwards from the sunspot. The wave pattern is repeated with a period of 5 min. The speed of propagation of the wave in both directions is about 0.5 km/s, and the wave has a horizontal wavelength of 2500 km. The reason for these discrepancies with our results is unclear. It is likely that the individual characteristics of sunspots determine the oscillatory processes to a much greater degree than we believe. Further studies of data for a large number of sunspots are necessary.

It is interesting that some of the patterns seen in the most recent numerical simulations of wave propagation in sunspots [38] are very similar to the chevron structures observed by us.

Let us now briefly summarize our results. Our analysis of variations in the line-of-sight velocity using 66 data series for 14 active regions does not support the hypothesis that the direct and inverse flows form a single system. We believe that the similarity of the oscillatory spectra for the direct and inverse flows in the period range 25–35 min could be associated with torsional oscillations of the sunspots. The chevron structure in the spatial–temporal distribution of the line-of-sight velocity in the sunspot convincingly indicates the presence of traveling waves in the umbra chromosphere. These waves propagate from the center of the sunspot outwards with a period of 2.8 min and a phase velocity of 45–60 km/s. The measured amplitude of the Doppler velocity is about 2 km/s. These waves disappear quite abruptly at the inner boundary of penumbra and, therefore, do not propagate into the penumbra itself. The spatial coherence of the motions does not exceed $2''$. There are periodic motions (with a period of 300 s) at the photospheric level directed from the inner boundary of the penumbra and from the superpenumbra toward the line of maximum Evershed velocity.

We believe that the wave motions observed in the sunspot umbra chromosphere are not directly related to traveling penumbral waves. This conclusion is supported by the absence of a clear continuation of the wave motions from the umbra to the penumbra, and also by the difference of the periods found in the studies cited above. However, oscillations with a period of about 300 s may sometimes be excited at the boundary between the umbra and penumbra when the amplitude of the umbral oscillations becomes sufficiently large (6–8 km/s), accompanied by the appearance of umbral flashes. Thus, this is not a direct continuation of the wave motions. An alternative explanation is that the observed traveling 5-min waves from the lower layers (photosphere) penetrate directly into the penumbra chromosphere. Longer time series of spectral observations (up to 2 h) supplemented by filtergrams in calcium lines and the Hα line, as well as information about the magnetic structures in the sunspots, are necessary for the further development of such investigations.

ACKNOWLEDGMENTS

This work was supported in part by the state support for Leading Scientific Schools of the Russian Federation (project no. 55-733.2003.2) and the Federal "Astronomy" Program (project no. 1105).

REFERENCES

1. E. X. Parker and J. Chinese, Astron. Astrophys. **1** (2), 99 (2001).
2. B. W. Lites, *Sunspot: Theory and Observations*, Ed. by J. H. Thomas and N. O. Weiss (1992), p. 261.
3. N. I. Kobanov, Solar Phys. **125**, 25 (1990).
4. E. B. Georgakilas and S. Koutchmy, Astron. Astrophys. **363**, 306 (2000).
5. B. Montesinos and J. H. Thomas, Nature **390**, 485 (1997).
6. R. Schlichenmaier and W. Schmidt, Astron. Astrophys. **349**, L37 (1999).
7. S. B. Pikel'ner and M. A. Lifshits, Astron. Zh. **41**, 1007 (1964) [Sov. Astron. **8**, 808 (1964)].
8. Yu. G. Zhugzhda and N. S. Dzhalilov, Astron. Astrophys. **132**, 45 (1984).
9. E. Marco and W. Mattig, in *Solar Magnetic Fields*, Ed. by M. Schussler and W. Schmidt (Cambridge Univ. Press, 1994), p. 257.
10. T. R. Rimmele, Astrophys. J. **445**, 511 (1995).
11. M. Sigwarth and W. Mattig, Astron. Astrophys. **324**, 743 (1997).
12. N. I. Kobanov, Astron. Zh. **77**, 233 (2000) [Astron. Rep. **44**, 202 (2000)].
13. N. I. Kobanov and D. V. Makarchik, *Current Theoretical Models and Future High Resolution Solar Observations: Preparing for ATST*, Ed. by A. A. Pevtsov and H. Utenbroek, ASP Conf. Ser. **286**, 251 (2003).

14. A. A. Georgakilas, E. B. Christopoulou, A. Skodras, *et al.*, Astron. Astrophys. **403**, 1123 (2003).
15. A. Nindos, C. E. Alissandrakis, G. B. Gelfreikh, *et al.*, Astron. Astrophys. **386**, 658 (2002).
16. J. G. Doyle, E. Dzifcakova, and M. S. Madjarska, Solar Phys. **218**, 79 (2003).
17. N. Brynildsen, P. Maltby, O. Kjeldseth-Moe, *et al.*, Astron. Astrophys. **398**, L15 (2003).
18. T. J. Bogdan, Solar Phys. **192**, 373 (2000).
19. J. Staude, *Third Advances in Solar Physics Euroconference: Magnetic Fields and Oscillations*, Ed. by B. Schmieder, A. Hofman, and J. Staude, ASP Conf. Ser. **184**, 113 (1999).
20. K. Tziotziou, G. Tsiropoula, and P. Mein, Astron. Astrophys. **381**, 279 (2002).
21. N. I. Kobanov, Astron. Zh. **77** (12), 940 (2000) [Astron. Rep. **44**, 830 (2000)].
22. N. I. Kobanov, Prib. Tekh. Éksp. **4**, 110 (2001).
23. A. A. Georgakilas and E. B. Christopoulou, Astrophys. J. **584**, 509 (2003).
24. R. van der Voort, Astron. Astrophys. **397**, 757 (2003).
25. S. I. Gopasyuk, Astron. Zh. **62**, 157 (1984) [Sov. Astron. **28**, 93 (1984)].
26. A. A. Pevtsov, Candidate's Dissertation (1992).
27. G. Tsiroloula, C. E. Alissandrakis, and P. Mein, Astron. Astrophys. **355**, 375 (2000).
28. E. B. Christopoulou, A. A. Georgakilas, and S. Koutchmy, Astron. Astrophys. **375**, 617 (2001).
29. V. Bumba, M. Klvana, and A. Garcia, in *Proceedings of the SOLMAG: Magnetic Coupling of the Solar Atmosphere* (2002), p. 365.
30. R. Schlichenmaier and W. Schmidt, Astron. Astrophys. **358**, 1122 (2000).
31. J. M. Beckers and P. E. Tallant, Solar Phys. **7**, 351 (1968).
32. R. L. Moore and F. Tang, Solar Phys. **41**, 81 (1975).
33. I. P. Turova, R. B. Teplitskaia, and G. V. Kuklin, Solar Phys. **87**, 7 (1983).
34. Rouppe vander Voort, P. J. Rutten, P. Sutterlin, *et al.*, Astron. Astrophys. **403**, 277 (2003).
35. R. G. Giovanelli, Solar Phys. **27**, 71 (1972).
36. H. Zirin and A. Stein, Astrophys. J. **178**, L85 (1972).
37. E. B. Christopoulou, A. A. Georgakilas, and S. Koutchmy, Astron. Astrophys. **354**, 305 (2000).
38. T. Bogdan, M. Carlsson, V. Hansteen, *et al.*, Astrophys. J. **599**, 626 (2003).

Translated by Yu. Dumin